STATISTICS IN MEDICINE
Statist. Med. 2004; **23**:2729–2744 (DOI: 10.1002/sim.1840)

Modelling the distribution of ischaemic stroke-specific survival time using an EM-based mixture approach with random effects adjustment

S. K. Ng[1,*,†], G. J. McLachlan[1], Kelvin K. W. Yau[2] and Andy H. Lee[3]

[1]*Department of Mathematics, University of Queensland, Brisbane, QLD 4072, Australia*
[2]*Department of Management Sciences, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong*
[3]*Department of Epidemiology and Biostatistics, Curtin University of Technology, GPO Box U 1987, Perth, WA 6845, Australia*

SUMMARY

A two-component survival mixture model is proposed to analyse a set of ischaemic stroke-specific mortality data. The survival experience of stroke patients after index stroke may be described by a subpopulation of patients in the acute condition and another subpopulation of patients in the chronic phase. To adjust for the inherent correlation of observations due to random hospital effects, a mixture model of two survival functions with random effects is formulated. Assuming a Weibull hazard in both components, an EM algorithm is developed for the estimation of fixed effect parameters and variance components. A simulation study is conducted to assess the performance of the two-component survival mixture model estimators. Simulation results confirm the applicability of the proposed model in a small sample setting. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: EM algorithm; GLMM; stroke-specific death; survival mixture; Weibull distribution

1. INTRODUCTION

Mixture models have been widely used to model failure-time data in a variety of situations [1–5]. As a flexible way of modelling data, the mixture approach is directly applicable in situations where the adoption of a single parametric family for the distribution of failure time is inadequate. For example, following open-heart surgery for heart valve replacement, the risk of death can be characterized by three merging phases [6]: an early phase immediately

*Correspondence to: Angus S. K. Ng, Department of Mathematics, University of Queensland, Brisbane, QLD 4072, Australia.
†E-mail: skn@maths.uq.edu.au

Contract/grant sponsor: Australian Research Council
Contract/grant sponsor: Research Grants Council of Hong Kong
Contract/grant sponsor: National Health and Medical Research Council of Australia

Received June 2003
Accepted December 2003

following surgery in which the risk of dying is relatively high, a middle phase of constant risk, and finally a late phase in which the risk of death starts to increase as the patient ages. These phases overlap each other in time and thus cannot be modelled satisfactorily by attempting to fit a separate parametric model to each discrete time period; see for example References [4, 7] who specified mixture models for the survival function with three components corresponding to the three phases of death.

In this paper, we focus on the modelling of survival experience of ischaemic stroke patients after index stroke. In Australia, there are about 50 000 stroke and transient ischaemic attacks every year, with ischaemic stroke accounting for about 60 per cent of all stroke events [8]. Of all stroke events each year, 70 per cent are first-ever strokes. Ischaemic stroke is diagnosed by the ICD9-CM [9] and the ICD10 [10] codes and is usually accompanied by acute and chronic phases [11]. In Western Australia, the annual incidence of stroke was estimated at 178 per 100 000. Approximately, around 70 per cent of acute stroke events result in hospitalization and 15 per cent of hospitalized ischaemic stroke patients die within one month [12]. By adopting a mixture model with two components corresponding to the acute and chronic phases, the survival function of time to death T is modelled as

$$S(t;x) = pS_1(t;x) + (1 - p)S_2(t;x) \tag{1}$$

where p denotes the proportion of patients belonging to the acute phase and $S_1(t;x)$ and $S_2(t;x)$ are the conditional survival functions given that the patient is within the acute and chronic phases, respectively. Here x is a vector of covariates associated with each patient. With this concomitant information, effects of demographic characteristics and co-morbidities on the survival of patients in the acute and chronic phases can then be determined.

One issue concerns the heterogeneity due to random hospital effects arising from the clustering of patients within the same hospital. Although clinical practice should conform to a designated standard if guidelines on treatment and care have strictly been followed, it is inevitable that some inherent differences will still exist among hospitals. Thus, it is expected that observations from patients admitted to the same hospital at index stroke will be correlated. In this paper, we assume that such inherent hospital effects are random and shared among patients admitted to the same hospital. We extend the survival model (1) to adjust for random hospital effects based on the generalized linear mixed model (GLMM) method of McGilchrist [13]. The method commences from the best linear unbiased predictor (BLUP) and extends to obtain approximate residual maximum likelihood (REML) estimators for the variance component [14]. In addition, we propose the use of the EM algorithm [15] to obtain the BLUP estimate. This EM-based mixture approach has a number of desirable properties, including its simplicity of implementation and reliable global convergence [16, Section 1.7]. Moreover, it does not require the calculation of second derivatives of a conditional likelihood as required with some Newton–Raphson approaches [17] and it does not require Monte Carlo approximation as required with some computationally intensive methods [18, 19].

This paper is organized as follows. In Section 2, we describe the ischaemic stroke-specific mortality data. The two-component survival mixture model with random effects is presented in Section 3. In Section 4, we describe how the EM algorithm is adopted to obtain the BLUP estimate. The analysis of the ischaemic stroke data is presented in Section 5, and in Section 6, we present a simulation study to assess the performance of the proposed EM-based mixture approach in a small sample setting. Finally, some concluding remarks are provided in Section 7.

2. ISCHAEMIC STROKE-SPECIFIC MORTALITY DATA

In this study, hospital separation records corresponding to all ischaemic stroke events from January 1996 to December 1998 were retrieved from the Western Australia Hospital Morbidity Data System. Only those individuals who had an initial hospitalization for ischaemic stroke during the first six months of 1996 constituted our study cohort and this initial hospitalization was defined as the index stroke. Death data were obtained from the Australian Bureau of Statistics mortality database. Record linkage was used to extract the medical history for each patient from the diagnostic information recorded on hospital separation summaries. A unique patient identifier attached by record linkage to all hospital separation records for the same individual facilitated the retrieval of a patient's medical history. A data set was then compiled containing one record for each person in the cohort discharged from hospital after suffering an index stroke. Each record contained information describing the patient's demographics at the time of the index stroke. Demographic information included age at admission (AGE), gender (SEX), and indigenous status (ABORIGINAL). Clinical information such as a history of diabetes (DIABETES) and atrial fibrillation (AF) was also included. The data set consists of 557 patients from 56 hospitals.

When death occurred, the date and cause of death was included on the patient's record. Since stroke is generally an old-age disease, death occurrence may be due to stroke or other causes. In this study, the effect of risk factors on the stroke-specific death hazard is investigated with death due to other causes treated as censored observations. The proportion of censored observations is 72.9 per cent.

As described in Section 1, the main aim of the study is to identify and assess risk factors affecting the survival of ischaemic stroke patients in the acute and chronic phases, respectively, with an adjustment for random hospital variation. The outcomes of the analysis will thus assist clinicians to rationalize their medical practice, as well as hospital management in terms of budgetary allocation and rehabilitation planning. By comparing the significant factors between the two phases, appropriate strategy and policies can be prescribed to improve the efficiency of service delivery and manage the cost of acute care. In addition, the analysis provides information on inter-hospital variation. As a result, the relative efficiency of hospitals may be evaluated based on the predicted random effects. From another clinical perspective, it may also be important to assess the effect of risk factors on the proportion of ischaemic stroke patients in the acute and chronic phases, say, via a logistic function [20] with possibly an adjustment for random hospital variation. Adjusted odds ratio can then be calculated to estimate the relative risk of belonging to the acute phase or the chronic phase.

3. TWO-COMPONENT SURVIVAL MIXTURE MODEL WITH RANDOM EFFECTS

Let T_{ij} denote the observable failure/censoring time of the jth individual within the ith hospital; let M denote the number of hospitals and n_i the number of observations in the ith hospital, the total number of observations is therefore $N = \sum_{i=1}^{M} n_i$. Let x_{ij} be a vector of covariates associated with T_{ij}. The survival function of T is modelled by a two-component mixture model as

$$S(t_{ij}; x_{ij}) = pS_1(t_{ij}; x_{ij}) + (1 - p)S_2(t_{ij}; x_{ij}) \quad (i = 1, \ldots, M, \ j = 1, \ldots, n_i) \tag{2}$$

where p denotes the proportion of patients belonging to the first component, the subpopulation of patients in an acute condition, and $S_g(t_{ij};x_{ij})$ is the conditional survival function of the gth component ($g=1,2$). Under Cox's proportional hazards model [21], the conditional hazard function for the gth component ($g=1,2$) is given by,

$$h_g(t_{ij};x_{ij}) = h_{g0}(t_{ij})\exp(\eta_g(x_{ij})) \quad (g=1,2) \tag{3}$$

where $h_{g0}(t_{ij})$ is the baseline hazard function and $\eta_g(x_{ij})$ is the linear predictor relating to the covariate x_{ij}. In this paper, the commonly used Weibull distribution is assumed for $h_{g0}(t_{ij})$ because it is flexible as either a monotonic increasing, constant, or monotonic decreasing baseline hazard. That is,

$$h_{g0}(t_{ij}) = \lambda_g \alpha_g t_{ij}^{\alpha_g - 1} \quad (i=1,\ldots,M,\ \ j=1,\ldots,n_i,\ \ g=1,2) \tag{4}$$

where $\lambda_g, \alpha_g > 0$ are unknown parameters. Alternatively, other general lifetime distributions may also be specified for $h_{g0}(t_{ij})$. For example, Larson and Dinse [22] assumed the baseline hazard to be piecewise constant, Gordon [23] modelled $h_{g0}(t_{ij})$ by the Gompertz distribution, while Peng *et al.* [24] considered a continuous baseline hazard in the generalized F distribution family.

As pointed out in Section 1, observations collected from the same hospital are often correlated. The dependence of clustered data (patients nested within hospitals) can lead to spurious associations and misleading inferences. In this paper, the GLMM method [13, 14] is adopted to adjust for the random hospital effects. An unobserved random term is introduced multiplicatively in each conditional hazard function to explain the variability shared by patients within a hospital. With reference to (3), the random effect U_{gi} of the ith hospital on the gth component hazard function can be accommodated through the linear predictor, via

$$\eta_g(x_{ij}) = x_{ij}^{\mathrm{T}}\beta_g + U_{gi} \quad (i=1,\ldots,M,\ \ j=1,\ldots,n_i,\ \ g=1,2) \tag{5}$$

where β_g is the vector of regression coefficients and the superscript T denotes vector transpose. The unobservable random hospital effects U_{gi} ($i=1,\ldots,M$) are taken to be i.i.d. $\mathrm{N}(0,\theta_g)$. A positive value of U_{gi} indicates that patients in the ith hospital will experience a higher risk of failure if they belong to the gth subpopulation ($g=1,2$). Thus if θ_g differs from zero, it implies a significant difference in the survival for patients of the gth subpopulation between the participating hospitals. Under the formulation based on (3)–(5), the vector of unknown parameters becomes

$$\psi = (p, \beta_1^{\mathrm{T}}, \beta_2^{\mathrm{T}}, u_1^{\mathrm{T}}, u_2^{\mathrm{T}}, \lambda_1, \lambda_2, \alpha_1, \alpha_2)^{\mathrm{T}}$$

where $u_1^{\mathrm{T}} = [U_{11}, U_{12}, \ldots, U_{1M}]$ and $u_2^{\mathrm{T}} = [U_{21}, U_{22}, \ldots, U_{2M}]$. The GLMM method commences with the BLUP at the initial step and proceeds to obtain approximate REML estimators of the parameters θ_g in the variance component [14, 25, 26]. For a given initial value of θ_g ($g=1,2$), the BLUP estimator of ψ maximizes $l = l_1 + l_2$, where

$$\begin{aligned} l_1 &= \sum_{i=1}^{M}\sum_{j=1}^{n_i}[D_{ij}\ \log\ f(t_{ij};x_{ij}) + (1-D_{ij})\ \log\ S(t_{ij};x_{ij})] \\ l_2 &= -(1/2)[M\ \log(2\pi\theta_1) + (1/\theta_1)u_1^{\mathrm{T}}u_1] + (-1/2)[M\ \log(2\pi\theta_2) + (1/\theta_2)u_2^{\mathrm{T}}u_2] \end{aligned} \tag{6}$$

Here $D_{ij}=1$ and $D_{ij}=0$ indicate a failure and a censored observation, respectively, and

$$f(t_{ij};x_{ij})=pf_1(t_{ij};x_{ij})+(1-p)f_2(t_{ij};x_{ij})$$

is the probability density function of T based on (2), where $f_g(t_{ij};x_{ij})$ is the conditional probability density function given that the patient belongs to the gth component ($g=1,2$). In (6), l_1 is the log likelihood based on the failure and censored times conditional on u_1 and u_2, and l_2 is the logarithm of the joint probability density function of u_1 and u_2, with u_1 and u_2 taken to be independent. The BLUP estimate of ψ is obtained as a solution of the equation $\partial l/\partial\psi=0$, which can be solved via the EM algorithm as presented in Section 4 below.

The approximate REML estimates of the variance components θ_1 and θ_2 are obtained by maximizing the restricted log likelihood function, which is the log likelihood obtained from a specified set of linearly independent error contrasts [27]. The details are provided in Appendix A.

4. EM ALGORITHM FOR ESTIMATION OF BLUP ESTIMATOR

The EM algorithm is a broadly applicable approach to the iterative computation of maximum likelihood (ML) estimates, useful in a variety of incomplete-data problems [16, 28]. In order to pose the estimation procedure as an incomplete-data problem, an unobservable random vector Z is introduced. For each observation t_{ij}, there is a corresponding two-dimensional indicator variable z_{ij}. For example, $z_{ij}=(1,0)^{\mathrm{T}}$ indicates that t_{ij} belongs to the first component. The random effects u_1 and u_2 are not introduced as incomplete variables in the complete-data framework, as this can slow down the EM algorithm considerably, especially when the variance components are relatively small [29]. Furthermore, when treating the random effects as incomplete variables, an analytical form of the log likelihood expression in the E-step involves high-dimensional integration, which is difficult to perform. On the $(k+1)$th iteration, the E-step of the EM algorithm involves the calculation of the Q-function, which is the expectation of the complete-data log-likelihood conditional on the current estimate of the parameter and the observed data. In particular, the Q-function can be decomposed as

$$Q(\psi,\psi^{(k)})=Q_p^{(k)}+Q_{\xi_1}^{(k)}+Q_{\xi_2}^{(k)}$$

with respect to the parameters p, $\xi_1=(\beta_1^{\mathrm{T}},u_1^{\mathrm{T}},\lambda_1,\alpha_1)^{\mathrm{T}}$, and $\xi_2=(\beta_2^{\mathrm{T}},u_2^{\mathrm{T}},\lambda_2,\alpha_2)^{\mathrm{T}}$, respectively. It implies that the estimates of p, ξ_1, and ξ_2 can be updated separately in the M-step of the EM algorithm by maximizing $Q_p^{(k)}$, $Q_{\xi_1}^{(k)}$, and $Q_{\xi_2}^{(k)}$, respectively. Hence, the proposed EM-based mixture approach possesses a number of desirable properties, including its simplicity of implementation and reliable global convergence. Moreover, it does not require the calculation of the second derivatives of a conditional likelihood as required with some Newton–Raphson approaches [17]. The latter methods usually involve also the calculations of inverse of matrices with large dimensions.

In Appendix B, we provide the detailed descriptions of the E- and M-steps of the EM algorithm for estimation of the BLUP estimator. We also describe the estimation procedure with the EM-based mixture approach.

Table I. Preliminary analysis.

(a) *Results of fitting a single Weibull and a two-component Weibull mixture without the random hospital effect adjustment*

	Regression coefficient (S.E.)		
		Two-component Weibull mixture	
Covariate	A single Weibull	1st component	2nd component
Constant ($\log \lambda_g$)	−9.533 (1.03)*	−3.365 (1.22)*	−12.620 (1.44)*
AGE (β_{1g})	0.082 (0.01)*	0.015 (0.02)	0.109 (0.02)*
SEX (β_{2g})	0.100 (0.20)	−0.934 (0.11)*	0.085 (0.29)
ABORIGINAL (β_{3g})	0.385 (0.72)	0.501 (1.41)	0.358 (0.68)
DIABETES (β_{4g})	0.179 (0.25)	0.198 (0.88)	0.445 (0.36)
AF (β_{5g})	−0.223 (0.19)	−1.623 (1.12)	−0.384 (0.37)

(b) *Model selection*

Number of components (g)	Log likelihood	AIC†	BIC‡	Worth indices§
1	−1101.87	2217.74	2248.00	—
2	−1066.69	2163.38	2228.22¶	(0.63, 0.37)¶
3	−1057.74	2161.48¶	2260.90	(0.47, 0.44, 0.09)

*Significant at 5 per cent level.
†Akaike's information criterion.
‡Bayesian information criterion.
§The number of components is chosen to be minimum value of g for which the sum of their worth indices exceeds 0.8.
¶The number of components selected by each model selection method.

5. ANALYSIS OF THE ISCHAEMIC STROKE DATA

The ischaemic stroke data described in Section 2 is initially fitted using, respectively, a single Weibull regression model and a two-component Weibull mixture regression model. Both models do not adjust for the random hospital effects. The results of this preliminary analysis are presented in Table I. It can be seen that, with a single Weibull regression model, the patient's age at index stroke is the only significant risk factor. By using a two-component Weibull mixture, it is found that the patient's age only has significant effect on patients' survival in the chronic phase. In addition, the gender of patient has a significant effect on the survival of patients in the acute phase. This risk factor has not been identified using a single Weibull model. With a single Weibull model, a decreasing hazard ($\alpha = 0.311$) is determined. In contrast, with a two-component Weibull mixture, an increasing hazard ($\alpha_1 = 1.584$) and a decreasing hazard ($\alpha_2 = 0.395$) are determined for the acute and chronic phases, respectively. The number of components may be chosen based on some information criteria in model selection [28, Chapter 6], [30] or the technique described in Reference [31], where

Table II. Results of fitting a two-component survival mixture model with random hospital effects to the ischaemic stroke mortality data.

	1st component	2nd component
p_g	0.100	0.900
α_g	1.707	0.419
θ_g	0.238 (0.30)	0.947 (0.44)*
Covariate	Coefficient (S.E.)	
Constant ($\log \lambda_g$)	−3.803 (1.51)*	−13.079 (1.47)*
AGE (β_{1g})	0.020 (0.02)	0.114 (0.02)*
SEX (β_{2g})	−1.478 (0.40)*	0.060 (0.29)
ABORIGINAL (β_{3g})	0.084 (1.76)	1.012 (0.82)
DIABETES (β_{4g})	0.543 (0.72)	0.512 (0.36)
AF (β_{5g})	−1.857 (0.99)	−0.268 (0.32)

*Significant at 5 per cent level.

the 'worth index' of each component was calculated to select the number of experts in the mixture of proportional hazards model. Result of applying these model selection methods is included in Table I. Based on this result, we choose a two-component mixture of survival model.

We then apply the EM-based two-component Weibull mixture approach with random hospital effects adjustment. The results are given in Table II. It can be seen that about 10 per cent of patients are identified as being in the acute phase. An increasing hazard ($\alpha_1 = 1.707$) is determined. The remaining 90 per cent of patients appear to be in the chronic phase and a decreasing hazard ($\alpha_2 = 0.419$) is observed for this second component. The gender of patient has a significant impact on the hazard of the first component, suggesting that male stroke patients have a lower risk of stroke-specific death during the acute phase. For the chronic phase, age is found to be the only significant risk factor implying that older-age patients experience a higher risk of ischaemic stroke-specific mortality.

The survival function of the first component is plotted against time by gender in Figure 1, while other covariates are set at their median values. In Figure 2, the estimate of the survival function of the second component is plotted against time for various levels of the admission age. It can be seen that increased age of patient is related to higher death rate in the chronic phase.

By allowing for random hospital effects in model (2), significant hospital variation is detected in the second component (Table II). It implies that heterogeneity in survival during the chronic phase is partially due to the differences among hospitals. The identification of risk factors, after accounting for the random hospital variation, provides useful information on how a patient's survival in the chronic phase is affected. In addition, the quantification of inter-hospital variation as measured by the predicted random effects provides additional insights to assess the variation among hospitals on stroke specific deaths. Predicted random hospital effects for the first component (acute phase) and the second component (chronic phase) are displayed in Figure 3. An inspection of hospital identity reveals no notable difference between the three tertiary hospitals and other hospitals in terms of the predicted hospital effects.

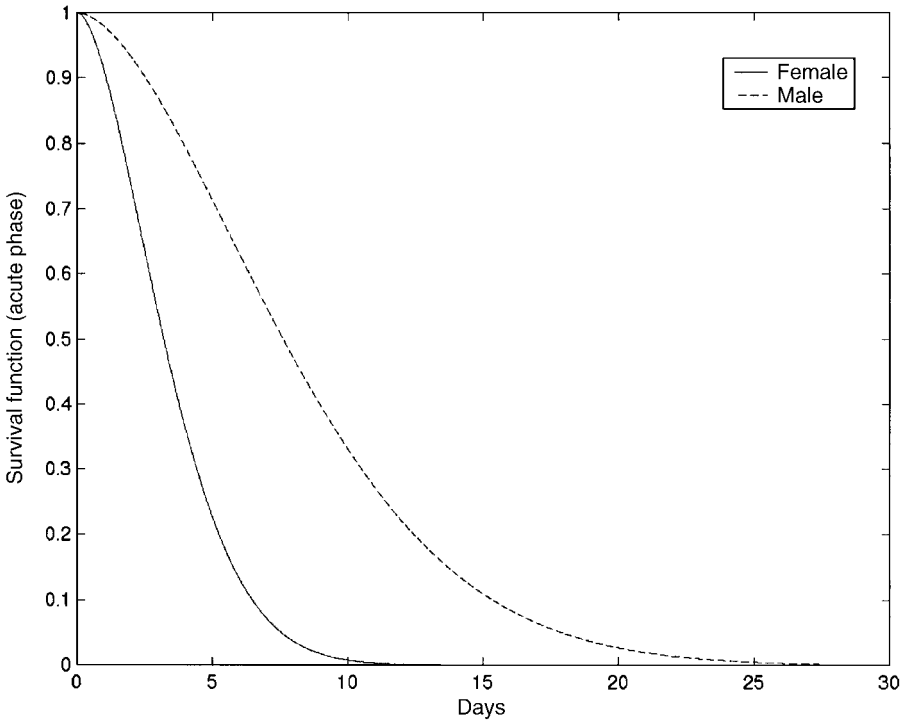

Figure 1. Survival function plot of the first component for time to stroke-specific death by gender.

6. SIMULATION STUDY

A simulation study is conducted based on a multi-centre clinical trial data structure. It is assumed that there are 20 hospitals, and within each hospital there are 25 patients. A continuous covariate variable x_{ij} $(i=1,\ldots,20,\ j=1,\ldots,25)$ is generated independently from $\mathrm{N}(0,1)$. Realizations of Z are simulated in which an individual has a probability of p to be from the first component, $z_{ij}=(1,0)^{\mathrm{T}}$, and has a probability of $(1-p)$ to be from the second component, $z_{ij}=(0,1)^{\mathrm{T}}$. Suppose an individual belongs to the first component, a sample failure time is then generated from the conditional probability density function $f_1(t;x_{ij})=h_{10}(t)\exp(\eta_1(x_{ij}))S_{10}(t)^{\exp\ \eta_1(x_{ij})}$, with U_{1i} generated from $\mathrm{N}(0,\theta_1)$. Similarly, for an individual belonging to the second component, a sample failure time is generated from the conditional probability density function $f_2(t;x_{ij})=h_{20}(t)\exp(\eta_2(x_{ij}))S_{20}(t)^{\exp\ \eta_2(x_{ij})}$, with U_{2i} generated from $\mathrm{N}(0,\theta_2)$. If the generated failure time is greater than a constant censoring time, C, it is taken as censored at time C. In the simulation study, we assume the baseline hazards for both components to follow a Weibull distribution (3), with different known parameters λ_g and α_g $(g=1,2)$. We fix $\lambda_1=0.05$, $\alpha_1=1.5$, $\beta_1=0.5$, $\lambda_2=0.01$, $\alpha_2=0.5$, $\beta_2=-0.5$, and $C=1000$ in all the settings, and consider three different sets of parameter value of p (0.1, 0.3, and 0.5). There are 500 replications in each setting considered.

The performances of the estimators are assessed in terms of their biases and standard errors. The simulation results are given in Table III, where SE_1 and SE_2 denote the average

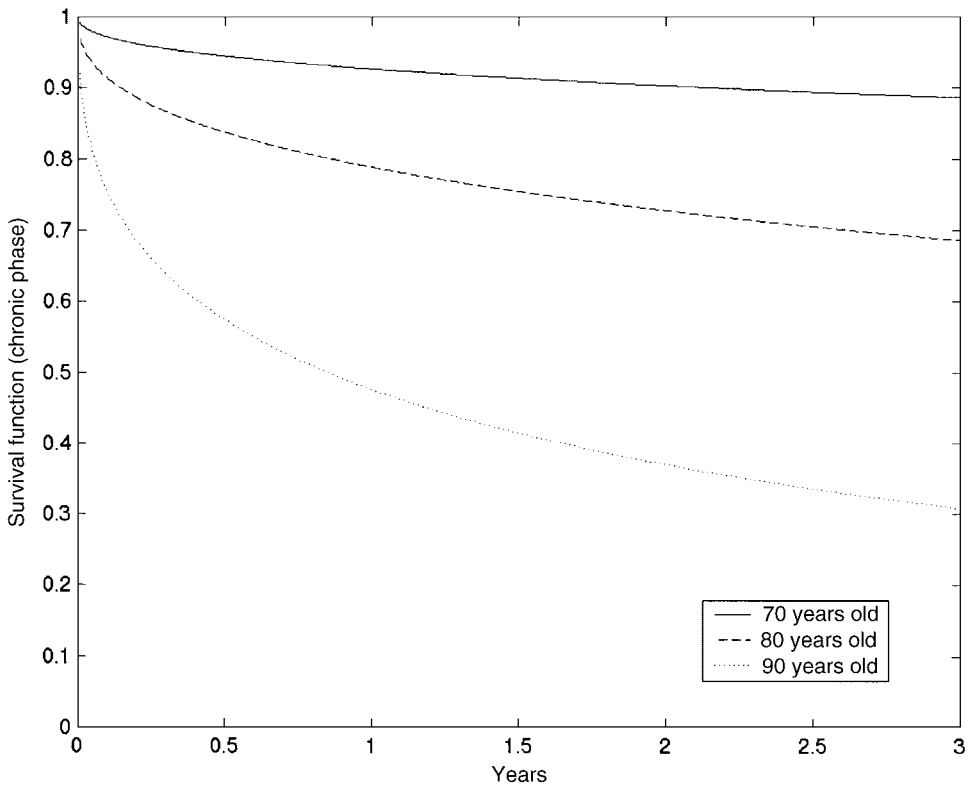

Figure 2. Survival function plot of the second component for time to stroke-specific death by three levels of age at index stroke admission.

of the standard error of the estimates and of the sample standard error of the estimates, respectively, over the 500 replications. From Table III, no appreciable bias is observed in all simulation settings, confirming the applicability of the proposed model in small sample situations. As anticipated, the estimate of p is slightly biased when its true parameter value is close to the boundary ($p = 0.1$) of the parameter space. A comparison of SE_1 and SE_2 provides information on whether the estimated standard error of the procedure is overestimated or underestimated. In general, good agreement between SE_1 and SE_2 for all the parameters is observed. When $p = 0.1$, the estimated standard errors of regression coefficients and variance component estimates appear to be slightly underestimated for the first component. Thus, caution should be exercised in interpreting the significance levels attached to these estimates when the estimated value of p is small.

7. DISCUSSION

We have proposed an EM-based survival mixture approach for modelling the distribution of survival time with random effects adjustment. The study demonstrates how random effects

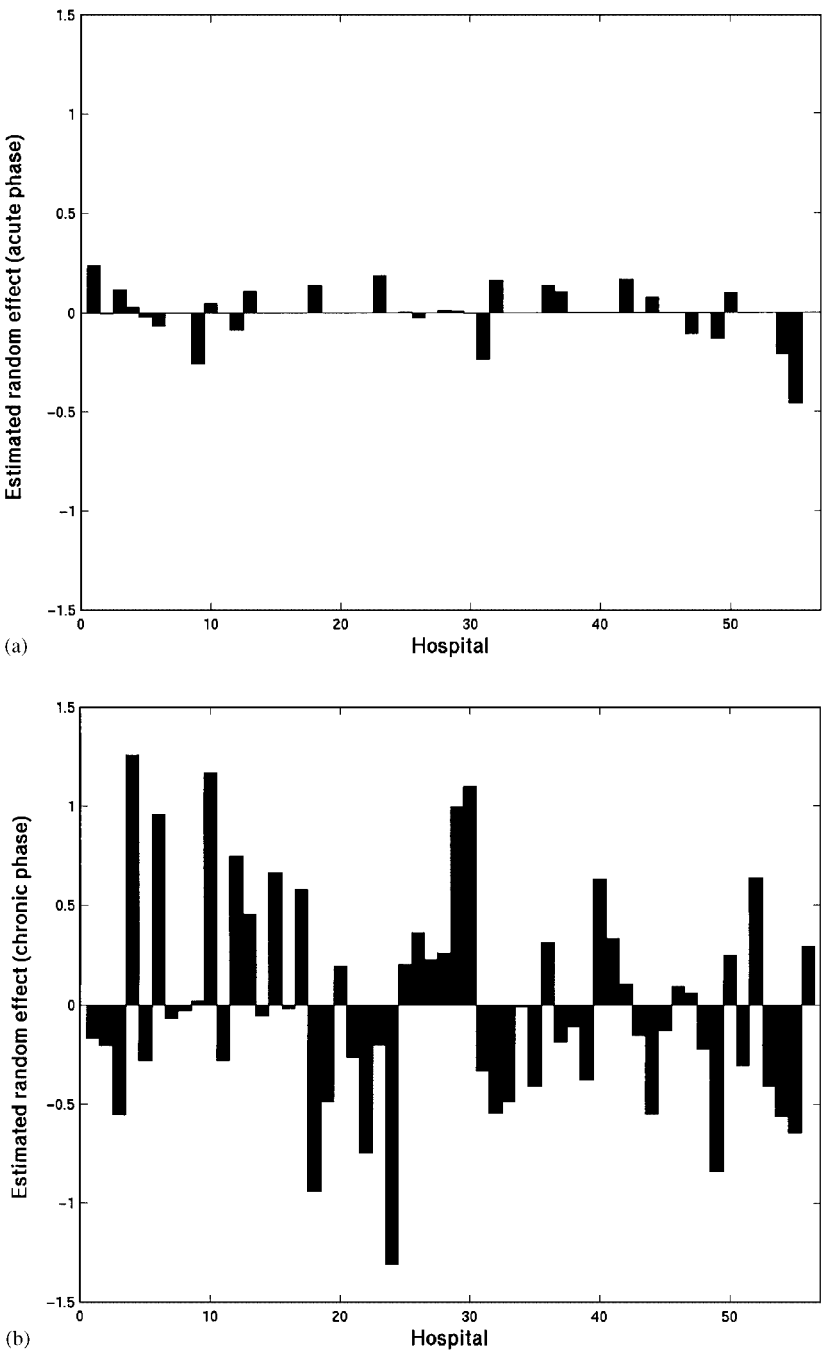

Figure 3. Prediction of random hospital effects for (a) the first component (acute phase) and (b) the second component (chronic phase).

Table III. Estimated biases and standard errors of REML estimators for two-component survival mixture model.

Parameter	True value	Average bias	SE_1	SE_2
$p=0.1$				
$\theta_1=0.5,\ \theta_2=0.5$				
p	0.1	0.022	0.023	0.020
β_1	0.5	−0.035	0.194	0.319
β_2	−0.5	−0.031	0.112	0.104
θ_1	0.5	0.091	0.366	0.393
θ_2	0.5	0.048	0.252	0.228
$\theta_1=1,\ \theta_2=1$				
p	0.1	0.021	0.023	0.020
β_1	0.5	−0.043	0.201	0.333
β_2	−0.5	−0.035	0.111	0.097
θ_1	1.0	0.074	0.573	0.739
θ_2	1.0	0.062	0.425	0.418
$p=0.3$				
$\theta_1=0.5,\ \theta_2=0.5$				
p	0.3	0.016	0.023	0.025
β_1	0.5	0.003	0.103	0.134
β_2	−0.5	−0.025	0.129	0.122
θ_1	0.5	0.062	0.243	0.271
θ_2	0.5	0.051	0.275	0.239
$\theta_1=1,\ \theta_2=1$				
p	0.3	0.015	0.022	0.026
β_1	0.5	0.001	0.105	0.140
β_2	−0.5	−0.025	0.129	0.117
θ_1	1.0	0.045	0.400	0.456
θ_2	1.0	0.035	0.440	0.431
$p=0.5$				
$\theta_1=0.5,\ \theta_2=0.5$				
p	0.5	0.009	0.021	0.024
β_1	0.5	0.014	0.075	0.096
β_2	−0.5	−0.023	0.155	0.148
θ_1	0.5	0.065	0.216	0.222
θ_2	0.5	0.066	0.320	0.283
$\theta_1=1,\ \theta_2=1$				
p	0.5	0.011	0.022	0.026
β_1	0.5	0.018	0.076	0.097
β_2	−0.5	−0.029	0.155	0.143
θ_1	1.0	0.093	0.394	0.405
θ_2	1.0	0.057	0.488	0.476

SE_1: average of standard error of estimates.
SE_2: sample standard error of estimates over 500 replications.

can be adjusted within a mixture-modelling framework in survival analysis. The estimation procedure is based on the GLMM approach [13, 32, 33]. Alternatively, exact ML approaches can be applied by integrating out the random effects in the joint likelihood. However, due to the intractability of the (marginal) likelihood function, computationally intensive numerical approximations are usually required to maximize the marginal likelihood [18, 34]. Comparative advantages of the different GLMM formulations have been discussed elsewhere [35].

In this paper, the component-hazard functions are assumed to be the Weibull distribution. The methodology described in Sections 3 and 4 can be readily modified to accommodate other lifetime distributions for both components. The model can also be extended to analyse survival data arising from other hierarchical settings than hospital clustering. For example, patients may be nested under different health regions or local districts within the state. As described in Section 2, the mixing proportion p may be specified as a function of covariates x_{ij}. The layout of the methodology should be sufficiently clear for the development of a survival model with random effect adjustment via the linear predictor in the functional form of $p(x_{ij})$. However, if the mixing proportion and the conditional survival functions are both expressed in terms of the same set of covariates x_{ij}, identifiability problems may occur when there is a large proportion of censoring observations [36].

The proposed model can be generalized to analyse cure rate problems and competing-risks data with nested random effects. With cure rate problems, a proportion of individuals are not susceptible to the failure risk [37–39]. These cured patients are referred to as long-term survivors with respect to the failure under study [17, 40]. Unlike the analysis of the present stroke-specific mortality data, where the aim is to identify and assess risk factors affecting patients' survival in different phases, studies on cure rate data focus mainly on the estimation of the cure proportion. With competing-risk data, each individual will die from one of multiple causes of failure [22, 36]. Survival model (2) may be adapted so that the components correspond to the different causes of failure. Under this setting, p denotes the proportion of patients who died from the first cause and D_{ij} defines failure types or a censored observation.

Analysis of the ischaemic stroke-specific mortality data has identified different risk factors affecting the survival of patients in the acute and chronic phases. The results provide useful information to establish hospital care strategy and policy for better utilization of resources according to these two phases. As described in Section 6, the estimated standard errors of the estimates given in Table II have been interpreted with caution since the estimated proportion of patients in the acute phase is small ($p = 0.1$). In the analysis, significant hospital variation is detected in a patient's survival during the chronic phase. The predicted random hospital effects facilitate the comparison of hospital performances in ischaemic stroke treatment and rehabilitation at the chronic phase, after adjustment for patient characteristics and clinical risk factors.

APPENDIX A: ESTIMATION OF VARIANCE COMPONENTS AND ASYMPTOTIC VARIANCES

The approximate REML estimates of the variance components θ_1, θ_2 and the asymptotic variances of $\hat{p}$, $\hat{\beta}_1$, $\hat{\beta}_2$ are obtained based on Reference [14]. With reference to (6), denote Ω the negative second derivative of $l = l_1 + l_2$ with respect to $p \,|\, \beta_1 \,|\, \beta_2 \,|\, u_1 \,|\, u_2$ in the BLUP

procedure. Let $\Omega^{-1}=(A_{ij})$, $(i=1,\ldots,5,\ j=1,\ldots,5)$, and the matrix is partitioned conformally to $p\,|\,\beta_1\,|\,\beta_2\,|\,u_1\,|\,u_2$, we have

$$\hat{\theta}_1 = M^{-1}(\operatorname{tr} A_{44} + \hat{u}_1^{\mathrm{T}}\hat{u}_1) \tag{A1}$$

$$\hat{\theta}_2 = M^{-1}(\operatorname{tr} A_{55} + \hat{u}_2^{\mathrm{T}}\hat{u}_2) \tag{A2}$$

$$\operatorname{var}\begin{pmatrix}\hat{p}\\ \hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix} = \begin{bmatrix}A_{11} & A_{12} & A_{13}\\ A_{21} & A_{22} & A_{23}\\ A_{31} & A_{32} & A_{33}\end{bmatrix} \tag{A3}$$

$$\operatorname{var}\begin{pmatrix}\hat{\theta}_1\\ \hat{\theta}_2\end{pmatrix} = 2\begin{bmatrix}\theta_1^{-2}(M-2\theta_1^{-1}\operatorname{tr}A_{44})+\theta_1^{-4}\operatorname{tr}(A_{44}^2) & \theta_1^{-2}\theta_2^{-2}\operatorname{tr}(A_{45}A_{54})\\ \theta_1^{-2}\theta_2^{-2}\operatorname{tr}(A_{45}A_{54}) & \theta_2^{-2}(M-2\theta_2^{-1}\operatorname{tr}A_{55})+\theta_2^{-4}\operatorname{tr}(A_{55}^2)\end{bmatrix}^{-1} \tag{A4}$$

APPENDIX B: IMPLEMENTATION OF THE EM-BASED MIXTURE APPROACH

From (6), the E-step involves the calculation of $Q(\psi,\psi^{(k)}) = Q_p^{(k)} + Q_{\xi_1}^{(k)} + Q_{\xi_2}^{(k)}$, where

$$Q_p^{(k)} = \sum_{i=1}^{M}\sum_{j=1}^{n_i}\left[\tau_{ij}^{(k)}\log\left(\frac{p}{1-p}\right)+\log(1-p)\right]$$

$$Q_{\xi_1}^{(k)} = \sum_{i=1}^{M}\sum_{j=1}^{n_i}[\tau_{ij}^{(k)}l_{11,ij}+l_{21,ij}]$$

$$Q_{\xi_2}^{(k)} = \sum_{i=1}^{M}\sum_{j=1}^{n_i}[(1-\tau_{ij}^{(k)})l_{12,ij}+l_{22,ij}]$$

and where

$$l_{1g,ij} = D_{ij}\log f_g(t_{ij};x_{ij}) + (1-D_{ij})\log S_g(t_{ij};x_{ij}) \quad (g=1,2)$$

$$l_{2g,ij} = -\frac{1}{2}\left(M\log(2\pi\theta_g)+\frac{u_g^{\mathrm{T}}u_g}{\theta_g}\right) \quad (g=1,2)$$

and

$$\tau_{ij}^{(k)} = E_{\psi^{(k)}}(Z_{ij}\,|\,t_{ij},x_{ij}) = \frac{p^{(k)}(f_1^{(k)})^{D_{ij}}(S_1^{(k)})^{(1-D_{ij})}}{p^{(k)}(f_1^{(k)})^{D_{ij}}(S_1^{(k)})^{(1-D_{ij})}+(1-p^{(k)})(f_2^{(k)})^{D_{ij}}(S_2^{(k)})^{(1-D_{ij})}} \tag{B1}$$

is the current estimated posterior probability that t_{ij} belongs to the first component, where $E_{\psi^{(k)}}$ denotes the expectation based on the current fit $\psi^{(k)}$, $f_g^{(k)} = f_g(t_{ij}; x_{ij}, \xi_g^{(k)})$, and $S_g^{(k)} = S_g(t_{ij}; x_{ij}, \xi_g^{(k)})$ $(g = 1, 2)$.

The M-step provides the updated estimate $\psi^{(k+1)}$ that maximizes $Q(\psi, \psi^{(k)})$ with respect to ψ and thus involves solving the non-linear equations

$$\begin{aligned}
&\text{for } \beta_g (g = 1,2): \quad \sum_{i=1}^{M} \sum_{j=1}^{n_i} (\tau_{ij}^{(k)})^{(2-g)} (1 - \tau_{ij}^{(k)})^{(g-1)} [D_{ij} + \log S_g(t_{ij}; x_{ij})] x_{ij} = 0 \\
&\text{for } u_g (g = 1,2): \quad \sum_{j=1}^{n_i} [(\tau_{ij}^{(k)})^{(2-g)} (1 - \tau_{ij}^{(k)})^{(g-1)} (D_{ij} + \log S_g(t_{ij}; x_{ij}))] - \frac{u_g}{\theta_g} = 0 \\
&\text{for } \lambda_g (g = 1,2): \quad \sum_{i=1}^{M} \sum_{j=1}^{n_i} [(\tau_{ij}^{(k)})^{(2-g)} (1 - \tau_{ij}^{(k)})^{(g-1)} \left[\frac{D_{ij}}{\lambda_g} - \exp(\eta_g(x_{ij})) t_{ij}^{\alpha_g} \right] = 0 \\
&\text{for } \alpha_g (g = 1,2): \quad \sum_{i=1}^{M} \sum_{j=1}^{n_i} (\tau_{ij}^{(k)})^{(2-g)} (1 - \tau_{ij}^{(k)})^{(g-1)} \left[\frac{D_{ij} + (D_{ij}\alpha_g - h_g(t_{ij}; x_{ij}) t_{ij}) \log t_{ij}}{\alpha_g} \right] = 0
\end{aligned} \tag{B2}$$

and the following closed-form equation for p:

$$p^{(k+1)} = \sum_{i=1}^{M} \sum_{j=1}^{n_i} \tau_{ij}^{(k)} / N \tag{B3}$$

The MINPACK routine HYBRD1 [41] is adopted to find a solution to (B2). The estimation procedure of the EM-based approach is summarized as follows:

1. Set initial values $\theta_1^{(0)}$, $\theta_2^{(0)}$, $p^{(0)}$, $\xi_1^{(0)}$, and $\xi_2^{(0)}$.
2. Calculate τ_{ij} using (B1), update $\xi_g (g = 1, 2)$ by (B2), and update p by (B3).
3. Repeat Step 2 until convergence.
4. Update θ_1 and θ_2 using (A1) and (A2), respectively.
4. Repeat Steps 2–4 until convergence.
5. Calculate the standard errors of $\hat{p}$, $\hat{\beta}_g$, and $\hat{\theta}_g (g = 1, 2)$ by (A3) and (A4).

In our analysis, we set initial values $\theta_1^{(0)} = \theta_2^{(0)} = 1$ and obtain $p^{(0)}$, $\xi_1^{(0)}$, and $\xi_2^{(0)}$ based on the preliminary result of the two-component Weibull mixture regression model without the random hospital effect adjustment (Section 5).

ACKNOWLEDGEMENTS

The authors wish to thank the Editor and the referees for helpful comments on the paper. The authors are grateful to the Health Information Centre, Health Department of Western Australia, for providing the ischaemic stroke mortality data. This work was supported in part by Grants from the Australian Research Council, the Research Grants Council of Hong Kong, and the National Health and Medical Research Council of Australia.

REFERENCES

1. De Angelis R, Capocaccia R, Hakulinen T, Soderman B, Verdecchia A. Mixture models for cancer survival analysis: application to population-based data with covariates. *Statistics in Medicine* 1999; **18**:441–454.
2. Farewell VT, Coates RA, Fanning MM *et al*. The probability of progression to AIDS in a cohort of male sexual contacts of men with HIV disease. *International Journal of Epidemiology* 1992; **21**:131–135.
3. Kuk AYC, Chen CH. A mixture model combining logistic-regression with proportional hazards regression. *Biometrika* 1992; **79**:531–541.
4. McLachlan GJ, McGiffin DC. On the role of finite mixture models in survival analysis. *Statistical Methods in Medical Research* 1994; **3**:211–226.
5. Phillips N, Coldman A, McBride ML. Estimating cancer prevalence using mixture models for cancer survival. *Statistics in Medicine* 2002; **21**:1257–1270.
6. Blackstone EH, Naftel DC, Turner ME. The decomposition of time-varying hazard into phases, each incorporating a separate stream of concomitant information. *Journal of the American Statistical Association* 1986; **81**:615–624.
7. McGiffin DC, Galbraith AJ, McLachlan GJ *et al*. Aortic valve infection—risk factors for death and recurrent endocarditis following aortic valve replacement. *Journal of Thoracic and Cardiovascular Surgery* 1992; **104**:511–520.
8. Hankey GJ. Transient ischaemic attacks and stroke. *Medical Journal of Australia* 2000; **172**:394–400.
9. Goldstein LB. Accuracy of ICD-9-CM coding for the identification of patients with acute ischaemic stroke. *Stroke* 1998; **29**:1602–1604.
10. National Centre for Classification in Health. *Australian Coding Standards. The International Statistical Classification of Diseases and Related Health Problems*, *10th Revision*, *Australian Modification*. National Centre for Classification in Health: Sydney, 1998.
11. Lee AH, Wang K, Yau KKW, Somerford PJ. Truncated negative binomial mixed regression modelling of ischaemic stroke hospitalizations. *Statistics in Medicine* 2003; **22**:1129–1139.
12. Anderson CS, Jamrozik ZK, Broadhurst RJ, Stewart-Wynne EG. Predicting surviving among different subtypes of stroke: experience from the Perth Community Stroke Study, 1989–1990. *Stroke* 1994; **25**:1935–1944.
13. McGilchrist CA. REML estimation for survival models with frailty. *Biometrics* 1993; **49**:221–225.
14. McGilchrist CA. Estimation in generalised mixed models. *Journal of the Royal Statistical Society Series B* 1994; **56**:61–69.
15. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B* 1977; **39**:1–38.
16. McLachlan GJ, Krishnan T. *The EM Algorithm and Extensions*. Wiley: New York, 1997.
17. Yau KKW, Ng ASK. Long-term survivor mixture model with random effects: application to a multicentre clinical trial of carcinoma. *Statistics in Medicine* 2001; **20**:1591–1607.
18. Booth JG, Hobert JP. Maximum generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society Series B* 1999; **61**:265–285.
19. Wei GCG, Tanner MA. A Monte Carlo implementation of the EM algorithm and the Poor Man's data augmentation algorithm. *Journal of the American Statistical Association* 1990; **85**:699–704.
20. Quantin C, Sauleau E, Bolard P *et al*. Modeling of high-cost patient distribution within renal failure diagnosis related group. *Journal of Clinical Epidemiology* 1999; **52**:251–258.
21. Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society Series B* 1972; **34**:187–220.
22. Larson MG, Dinse GE. A mixture model for the regression analysis of competing risks data. *Applied Statistics* 1985; **34**:201–211.
23. Gordon NH. Application of the theory of finite mixtures for the estimation of 'cure' rates of treated cancer patients. *Statistics in Medicine* 1990; **9**:397–407.
24. Peng YW, Dear KBG, Denham JW. A generalized F mixture model for cure rate estimation. *Statistics in Medicine* 1998; **17**:813–830.
25. Breslow NE, Clayton DG. Approximate inference in generalised linear mixed models. *Journal of the American Statistical Association* 1993; **88**:9–25.
26. Schall R. Estimation in generalized linear mixed models with random effects. *Biometrika* 1991; **78**:719–727.
27. Patterson HD, Thompson R. Recovery of inter-block information when block sizes are unequal. *Biometrika* 1971; **58**:545–554.
28. McLachlan GJ, Peel D. *Finite Mixture Models*. Wiley: New York, 2000.
29. Meng XL, van Dyk DA. Fast EM-type implementations for mixed effects models. *Journal of the Royal Statistical Society Series B* 1998; **60**:559–578.
30. Gelfand AE, Ghosh SK, Christiansen C *et al*. Proportional hazards models: a latent competing risk approach. *Applied Statistics* 2000; **49**:385–397.
31. Rosen O, Tanner M. Mixtures of proportional hazards regression models. *Statistics in Medicine* 1999; **18**: 1119–1131.

32. Yau KKW, McGilchrist CA. ML and REML estimation in survival analysis with time dependent correlated frailty. *Statistics in Medicine* 1998; **17**:1201–1213.
33. Yau KKW. Multi-level models for survival analysis with random effects. *Biometrics* 2001; **57**:96–102.
34. McCulloch CE. Maximum likelihood algorithms for generalised linear mixed models. *Journal of the American Statistical Association* 1997; **92**:162–170.
35. Yau KKW, Kuk AYC. Robust estimation in generalised linear mixed models. *Journal of the Royal Statistical Society Series B* 2002; **64**:101–117.
36. Ng SK, McLachlan GJ. An EM-based semiparametric mixture model approach to the regression analysis of competing-risks data. *Statistics in Medicine* 2003; **22**:1097–1111.
37. Maller RA, Zhou S. Testing for the presence of immune or cured individuals in censored survival data. *Biometrics* 1995; **51**:1197–1205.
38. Peng YW, Dear KBG. A nonparametric mixture model for cure rate estimation. *Biometrics* 2000; **56**:237–243.
39. Sy JP, Taylor JMG. Estimation in a Cox proportional hazards cure model. *Biometrics* 2000; **56**:227–236.
40. Tsodikov A. Semi-parametric models of long- and short-term survival: an application to the analysis of breast cancer survival in Utah by age and stage. *Statistics in Medicine* 2002; **21**:895–920.
41. Moré JJ, Garbow BS, Hillstrom KE. *User Guide for MINPACK-1*, *ANL-80-74*. Argonne National Laboratory: Chicago, 1980.

