

Modelling clustered survival data from multicentre clinical trials

David V. Glidden^{*,†} and Eric Vittinghoff

Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94143-0560, U.S.A.

SUMMARY

In randomized clinical trials, subjects are recruited at multiple study centres. Factors that vary across centres may exert a powerful independent influence on study outcomes. A common problem is how to incorporate these centre effects into the analysis of censored time-to-event data. We survey various methods and find substantial advantages in the gamma frailty model. This approach compares favourably with competing methods and appears minimally affected by violation of the assumption of a gamma-distributed frailty. Recent computational advances make use of the gamma frailty model a practical and appealing tool for addressing centre effects in the analysis of multicentre trials. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: centre effects; frailty models; multicentre trials; survival analysis; treatment-by-centre interaction

1. INTRODUCTION

In randomized clinical trials, subjects are recruited at multiple study centres. The multicentre design can provide adequate sample sizes and enhance generalizability of study results. However, factors that vary by centre, including patient characteristics and medical practice patterns, may exert a powerful influence on study outcomes. Such centre effects potentially lead to clustering, or dependence between outcomes at each centre. If these effects are sufficiently powerful, then inferences that ignore the clustering can be seriously misleading. Data analysts must choose when and how to incorporate centre effects into the analysis.

These issues have been addressed in the context of clustered continuous or binary response data [1–8]. This paper discusses options for the analysis of clustered time-to-event data. We review various approaches and call particular attention to the utility of random effects or frailty modelling. Frailty modelling provides a flexible, efficient framework for accommodating centre effects in a variety of practical settings.

*Correspondence to: David V. Glidden, Department of Epidemiology and Biostatistics, University of California, San Francisco, 500 Parnassus Avenue, MU-420 West, Box 0560, San Francisco, CA 94143-0560, U.S.A.

†E-mail: dave@biostat.ucsf.edu

Issues that arise in the analysis of multicentre clinical trials include: (i) the detection of centre effects, (ii) estimation of a common treatment effect in the presence of centre effects, (iii) estimation of centre effects, and (iv) addressing treatment-by-centre interaction. This paper will review existing approaches, present the results of simulations comparing several of these methods, and point to areas where further development is needed.

1.1. Data and notation

Suppose that we observe censored time-to-event data from a study with K clusters (or centres) and n_k subjects per cluster ($k = 1, \dots, K$), so that $N = \sum_{k=1}^K n_k$ is the total sample size. Let T_{ki} be the latent failure time with C_{ki} denoting the censoring time for subject i in cluster k . Also assume we observe a p -dimensional vector of covariates Z_{ki} . Let \mathbf{T}_k be the vector $(T_{ki}, i = 1, \dots, n_k)$ with \mathbf{C}_k and \mathbf{Z}_k defined similarly. Assume that \mathbf{T}_k , \mathbf{C}_k , and \mathbf{Z}_k are independent across centres and $(\mathbf{T}_k, \mathbf{C}_k)$ ($k = 1, \dots, K$) are independent and identically distributed with the components of \mathbf{T}_k and \mathbf{C}_k conditionally independent given \mathbf{Z}_k . We observe $X_{ki} = \min(T_{ki}, C_{ki})$ and $\Delta_{ki} = I\{T_{ki} \leq C_{ki}\}$, the follow-up time and failure indicator, respectively. Let the function $Y_{ki}(t)$ represent the at-risk function $I\{X_{ki} \geq t\}$ and denote as $\mathcal{H}_k(t)$ the history of failure, censoring and covariate values for the entire k th cluster observed up to time t . We also assume $\max_k \sum_{i=1}^{n_k} \Delta_{ki} > 1$, that is, at least one cluster has more than a single failure.

1.2. Conditional versus marginal Cox modelling

In the presence of the dependence induced by centre effects, two distinct approaches are available: conditional (or centre-specific) and marginal (or population-averaged) models. The two strategies differ in methods for estimation as well as interpretation. The virtues of each approach have been extensively debated [9–11].

To highlight the conceptual differences, consider a simple example. Suppose a Cox model with a single predictor Z , coded 0 or 1, is used to make an unadjusted two-group treatment comparison in a multicentre clinical trial. The conditional approach explicitly considers centres in the formulation of the hazard of failure for subject i in centre k :

$$\lambda_{ki}(t | Z_{ki}) := \lim_{h \downarrow 0} h^{-1} \text{pr}\{t \leq T_{ki} < t + h | \mathcal{H}_k(t)\} = \lambda_{0k}(t) \exp(\beta^T Z_{ki}) \quad (1)$$

Centre effects are incorporated in equation (1) through the centre-specific baseline hazard $\lambda_{0k}(t)$. The model is conditional in the sense that the hazard ratio for treatment is with respect to—that is, conditional on—the centre-specific baseline hazard. Specifically, e^β is the hazard ratio comparing two subjects from the same centre, one treated and one untreated. More generally, conditional regression parameters model the effects of covariate differences between two members of the same centre. A key feature of the conditional Cox model is the decomposition of the hazard into a term for the centre-specific baseline hazard and an exponential term which models the multiplicative effect of covariates on this baseline risk. Variants of this conditional model arise from postulating different structures for the heterogeneity of the baseline hazard across centres.

In the marginal approach [12–14], centre effects are ‘averaged out’ by the model. The marginal hazard of failure for subject i in centre k is

$$\mu_{ki}(t | Z_{ki}) := \lim_{h \downarrow 0} h^{-1} \text{pr}\{t \leq T_{ki} < t + h | T_{ki} \geq t, Z_{ki}\} = \mu_0(t) \exp(\gamma^T Z_{ki}) \quad (2)$$

In this model, the baseline hazard $\mu_0(t)$ is not specific to centre, and an averaged effect of treatment is modelled by the exponential term. Thus the marginal hazard, which does not explicitly incorporate centre effects, represents the average hazard in the subset of the population remaining at risk at time t , and e^β is the hazard ratio comparing the risk of failure of two randomly selected members of the population: one treated and one untreated.

Note that while centre effects do not enter (2) explicitly, valid inferences from model (2) would require that clustering be taken into account. ‘Robust’ Lee *et al.* [14] standard errors implemented in some statistical packages are one method for doing this.

The contrast between models is well-illustrated by comparing the marginal hazard that would be obtained under the gamma frailty model, a mathematically tractable conditional model, with the hazard under the simple marginal model. In the gamma frailty model, $\lambda_{0k}(t) = \xi_k \lambda_0(t)$, where the ξ_k ($k = 1, \dots, K$) are centre effects distributed as independent and identically distributed gamma random variables with mean 1 and variance θ_0 . The variance parameter is interpretable as a measure of the heterogeneity across centres in baseline risk. When θ_0 is small, then values of ξ are closely concentrated around 1 and the centre effects are small. If θ_0 is large, then values of ξ are more dispersed, inducing greater heterogeneity in the centre-specific baseline hazards $\xi \lambda_0(t)$. The centre-specific baseline hazards are all proportional to $\lambda_0(t)$.

Under this model, the marginal hazard function can be expressed as the expectation of the hazard function [15], conditional on being at risk at t and covariate value Z_{ki} ; that is

$$\mu(t | Z) = E(\xi | T \geq t, Z) \lambda_0(t) \exp(\beta^T Z)$$

Taking the expectation represents an averaging over centre-specific sub-populations. Since the expectation is conditional on being at risk at time t , it constitutes averaging over a subset of the originally randomized cohort. Because centres with larger values of ξ_k have higher hazards, smaller proportions of the subsamples from these centres remain at risk at t , lowering the average hazard among the subsets still in follow-up. To make this more clear, consider the form of the conditional expectation for the gamma distributed random effect:

$$E(\xi | T \geq t, Z) = \{1 + \theta_0 \Lambda_0(t) \exp(\beta^T Z)\}^{-1}$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$. Clearly, the average frailty value is a decreasing function of time. The decrease is faster in situations with greater heterogeneity of the centre effects, as measured by θ_0 , a higher cumulative baseline hazard $\Lambda_0(t)$, and larger increases in risk due to covariates, which are modelled by $\exp(\beta^T Z)$. All these factors accelerate the ‘frailty selection’ which results from heterogeneity in the baseline risks.

Frailty selection also affects the marginal hazard ratio for treatment. When the data follows the conditional gamma frailty model (1), the marginal hazard ratio is [16]

$$\frac{\mu(t | Z = 1)}{\mu(t | Z = 0)} = e^\beta \left\{ \frac{1 + \theta_0 \Lambda_0(t)}{1 + e^\beta \theta_0 \Lambda_0(t)} \right\} \tag{3}$$

A critical point here is that under the proportional hazards specification for the conditional model, the ratio of the treatment-specific marginal hazards is not generally time-invariant; unless θ_0 or β is 0, the marginal hazard ratio (3) is a decreasing function of time. Only at time 0 is the marginal hazard ratio equal to the true conditional hazard ratio e^β , as t increases and $\Lambda_0(t) \rightarrow \infty$, the ratio tends to 1.

Under these conditions, the time-invariant hazard ratio obtained by applying a simple marginal model is attenuated from the conditional hazard ratio e^β . The attenuation is increased in situations with larger values of θ_0 , $\Lambda_0(\cdot)$, and β , representing the three factors which accelerate frailty selection. Conversely, the attenuation is modest if any of these parameters is near zero, in that case the centre-specific and population-averaged parameters are relatively close. Formal expressions and approximations were obtained by Henderson and Oman [17], who considered the effects of fitting a marginal model to conditional covariate effects. Interestingly, it can be shown that marginal and conditional effects are identical in the accelerated failure time model.

2. CONDITIONAL MODELS FOR SURVIVAL DATA

In this section we present three approaches to modelling centre effects through conditional models. All the models we consider have the form

$$\lambda_{ki}(t) = \lambda_{0k}(t) \exp(\beta^T Z_{ki}) \quad (4)$$

These models allow for variation in baseline risk across centres, but assume that the treatment effect is the same in all centres. Treatment-by-centre interaction is addressed in Section 3.4.

2.1. The stratified Cox model

The stratified Cox model [18] takes the form of (4) with the K baseline hazard functions $\lambda_{0k}(\cdot)$, ($k = 1, \dots, K$) completely unspecified. The hazard functions could have different shapes with some or all the hazard functions unequal. The lack of structure for the centre effects makes the stratified Cox model the most general of the three conditional models. Asymptotic theory for this model is valid for frameworks where n_k remains bounded as the total number of centres K increases, where the cluster size increases with K , and where K is fixed and n_k increases. The ease of computation and the applicability across a wide variety of settings make the stratified Cox model an appealing tool, especially if clustering is of no intrinsic interest or if frailties act non-proportionally on the baseline risk.

However, this approach requires discarding a considerable amount of information from the sample. As in the unstratified Cox model, the baseline hazards λ_{0k} play no role in the estimation. This means that no between-centre comparisons are attempted, and all information about the treatment effect comes from within-centre comparisons. For a fixed sample size, the loss of information increases with the number of centres. At the extreme, in matched pair survival data with one treated and one untreated subject per centre, the estimate of the treatment effect only uses information from pairs in which one member fails while the other remains at risk, and the stratified Cox model reduces to a simple—and relatively insensitive—sign test. This suggests that a method which makes inter- as well as intra-centre covariate comparisons will be more efficient.

2.2. The fixed effects Cox model

Additional structure can be added to the model (4) by assuming the centres act proportionally on the risk of failure. Specifically, it is assumed that $\lambda_{0k}(t) = \xi_k \lambda_0(t)$ where the ξ_k represent centre effects and $\lambda_0(t)$ is a single baseline hazard function. In the previous section

we saw that the gamma frailty model has this form. It can also result from modelling fixed effects for centre. Arbitrarily setting cluster 1 as the reference cluster, we obtain

$$\lambda_{0k}(t) = \lambda_0(t) \exp\{\alpha_k\} \quad (5)$$

for $(k = 1, \dots, K)$, with $\alpha_1 := 0$. This model can be estimated by including indicator variables for centre in an unstratified Cox model. When the number of clusters is small relative to the sample size, this is an attractive approach, especially if the centre effects are of intrinsic interest. However, large numbers of centres relative to the sample size can cause trouble. The asymptotics break down when $K \rightarrow \infty$ as N increases. For instance, in a matched pair experiment with 200 pairs, a fixed effects model would entail estimation of 199 centre parameters. In the context of binary outcomes, this has long been known to result in estimates of the treatment parameter which are strongly biased away from the null.

2.3. The random effects or frailty model

The random effects or frailty model, like the fixed effects model, assumes centre effects act proportionally on the baseline risk of failure. However, instead of treating the centre effects as parameters, the frailty model treats them as a sample from a member of a family of probability distributions. Here, the hazard for subject i in cluster k is

$$\lambda_{ki}(t) = \xi_k \lambda_0(t) \exp(\beta^T Z_{ki}) \quad (6)$$

The frailties (ξ_1, \dots, ξ_K) represent the unmeasured factors, independent of the measured covariates, which affect centre-specific baseline risks, and are assumed to act multiplicatively on the average baseline risk $\lambda_0(t)$.

Under the gamma frailty model introduced in Section 1, the frailties are assumed to be gamma distributed with mean 1 and variance θ_0 . The mean is constrained to 1 to make the average hazard identifiable. The variance parameter θ_0 indexes the degree of between-cluster variability and thus the level of within-cluster dependence. When $\theta_0 = 0$, the frailties are identically equal to 1. In that setting, centre effects are absent and failures are independent within as well as across centres. As θ_0 increases, frailties become more dispersed and dependence increases. The parameter θ_0 can be interpreted as a variance component and $1 + \theta_0$ as an odds ratio for failure between members of the same cluster [19]. Under the gamma frailty model, $\theta_0/(2 + \theta_0)$ is the value of the non-parametric intraclass correlation coefficient, Kendall's τ . Other frailty distributions include the positive stable [20], inverse Gaussian, compound Poisson [17] and log-normal frailties [21–23]. All these models can be represented in form (6) and use a single parameter to index the degree of dependence. Thus these models share the advantage of parsimony; in contrast to the fixed effects model, the number of parameters to describe cluster or centre effects does not grow with the number of clusters. In the gamma frailty model, for example, only θ_0 needs to be estimated in addition to β .

3. APPLICATION TO MULTICENTRE STUDIES

In this section, we contrast the application of stratified, fixed effects and random effects approaches to detecting and estimating centre effects as well as treatment-by-centre interaction.

We review the technical details involved, including methods for fitting various models, and discuss the relative strengths and weaknesses of the methods in different settings.

3.1. Model fitting and estimation of covariate effects

The stratified Cox model is fit by an extension of standard partial likelihood in which separate partial likelihoods specific to centre are combined. This means that the Cox model risk sets are comprised of subjects from one centre only. The stratified log-partial likelihood is

$$\sum_{K=1}^K \sum_{i=1}^{n_k} \Delta_{ki} \left[\beta^T Z_{ki} - \log \left\{ \sum_{j=1}^{n_k} \exp(\beta^T Z_{kj}) Y_{kj}(X_{ki}) \right\} \right]$$

Estimation for the model is straightforward using a Newton–Raphson algorithm in the major statistical software packages. A potential complication is monotone likelihood, which can occur in small samples with heavy censoring. This can be mitigated by a penalized partial likelihood [24] which penalizes extreme values of β .

In the fixed effects model, a reference centre is specified and $K - 1$ indicator functions are defined for the other centres. Estimation is again done by maximum partial likelihood, generally using Newton–Raphson. This is straightforward in principle, but may involve high-dimensional matrix operations beyond the limits of some software packages. Other complications can arise. For instance, if all events times in k th cluster are smaller or larger than event times in all other clusters, then the estimate of α_k is $-\infty$ or $+\infty$, respectively; if this occurs, the log-likelihood converges even as the estimate of α_k diverges. However, once the model is fit, estimates for both treatment and centre effects are available and their interpretations are simple and appealing. In Section 4 we compare the performance of these centre effect estimates with those of the gamma frailty model.

Frailty models cannot generally be estimated by extensions of partial likelihood. Unless the frailty distribution is carefully chosen, the likelihood may not have closed form and likelihood evaluations require numerical or Monte-Carlo approximation. A closed form expression for the likelihood, when available, may still be complex. For instance, the positive stable model likelihood involves a polynomial of order n_k , which is complex for large clusters. Also any likelihood expressions, closed form or not, involve $\Lambda_0(\cdot)$, posing further computational and theoretical challenges. The computational challenges arise because of maximizing over an infinite-dimensional parameter space. The theoretical challenges arise because the parameter space grows with the sample size so standard asymptotic theory is not available to understand the large sample behaviour.

The gamma frailty model is the most extensively studied model because it largely overcomes these obstacles. It yields a closed form likelihood which can be readily maximized. The maximum likelihood estimators have theoretical justification [25–27]. While the asymptotic theory is based on the number of independent clusters tending to infinity, limited simulation studies, including ours, suggest that the small-sample performance is good [15].

Until recently, maximum likelihood estimation of the gamma frailty model involved a conceptually simple but slow EM algorithm [15, 28]. However, Therneau and Grambsch [29] have recently shown that a fast two-step algorithm for maximization of a penalized partial likelihood converges to the EM maximum likelihood estimates. For the log-normal frailty model, use of the penalized likelihood is motivated by the Laplace approximation to the full likelihood, similar to arguments used in the context of generalized linear mixed models [21].

Both methods for fitting frailty models are now implemented in major software packages. SAS macros by J.P. Klein implement the EM algorithm for the gamma and positive stable frailty models, and are available at

<http://www.biostat.mcw.edu/SoftMenu.html>.

The most recent release of S-plus implements the penalized partial likelihood algorithm for the gamma and log-normal frailty models.

3.2. Tests for the presence of centre effects

An important preliminary issue is the detection of the presence of centre effects (or heterogeneity) and then describing them. The stratified Cox model, in which centre-to-centre variability is treated as a nuisance, does not provide a framework for testing or describing dependence or heterogeneity. Thus we contrast the fixed and random effects models, which lead to different tests of independence.

The fixed effects approach parameterizes the centre effects with the parameters $(\alpha_2, \dots, \alpha_K)$ and the null hypothesis of homogeneity is

$$\alpha_2 = \dots = \alpha_K = 0$$

This hypothesis can be tested using a likelihood ratio test on $K - 1$ degrees of freedom. Wald and score tests are also available, the former included in some major statistical packages. At least in simpler settings, these tests are less reliable than the likelihood ratio test.

Multiple tests of homogeneity are also available for random effects models. The most interesting of these are score tests involving the first and second derivatives of the log-likelihood, evaluated under the null hypothesis of homogeneity. In the gamma frailty model, this is a test of $\theta_0 = 0$. The resulting test statistic can be simply computed in some statistical packages using standard output from the Cox model fitting routine. In addition, different frailty distributions lead to the same test, and the validity of the test is not dependent on any particular frailty distribution. Nearly identical score tests of homogeneity were proposed independently by Gray [30, 31] and Commenges and Andersen [32].

The performance of fixed versus random effects tests of independence were compared in extensive simulations studies conducted by Andersen *et al.* [33]. Their simulations showed that in the fixed effects model, the likelihood ratio test is frequently anticonservative. Andersen *et al.* concluded that to give significance levels close to the nominal level, the fixed effects likelihood ratio test requires many more subjects in each centre than are typically available in practice. The score test, however, performed well in a broad range of settings and they recommend its use. A SAS macro by J.P. Klein to implement the Commenges and Andersen test is available at

<http://www.biostat.mcw.edu/SoftMenu.html>.

Likelihood ratio and Wald tests of homogeneity are also available for the random effects models. These tests are computationally more complex and model dependent than the score test because they require fitting a particular frailty model. Fitting the frailty model, however, can provide additional useful information. Even if the null hypothesis of homogeneity is not rejected, the data may be consistent with some level of clustering. By fitting a frailty model, profile likelihood functions, point estimates, and confidence intervals will provide additional

information for inference. However, this information may be to an unknown degree model dependent.

3.3. Estimation of centre effects

For both the fixed effects and frailty models, estimators of the centre effects are available. In the fixed effects model, these are directly estimated. Specifically, the (relative) centre effects are estimated by $\exp(\hat{\alpha}_k)$ ($k=2, \dots, K$). For the frailty model, the individual centre effects can be estimated by the conditional expectation

$$\tilde{\xi}_k = E(\xi | \text{Data for the } k\text{th cluster}) = E(\xi | \mathcal{H}_k(\infty))$$

Nielsen *et al.* [15] showed that this is equal to

$$\tilde{\xi}_k = \frac{\theta_0^{-1} + \sum_{i=1}^{n_k} \Delta_{ki}}{\theta_0^{-1} + \sum_{i=1}^{n_k} \exp(\beta^T Z_{ki}) \Lambda_0(X_{ki})} \quad (7)$$

The estimated frailties, $\hat{\xi}_k$, ($k=1, \dots, K$) can be obtained by substituting the maximum likelihood estimates of θ_0 , β and $\Lambda_0(\cdot)$ into equation (7). Like the familiar best linear unbiased predictor estimates for linear models, the frailty estimates are ‘shrunk’ from the non-parametric Poisson-type estimator

$$\frac{\sum_{i=1}^{n_k} \Delta_{ki}}{\sum_{i=1}^{n_k} \exp(\beta^T Z_{ki}) \Lambda_0(X_{ki})}$$

towards the mean value of the mixing distribution, 1. The degree of shrinkage depends on the amount of data available in the cluster and the value of $\hat{\theta}$. If the cluster is small or there is limited follow-up, the estimated frailty will be shrunk closer to 1. Shrinkage is also greater if $\hat{\theta}$ is near zero. The estimate $\hat{\theta}$ incorporates information from the entire sample in estimating the frailty for each cluster. Analogous empirical Bayes estimators are widely useful in biostatistics [34].

3.4. Treatment-by-centre interaction

It is plausible that factors which differ by centre may affect the magnitude of the treatment effect as well as baseline rates of failure. Such treatment-by-centre interaction may reflect differences in patient characteristics and in implementation of the protocol. Understanding and modelling treatment-by-centre interaction is a challenging aspect of the analysis of multicentre clinical trials. There are a number of open questions in this area, and work for censored time-to-event data lags behind developments for continuous and binary data. Issues include estimation of the (possibly weighted) mean treatment effect, estimation of centre-specific treatment effects, and testing for treatment-by-centre interaction.

A general conditional Cox model for treatment-by-centre interaction has the form

$$\lambda_{ki}(t) = \lambda_{0k}(t) \exp(\tilde{\beta}_k Z_{ki}) \quad (8)$$

where Z_{ki} indicates treatment assignment. The vectors $(\tilde{\beta}_1, \dots, \tilde{\beta}_K)$ describe the centre-specific treatment effects. Covariates could be added to this model in the obvious way. The treatment effect for the k th cluster $\tilde{\beta}_k$ can be written as the sum of a mean treatment effect $\bar{\beta}$ and a mean zero random variable $\omega_k := \tilde{\beta}_k - \bar{\beta}$.

Equation (8) implies two model components. First is a model for the variation in baseline hazards. As we have seen, choices for the baseline hazard model include a common hazard function (no centre effects), stratification by centre, and proportional fixed or random effects. The second component is a model for the centre-to-centre variation in the treatment effect (a form for β_k). For the latter component, fixed and random effects strategies are potentially available.

Treating $(\tilde{\beta}_1, \dots, \tilde{\beta}_K)$ as fixed effects is easily implemented with any of the four baseline hazard specifications. A corresponding $K - 1$ degree of freedom likelihood ratio test can be used to test the null hypothesis $\tilde{\beta}_1 = \dots = \tilde{\beta}_K$. Moreover, the method of Gail and Simon [35] provides a test of the more specific alternative hypothesis of qualitative interaction. However, a fixed effects approach to treatment-by-centre interaction is likely to perform poorly when the number of clusters is large and there are relatively few subjects per cluster. In this context, fixed effects estimates of treatment-by-centre would be expected to demonstrate the same weaknesses as fixed effects estimates of centre-to-centre heterogeneity, including anticonservative tests, low power and point estimates which can be both biased and highly variable.

Frailty models where the interaction is modelled as a random effect may be a potentially useful alternative, in particular for sparse data. These would have the form

$$\lambda_{ki} = \lambda_{0k}(t) \exp(\bar{\beta}Z_{ki} + \omega_k Z_{ki})$$

where ω_k ($k = 1, \dots, K$) could be chosen to follow a log-gamma or normal distribution. Similar models have been proposed for multicentre binary [36] and continuous data [5]. If the centre effects are handled using stratification of the baseline hazards and the interaction is modelled as log-gamma or Gaussian, then the penalized likelihood approach could be used, possibly with only minor modifications to available software. Furthermore, this approach would allow for estimation of the mean treatment effect, a test for treatment-by-centre interaction, and empirical Bayes estimates of the centre-specific treatment effects.

However, a drawback is that random effects models for the interaction cannot be readily implemented in current frailty model implementations, even though such an extension is straightforward. Furthermore, the model becomes considerably more complicated if the centre effects are also treated as random. In that case, possible approaches include an additive gamma frailty model [37], Markov chain Monte Carlo [38] and adaptations of penalized likelihood [39, 22].

As an alternative, permutation-based statistics for treatment-by-centre interaction have been proposed [40] for survival data. The major limitation of this method is that it is not a pure test for treatment-by-centre interaction. It checks for absence of any centre effects (either as a main effect or as a treatment-by-centre interaction). Thus, the test may reject based on centre effects even if there is no treatment-by-centre interaction.

A number of strategies developed in other contexts [41–43] might also be adapted for tests of treatment-by-centre interaction for survival data. For example, Liang and Self [41] developed tests for homogeneity for sparse, stratified 2×2 tables. For censored survival data

analogous test statistics would take the forms

$$\sum_{k=1}^K \{\hat{S}_k(\hat{\beta})\}^2 \quad \text{and} \quad \sum_{k=1}^K \{\hat{S}_k(\hat{\beta})^2 - \hat{I}_k(\hat{\beta})\}$$

where

$$\hat{S}_k(\hat{\beta}) = \sum_{i=1}^{n_k} \Delta_{ki} \left\{ Z_{ki} - \frac{\sum_{j=1}^{n_k} Z_{kj} \exp(\hat{\beta}^T Z_{kj}) Y_{kj}(X_{ki})}{\sum_{l=1}^{n_k} \exp(\hat{\beta}^T Z_{kl}) Y_{kl}(X_{ki})} \right\}$$

Here, $\hat{\beta}$ is the stratified Cox estimate assuming no treatment-by-centre interaction and $\hat{I}_k(\hat{\beta})$ is minus the derivative of $\hat{S}_k(\cdot)$ evaluated at $\hat{\beta}$. The properties of these tests, including their null distribution, should be further explored. However, they are relatively simple and involve quantities which can be easily extracted from a Cox model fit. Liang and Self report good size and power for similar tests.

Cai *et al.* [44, 45] considered the problem of estimating the unweighted mean treatment effect $\bar{\beta}$. They used a model with stratified baseline hazards and treatment modelled by fixed effects, developed an estimating function for $\bar{\beta}$ as well as other regression coefficients, and estimated $\bar{\beta}$ by $K^{-1} \sum_{k=1}^K \hat{\beta}_k$. The estimator is very simple, especially if there are no additional covariates. In that case, the approach is equivalent to estimating treatment effects separately for the K centres and averaging the results. With additional covariates, this approach requires a robust variance estimator to deal with intracluster dependence.

In summary, methods for detecting and estimating treatment-by-centre interaction require further development. Adaptation of developments in other areas, including tools from binary data and meta-analysis [46, 47], may lead to useful techniques.

4. SIMULATION STUDIES

4.1. Estimation of the treatment effect

Simulations were undertaken to compare the performance of the three centre-specific models (stratified and fixed effects Cox models, and the gamma frailty model) with respect to bias, root mean squared error (MSE) and empirical coverage of 95 per cent confidence intervals. Settings varied with respect to total sample size ($N = 100, 400$), number of subjects per centre ($n = 2, 10, 20$) magnitude of intra-centre dependence, and the frailty distribution. We also examined the performance of the marginal Cox model in the setting of no centre effects. No results are presented for the fixed effects model with $N = 400$ and $n = 2$ because the models did not reliably converge.

Intra-centre dependence was simulated from a general conditional model

$$\lambda_{ki}(t) = \lambda_0(t) \exp(\beta_0 Z_{ki} + \varepsilon_k) \quad (9)$$

Note, here we have incorporated frailties on the log scale, a formulation equivalent to (6). We denote the log-frailties by $\varepsilon_k = \log(\zeta_k)$ to avoid confusion. The baseline hazard function $\lambda_0(t)$ followed a Weibull form with shape parameter 1.3 and scale parameter 5.0 in all settings. In addition to a simulation with independent data ($\varepsilon = 0$), we generated frailties from the gamma,

Table I. Comparison of four methods for estimation of treatment effects when failures across centres are independent (no centre effects).

		N = 100			N = 400		
		n = 2	n = 10	n = 20	n = 2	n = 10	n = 20
Marginal Cox							
	Mean β	0.696	0.701	0.694	0.693	0.696	0.693
	Root MSE	0.255	0.254	0.254	0.125	0.125	0.123
	95 per cent CI Cov	0.944	0.922	0.840	0.944	0.937	0.932
Fixed effects							
	Mean β	1.362	0.770	0.730		0.773	0.730
	Root MSE	0.916	0.306	0.275		0.165	0.138
	95 per cent CI Cov	0.574	0.922	0.941		0.889	0.932
Stratif Cox							
	Mean β	0.716	0.703	0.702	0.700	0.698	0.696
	Root MSE	0.374	0.288	0.271	0.176	0.140	0.133
	95 per cent CI Cov	0.954	0.947	0.948	0.959	0.949	0.951
Gamma frailty							
	Mean β	0.720	0.709	0.698	0.702	0.700	0.696
	Root MSE	0.270	0.258	0.256	0.129	0.126	0.123
	95 per cent CI Cov	0.942	0.951	0.954	0.939	0.947	0.950

inverse Gaussian and positive stable densities, in each case using parameter values that give Kendall's τ of 0.50. For the gamma frailty model this requires a model with shape and scale parameter of 2; for the inverse Gaussian and positive stable distributions, parameter values of 5.0 and 0.50 were used, respectively. A simulation was also conducted with a positive stable frailty with parameter 0.75 (Kendall's $\tau = 0.25$) to examine the effect of the level of dependence on the estimates.

In all settings the treatment effect parameter β_0 was $\log(2) = 0.693$. Half of the subjects in each centre were assigned to treatment, reflecting the stratified randomization in most multicentre clinical trials. Censoring was uniform over 0–0.48, 1.7, 11.5, 0.57 and 0.50 for the independence, gamma, inverse Gaussian and positive stable (Kendall's $\tau = 0.50, 0.25$) frailties, respectively. This yielded 30 per cent censoring in each setting.

Table I shows that in the absence of centre effects, the population-averaged Cox model with a robust standard error [14] is clearly the best analytic choice, except when the number of centres is small. It is virtually unbiased in all settings. However with only five centres of size 20, the empirical coverage of the 95 per cent confidence interval is only 0.84. The stratified Cox model and gamma frailty model show good small-sample performance with negligible bias and excellent coverage of 95 per cent confidence intervals in all settings. Because the stratified Cox model uses no intracentre information, it is only approximately $(0.255/0.374)^2 = 46$ per cent efficient relative to the standard Cox model when $n = 2$ and $N = 100$. In contrast, the relative efficiency of the gamma frailty model is good and improves with centre size, ranging from 0.89 to 0.99. The fixed-effects approach performs the most poorly. It is computationally intensive in the setting when $N = 100$ and $n = 2$ and was not

Table II. Comparison of four methods for estimation of treatment effects when centre effects are gamma distributed.

		$N = 100$			$N = 400$		
		$n = 2$	$n = 10$	$n = 20$	$n = 2$	$n = 10$	$n = 20$
Fixed effects							
	Mean β	1.259	0.753	0.728	0.755	0.719	
	Root MSE	0.815	0.299	0.275	0.153	0.134	
	95 per cent CI Cov	0.631	0.926	0.949	0.909	0.940	
Stratif Cox							
	Mean β	0.710	0.704	0.708	0.701	0.696	0.694
	Root MSE	0.379	0.290	0.275	0.176	0.139	0.132
	95 per cent CI Cov	0.948	0.952	0.953	0.959	0.954	0.957
Gamma frailty							
	Mean β	0.683	0.697	0.696	0.695	0.698	0.694
	Root MSE	0.296	0.263	0.267	0.146	0.132	0.131
	95 per cent CI Cov	0.949	0.954	0.948	0.946	0.948	0.942

practical when $N = 400$ and $n = 2$. It shows considerable anti-conservative bias which persists even when the number of subjects per centre is as large as 20. With a small number of centres, the frailty model is only slightly less efficient than the population averaged model, and gives confidence intervals with considerably better coverage properties. In a more general sense, of course, the choice between these alternatives should be based on whether the population-averaged or centre-specific parameters more directly address the scientific research question.

Table II summarizes the results for gamma distributed centre effects. The models are clearly ordered in terms of root MSE, with the gamma frailty model having the lowest root MSE for all values of K and n . The stratified Cox model has the second smallest root MSE, with negligible bias and confidence interval coverage near nominal levels. The fixed effects model performs worst, with badly biased point estimates in the setting of two subjects per centre. With $N = 400$ and $n = 2$, estimates cannot reliably be found.

In Table II, the superior performance of the gamma frailty model fit would be expected, since the data were generated with gamma frailties. However, the choice of a gamma model is based on theoretical and computational tractability, not biological plausibility. Thus it is important to examine the performance of a gamma frailty model when the centre effects have other distributions. Summarized in Tables III and IV are the results for the inverse Gaussian and positive stable frailties with overall dependence similar to the gamma frailty data (Kendall's $\tau = 0.50$). The results are very similar to each other and to the results in Table II. The ordering of the models by root MSE is the same. The gamma frailty fit has the lowest MSE and has confidence interval coverage close to nominal levels. There is negative bias for the gamma frailty fit to both non-gamma models when n is small, but this bias drops off for $n = 10$ or 20. Again, the fixed effects approach performs poorly except when N is large and the number of centres is small. In simulations not reported, we found that the pattern of bias, root MSE and confidence interval coverage holds for a variety of non-gamma frailty distributions.

Table III. Comparison of four methods for estimation of treatment effects when centre effects are inverse gaussian distributed.

		N = 100			N = 400		
		n = 2	n = 10	n = 20	n = 2	n = 10	n = 20
Fixed effects							
	Mean β	1.258	0.752	0.725	0.751	0.718	
	Root MSE	0.810	0.291	0.275	0.152	0.134	
	95 per cent CI Cov	0.631	0.933	0.939	0.912	0.942	
Stratif Cox							
	Mean β	0.710	0.705	0.708	0.698	0.697	0.693
	Root MSE	0.370	0.286	0.276	0.178	0.140	0.133
	95 per cent CI Cov	0.955	0.947	0.950	0.955	0.951	0.955
Gamma frailty							
	Mean β	0.626	0.683	0.695	0.635	0.680	0.687
	Root MSE	0.281	0.263	0.263	0.150	0.127	0.128
	95 per cent CI Cov	0.949	0.949	0.951	0.930	0.952	0.943

Table IV. Comparison of four methods for estimation of treatment effects when centre effects are positive stable distributed with parameter 0.50.

		N = 100			N = 400		
		n = 2	n = 10	n = 20	n = 2	n = 10	n = 20
Fixed effects							
	Mean β	1.249	0.751	0.728	0.751	0.720	
	Root MSE	0.814	0.297	0.283	0.154	0.137	
	95 per cent CI Cov	0.641	0.939	0.942	0.913	0.938	
Stratif Cox							
	Mean β	0.706	0.704	0.709	0.696	0.697	0.696
	Root MSE	0.373	0.289	0.285	0.181	0.143	0.136
	95 per cent CI Cov	0.953	0.954	0.948	0.954	0.953	0.950
Gamma frailty							
	Mean β	0.620	0.679	0.689	0.633	0.679	0.689
	Root MSE	0.291	0.269	0.268	0.153	0.132	0.129
	95 per cent CI Cov	0.944	0.947	0.950	0.933	0.948	0.949

In Table V, we show results for the positive stable model with lower overall dependence (Kendall's $\tau = 0.25$). The lower level of dependence does not appreciably affect the fixed effects or stratified models. There is a slight improvement in the bias of the gamma frailty model when $n = 2$. However, the overall ordering of root MSE is unchanged.

In summary, the fixed effects model is biased in any setting where K is large and n is small. In addition, the fixed effects approach proved to be generally the most computationally demanding method. In contrast, the simulations suggest that both the gamma frailty and

Table V. Comparison of four methods for estimation of treatment effects when centre effects are positive stable distributed with parameter 0.75.

		$N = 100$			$N = 400$		
		$n = 2$	$n = 10$	$n = 20$	$n = 2$	$n = 10$	$n = 20$
Fixed effects							
	Mean β	1.298	0.758	0.730		0.759	0.723
	Root MSE	0.849	0.294	0.274		0.157	0.135
	95 per cent CI Cov	0.603	0.933	0.943		0.906	0.939
Stratified Cox							
	Mean β	0.706	0.704	0.706	0.696	0.697	0.695
	Root MSE	0.366	0.284	0.274	0.178	0.139	0.133
	95 per cent CI Cov	0.953	0.954	0.950	0.956	0.952	0.953
Gamma frailty							
	Mean β	0.653	0.689	0.686	0.664	0.686	0.690
	Root MSE	0.276	0.262	0.260	0.137	0.129	0.128
	95 per cent CI Cov	0.947	0.953	0.952	0.941	0.951	0.945

stratified Cox model estimates are nearly unbiased and have good nominal confidence interval coverage in all settings studied. The efficiency of the stratified approach is dependent on centre size, while the efficiency of the gamma frailty model is competitive in nearly every setting. Even at independence, the frailty model is at least 89 per cent efficient relative to the standard Cox model.

4.2. Estimation of centre effects

We also performed a more limited simulation study of the fixed effect and gamma frailty estimators of the centre effects, in the setting with $N = 100$ total subjects and $n = 2, 10, 20$ and 50 subjects per centre. Centre effects were generated from the positive stable distribution with parameter 0.50. We also studied complete independence with $\varepsilon = 0$. We simulated 5000 data sets and calculated the root MSE of the centre effect estimates in each setting. The frailty estimates are scaled to have mean 1, while the fixed effects estimates are scaled relative to the first centre. To make them comparable, the log-frailties $(\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)$ were scaled to $\tilde{\alpha}_j = \hat{\varepsilon}_j - \hat{\varepsilon}_1$ ($j = 2, \dots, n$).

The results are displayed in Table VI. They show that the gamma-based estimators are appreciably closer to the true centre effects, despite the fact that the centre effects are markedly non-gamma distributed. The fixed effects approach, which makes no assumptions about the distribution of the centre effects, is more precise for smaller centre effects and as the number of subjects per centre increases. However, even in a setting with two moderately large centres, the gamma frailty approach does better. This superior performance exemplifies the well-known result of James and Stein [34,48,49] for empirical Bayes or shrinkage estimators.

Table VI. Comparison of average root MSE of estimates of the magnitude of centre effects (gamma frailty versus fixed effects): using positive stable frailty distribution with parameter 0.50 and $N = 100$.

	Positive stable frailty		Independence	
	Gamma frailty	Fixed effects	Gamma frailty	Fixed effects
2/centre	1.83	3.61	0.0767	2.53
10/centre	0.773	1.41	0.0590	0.568
20/centre	0.576	0.737	0.0473	0.372
50/centre	0.427	0.520	0.0296	0.196

5. DATA EXAMPLE

Raloxifene is a selective estrogen receptor modulator which blocks estrogen in the breast and endometrium and has estrogen-like effects on bone and lipids. Hence, it may have multiple beneficial effects including the prevention of breast cancer and fractures. The randomized clinical trial, multiple outcomes of raloxifene evaluation (MORE) [50], began in 1994. The study enrolled post-menopausal women aged at most 81 years with clinical or radiographic evidence of osteoporosis. The subjects were randomized to either placebo, raloxifene 60 mg 1/day, or raloxifene 60 mg 2/day. In the data analyses, the two raloxifene dose groups were combined. The study randomized 7705 women (5126 to raloxifene and 2576 to placebo) at 174 clinical centres in 25 countries. Centre size ranged from 2 to 743 subjects.

Subjects were followed for multiple outcomes including the development of a new fracture during the follow-up period. Here, we compare the time to ankle fracture between the raloxifene and placebo groups. The study ascertained 62 incident fractures during the 4-year follow-up period: 28 on placebo and 34 on raloxifene. The range of event times was 4.5–42 months, with the bulk of events (50/62, 80 per cent) occurring between 32 and 36 months of follow-up. Ankle fractures per centre varied between 0 and 9/centre, with no events at 136/174 (78 per cent) centres.

The four methods in the paper give slightly different results. The working independence estimator of the log-relative hazard of fracture on the combined raloxifene groups compared to placebo is -0.529 with a robust standard error (SE) of 0.211 ($p = 0.012$). However, the stratified Cox estimate is -0.471 with a SE = 0.271 ($p = 0.08$) and the fixed effects Cox estimate is -0.459 with SE = 0.262 ($p = 0.08$). The difference between these estimates may reflect discrepancies between the marginal and conditional parameters or the poor performance of these methods in a setting with many centres. The frailty based estimate is -0.526 with a SE of 0.257 ($p = 0.04$). Here, the frailty model is particularly useful because it provides a direct estimate of the strength of centre effects. The estimated θ is equal to 0.46. Hence, it appears the centre effects are weak and that the marginal results are more credible than the fixed effects or stratified models, which we expect to perform poorly in this setting. In Figure 1, we plot the profile likelihood for θ , which suggests that while centre effects do not appear strong, the 95 per cent confidence interval for θ ranges from independence to 4.30.

In this study, most centres had no events. The stratified Cox model discards these centres entirely and they contribute no information on the treatment effect. The working-independence approach uses all subjects regardless of centre membership. The fixed effects and frailty models use information from centres with no events but downweight their emphasis. The frailty model

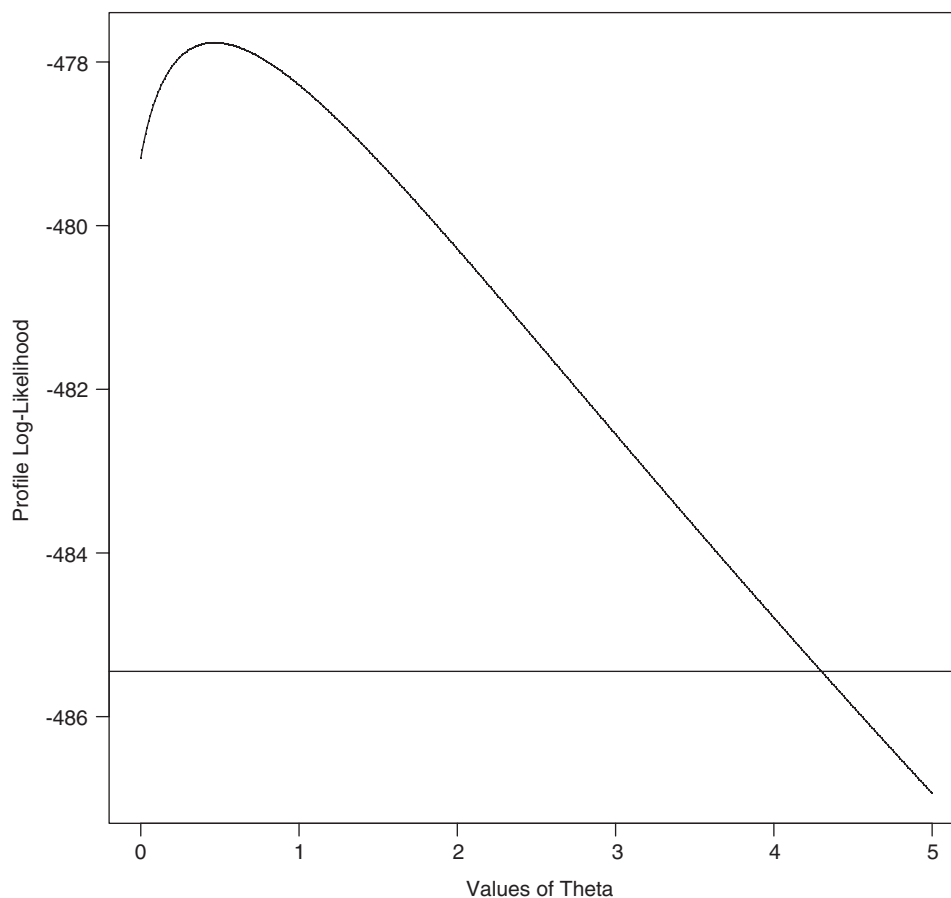


Figure 1. The profile log-likelihood in θ for the MORE ankle fracture data. The intersection of the horizontal line with the likelihood marks the 95 per cent confidence interval in θ .

dampens more gently, basing the degree of cross-cluster comparison on the amount of cluster-level variation. Here, the θ is small and the frailty estimate is almost identical to the working independence estimate. The fixed-effect model behaves like a frailty model with a large value of θ and emphasizes within-centre comparisons more than across-centre comparisons.

The frailty model provides estimates of the frailties or failure rates in the various clinical centres, which are 'smoothed' in inverse proportion to the information available from each. Figure 2 is a scatterplot of log estimated frailties for each centre against the (log) estimated overall fracture rate. The size of the plotting symbol is inversely proportional of the SE. Here, we have excluded centres with no reported fractures. Values to the left of the diagonal line indicate centre-specific estimates of failure rate which are lower than their corresponding frailty estimate, while values to the right of the line have centre-specific fracture rates which are larger. Extreme values of failure rates are shrunk toward the overall failure rate with small centres (indicated by the smaller plotting symbols) pulled more strongly to the mean.

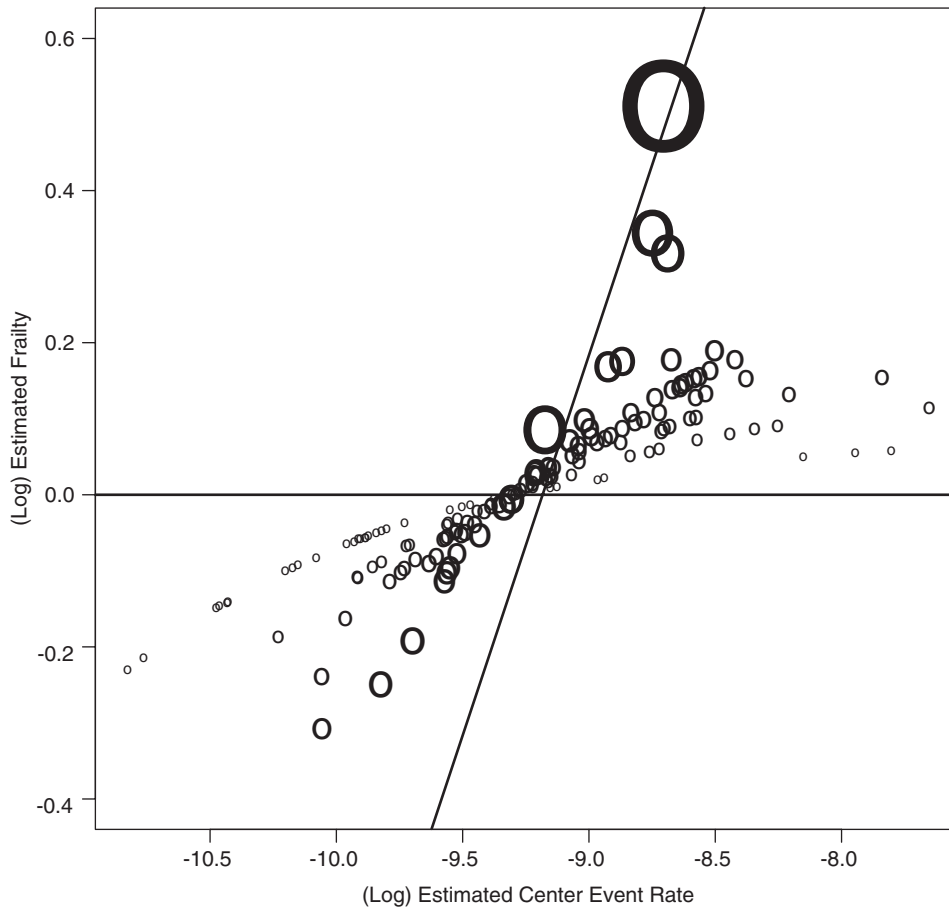


Figure 2. Plot of (log) estimated centre effect estimated based on a gamma frailty model against the estimated centre-specific (log) ankle fracture rate in MORE.

6. DISCUSSION

This paper has surveyed approaches to multicentre clinical trials for censored time to event data. We illustrated both conceptually and through simulations that random effects modelling can be a useful tool in those settings.

In previous reviews, authors have suggested that the primary criterion in selecting between fixed and random effects approaches is interpretation [1, 8, 51]. That is, fixed effects should be used when inference is intended only to apply to the selected centres. If inferences are intended to apply more generally, then it is natural to approach the data using random effects.

We find that in a broad range of settings, the gamma frailty model, an example of the random effects approach, produces estimates with lower MSE than the fixed effects or stratified approaches. In addition, our simulations suggest that misspecification of the frailty distribution may not greatly disturb the performance of the regression coefficient estimators, despite the fact

that different frailty distributions can lead to appreciably different association structures [52] detectable in practical sample sizes [53]. The lack of sensitivity is consistent with analogous theoretical and numerical results for generalized linear models [54]. This suggests that the gamma frailty approach is a good choice in terms of performance even when the random effects interpretation is not appealing.

Our review also shows that approaches to treatment-by-centre interaction are less well-developed for survival data than for binary data and meta-analysis. Adapting approaches from these areas should be straightforward and lead to useful new methods.

Until recently, user-friendly widely available software was not available for fitting the gamma frailty model. However, the penalized likelihood framework has greatly simplified computation of the maximum likelihood estimator. With the recent implementation of the penalized likelihood approach in S-plus, the gamma frailty approach is now a practical reality for data analysis. Some gaps remain, especially in the use of random effects models for treatment-by-centre interaction. The development of computation and theory for such extended frailty models is a useful area for future development.

ACKNOWLEDGEMENTS

This research was supported by grant R01 HL065411. The authors thank the Department of Biostatistics, Johns Hopkins School of Public Health for hosting David Glidden and the MORE investigators for sharing their data. The paper was much improved by the comments and suggestions of three anonymous reviewers.

REFERENCES

1. Agresti A, Hartzel J. Strategies for comparing treatment on a binary response with multicentre data. *Statistics in Medicine* 2000; **19**(8):1115–1139.
2. Localio AR, Berlin JA, Ten Have TR, Kimmel SE. Adjustments for centre in multicentre studies: an overview. *Annals of Internal Medicine* 2001; **135**(2):112–123.
3. Gould AL. Multi-centre trial analysis revisited. *Statistics in Medicine* 1998; **17**(15, 16):1779–1797.
4. Jones B, Teather D, Wang J, Lewis JA. A comparison of various estimators of a treatment difference for a multi-centre clinical trial. *Statistics in Medicine* 1998; **17**(15-16):1767–1777.
5. Senn S. Some controversies in planning and analysing multi-centre trials. *Statistics In Medicine* 1998; **17**(15-16):1753–1765.
6. Beitley PJ, Landis JR. A mixed-effects model for categorical data (correction 42: 1009). *Biometrics* 1985; **41**(4):991–1000.
7. Fleiss JL. Analysis of data from multiclinic trials. *Controlled Clinical Trials* 1986; **7**(4):267–275.
8. Grizzle JE. Letter to the editor. *Controlled Clinical Trials* 1987; **8**(4):392–393.
9. Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review* 1991; **59**(1):25–35.
10. Lindsey JK, Lambert P. On the appropriateness of marginal models for repeated measurements in clinical trials. *Statistics in Medicine* 1998; **17**(4):447–469.
11. Wei LJ, Glidden DV. An overview of statistical methods for multiple failure time data in clinical trials. *Statistics in Medicine* 1997; **16**(8):833–839.
12. Lin DY. Cox regression analysis of multivariate failure time data—the marginal approach. *Statistics in Medicine* 1994; **13**(21):2233–2247.
13. 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**(408):1065–1079.
14. Lee EW, Wei LJ, Amato DA. Cox-type regression analysis for large number of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, Klein JP, Goel PK (eds). Kluwer: Dordrecht, 1992.
15. Nielsen GG, Gill RD, Andersen PK, Sørensen TIA. A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics* 1992; **19**(1):25–43.
16. Aalen OO. Heterogeneity in survival analysis. *Statistics in Medicine* 1988; **7**(1):1109–1120.

17. Henderson R, Oman P. Effect of frailty on marginal regression estimates in survival analysis. *Journal of the Royal Statistical Society, Series B* 1999; **61**(2):367–379.
18. Holt JD, Prentice RL. Survival analyses in twin studies and matched pair experiments. *Biometrika* 1974; **61**(1):17–30.
19. Clayton DG. Model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 1978; **65**(1):141–151.
20. Hougaard P. Survival models for heterogeneous populations derived from stable distributions. *Biometrika* 1986; **73**(3):387–396.
21. McGilchrist CA. REML estimation for survival models with frailty. *Biometrics* 1993; **49**(1):221–225.
22. Ripatti S, Palmgren J. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* 2000; **56**(4):1016–1022.
23. Vaida F, Xu RH. Proportional hazards model with random effects. *Statistics in Medicine* 2000; **19**(24):3309–3324.
24. Heinze G, Schemper L. A solution to the problem of monotone likelihood in Cox regression. *Biometrics* 2001; **57**(1):114–119.
25. Murphy SA. Consistency in a proportional hazards model incorporating a random effect. *Annals of Statistics* 1994; **22**(2):712–731.
26. Murphy SA. Asymptotic theory for the frailty model. *Annals of Statistics* 1995; **23**(1):182–198.
27. Parner E. Asymptotic theory for the correlated gamma-frailty model. *Annals of Statistics* 1998; **26**(1):183–214.
28. Klein JP. Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* 1992; **48**(3):795–806.
29. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. Springer: New York, 2000.
30. Gray RJ. Tests for variation over groups in survival data. *Journal of the American Statistical Association* 1995; **90**(429):198–203.
31. Gray RJ. On tests for group variation with a small to moderate number of groups. *Lifetime Data Analysis* 1998; **4**(2):139–148.
32. Commenges D, Andersen PK. Score test of homogeneity for survival data. *Lifetime Data Analysis* 1995; **1**(2):145–160.
33. Andersen PK, Klein JP, Zhang MJ. Testing for centre effects in multi-centre survival studies: a Monte Carlo comparison of fixed and random effects tests. *Statistics in Medicine* 1999; **18**(12):1489–1500.
34. Louis TA. Using empirical Bayes methods in biopharmaceutical research. *Statistics in Medicine* 1991; **10**(6):811–829.
35. Gail M, Simon R. Testing for qualitative interaction between treatment effects and patient subsets. *Biometrics* 1985; **41**(2):361–372.
36. Raghunathan TE, Li Y. Analysis of binary data from a multicentre clinical-trial. *Biometrika* 1993; **80**(1):127–139.
37. Petersen JH. An additive frailty model for correlated life times. *Biometrics* 1998; **54**(2):646–661.
38. Gray RJ. A Bayesian analysis of institutional effects in a multicentre cancer clinical-trial. *Biometrics* 1994; **50**(1):244–253.
39. Yamaguchi T, Ohashi Y. Investigating centre effects in a multicentre clinical trial of superficial bladder cancer. *Statistics in Medicine* 1999; **18**(3):1961–1971.
40. Potthoff RF, Peterson BL, George SL. Detecting treatment-by-centre interaction in multicentre clinical trials. *Statistics in Medicine* 2001; **20**(2):193–213.
41. Liang K-Y, Self SG. Testing for homogeneity of odds ratio when the data are sparse. *Biometrika* 1985; **72**(2):353–358.
42. Zelen M. The analysis of several 2×2 contingency tables. *Biometrika* 1971; **58**(1):129–137.
43. Lipsitz SR, Dear KBG, Laird NM, Molenberghs G. Tests for homogeneity of the risk difference when data are sparse. *Biometrics* 1998; **54**(1):148–160.
44. Cai JW, Sen PK, Zhou HB. A random effects model for multivariate failure time data from multicentre clinical trials. *Biometrics* 1999; **55**(1):182–189.
45. Cai JW, Zhou HB, Davis CE. Estimating the mean hazard ratio parameters for clustered survival data with random clusters. *Statistics in Medicine* 1997; **16**(17):2009–2020.
46. DerSimonian R, Laird N. Metaanalysis in clinical trials. *Controlled Clinical Trials* 1986; **7**(3):177–188.
47. Norman SLT. Meta analysis: formulating, evaluating, combining and reporting. *Statistics in Medicine* 1999; **18**(3):321–359.
48. James W, Stein C. Estimation with quadratic loss. *Proceeding of the 4th Berkeley Symposium on Mathematical Statistics and Probabilities*, vol. 1. University of California Press: California, 1961; 361–379.
49. Morris C. Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association* 1983; **78**(381):47–59.
50. Cummings SR, Eckert S, Krueger KA, Grady D, Powles TJ, Cauley JA, Norton L, Nickelsen T, Bjarnason NH, Morrow M, Lippman ME, Black D, Glusman JE, Costa A, Jordan VC. The effect of raloxifene on risk of breast cancer in postmenopausal women—results from the MORE randomized trial. *Journal of the American Medical Association* 1999; **281**(23):2189–2197.

51. Hougaard P. *Analysis of Multivariate Survival Data*. Springer: New York, 2000.
52. Shih JH, Louis TA. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* 1995; **51**(4):1384–1399.
53. Glidden DV. Checking the adequacy of the gamma frailty model for multivariate failure times. *Biometrika* 1999; **86**(2):381–393.
54. Neuhaus JM, Hauck WW, Kalbfleisch JD. The effects of mixture distribution misspecification when fitting mixed-effects logistic-models. *Biometrika* 1992; **79**(4):755–762.