Elsevier Editorial System(tm) for Computers in Biology and Medicine

Manuscript Draft

Manuscript Number:

Title: A computer graphical user interface for survival mixture modelling of recurrent infections

Article Type: Full Length Article

Keywords: Accelerated failure time; Graphical user interface; Mixture model; Random effects; Recurrent infections; Survival data

Corresponding Author: Professor Andy H Lee, Ph.D.

Corresponding Author's Institution: Curtin University of Technology

First Author: Andy H Lee, Ph.D.

Order of Authors: Andy H Lee, Ph.D.; Yun Zhao, Ph.D.; Kelvin K Yau, Ph.D.; Shu-Kay Ng, Ph.D.

Abstract: Recurrent infections data are commonly encountered in medical research, where the recurrent events are characterised by an acute phase followed by a stable phase after the index episode. Two-component survival mixture models, in both proportional hazards and accelerated failure time settings, are presented as a flexible method of analysing such data. To account for the inherent clustering and dependency of the recurrent observations, random effects are incorporated within the conditional hazard function, in the manner of generalised linear mixed models. Assuming a Weibull or log-logistic baseline hazard in both mixture components of the survival mixture model, an EM algorithm is developed for the residual maximum quasi-likelihood estimation of fixed effect and variance components parameters. The methodology is implemented as a graphical user interface coded using Microsoft visual C++. Application to model recurrent urinary tract infections for elderly women is illustrated, where significant individual variations are evident at both acute and stable phases. The survival mixture methodology developed enable practitioners to identify pertinent risk factors affecting the recurrent times and to draw valid conclusions inferred from these clustered and heterogeneous survival data.

A computer graphical user interface for survival mixture modelling of recurrent infections

Andy H. Lee, Yun Zhao

Department of Epidemiology and Biostatistics, School of Public Health, Curtin University of Technology, Perth, WA, Australia

Kelvin K.W. Yau

Department of Management Sciences, City University of Hong Kong, Hong Kong

S.K. Ng

School of Medicine, Griffith University, Meadowbrook, QLD, Australia

---

Address for correspondence:

Professor Andy H. Lee

Department of Epidemiology and Biostatistics

School of Public Health

Curtin University of Technology

GPO Box U 1987, Perth, WA, 6845

Australia

Phone: +61-8-92664180

Fax: +61-8-92662958

E-mail: Andy.Lee@curtin.edu.au

A computer graphical user interface for survival mixture modelling of recurrent infections

ABSTRACT

Recurrent infections data are commonly encountered in medical research, where the recurrent events are characterised by an acute phase followed by a stable phase after the index episode. Two-component survival mixture models, in both proportional hazards and accelerated failure time settings, are presented as a flexible method of analysing such data. To account for the inherent clustering and dependency of the recurrent observations, random effects are incorporated within the conditional hazard function, in the manner of generalised linear mixed models. Assuming a Weibull or log-logistic baseline hazard in both mixture components of the survival mixture model, an EM algorithm is developed for the residual maximum quasi-likelihood estimation of fixed effect and variance components parameters. The methodology is implemented as a graphical user interface coded using Microsoft visual C++. Application to model recurrent urinary tract infections for elderly women is illustrated, where significant individual variations are evident at both acute and stable phases. The survival mixture methodology developed enable practitioners to identify pertinent risk factors affecting the recurrent times and to draw valid conclusions inferred from these clustered and heterogeneous survival data.

Keywords: Accelerated failure time; Graphical user interface; Mixture model; Random effects; Recurrent infections; Survival data

1. Introduction

Survival mixture models are often used to model heterogeneous failure time data in medical research [1-4]. Recently, a two-component Weibull survival mixture model has been proposed to analyse ischaemic stroke-specific survival time [5], in which patients are grouped into acute and chronic phases after the index stroke event. The survival mixture approach is applicable because the two phases overlap each other in time and thus the risk of death cannot be described satisfactorily by fitting separate parametric model to each time period [5]. Within the class of survival mixture models, the hazard rates are often assumed to be proportional. Although the proportional hazards assumption is appropriate in many situations, an attractive alternative is the accelerated failure time (AFT) model [6], whereby the covariates can affect the survival experience of patients by speeding up or slowing down the survival time. The AFT model relates covariates linearly to the logarithm of the survival time and provides a wide range of parametric forms for the hazard function. It is more sensible for wear-related processes if a constant load differs between individuals. The AFT model also provides robust parameter estimates toward neglected covariates (i.e. when not all relevant covariates are included) [7-9].

An important issue concerns the heterogeneity due to random hospital effects arising from the clustering of observations within the same hospital or centre [10]. Such random effects can be adjusted within the two-component survival mixture framework [5]. In the limiting case, the two-component model reduces to a long-term survivor mixture model with random effects [11]. A similar survival model with a cure fraction has also been developed [12]. In the AFT setting, an AFT model with normal random effects was introduced by Klein *et al* [13]. A frailty-based AFT model was also available to analyse recurrent infections for kidney patients, where the frailty was drawn from a gamma distribution [14]. Moreover, parametric AFT

models have been suggested to model kidney transplant survival data, with different random effect distributions and baseline hazards for the underlying hazard function [15].

Recurrent infections data are commonly encountered in medical research, where the recurrent events are characterised by an acute phase followed by a stable phase after the index episode. Although existing survival mixture models allow the specification of a cured proportion and/or the mixing of survival functions for lifetime distribution with overlapping phases, the issue of natural clustering of recurrent observations has not been addressed satisfactorily in the literature. On the other hand, current survival frailty models are essentially limited to a single component survival function. In the presence of simultaneous heterogeneity and dependency, application of such procedures may lead to inaccurate hazard rates and consequently incorrect inferences concerning pertinent risk factors and the effectiveness of preventive treatment or interventions [5].

The purpose of this study is to present a unified and flexible approach of modelling recurrent infections data by a finite mixture of survival distributions incorporating random effects. In the next section, two-component survival mixture models are proposed in both proportional hazards and AFT settings to correspond to the acute and stable phases of recurrent infections. Random effects are accommodated within each conditional hazard function, in the manner of generalised linear mixed models [16]. Assuming a Weibull or log-logistic baseline hazard in both mixture components of the survival mixture model, an EM algorithm is developed for the residual maximum quasi-likelihood estimation of fixed effect and variance components parameters. In section 3, the methodology is implemented as a graphical user interface coded using the Microsoft visual C++ language. Application of the procedure to model recurrent urinary tract infections for elderly women are illustrated in Section 4, the results of which will

enable medical practitioners to concentrate on the substantive issues and to draw valid conclusions from the recurrent outcomes. Finally, some concluding remarks, including an outline of further extensions, are provided in Section 5.

2. Two-component survival mixture models with random effects

Let Y_{ij} denote the jth recurrent time (j = 1, 2, …, n_i) within cluster i (i = 1, 2, …, M), with $N = \sum_{i=1}^{M} n_i$ being the total number of observations. In addition to $T_{ij} = \min(C_{ij}, Y_{ij})$, where C_{ij} represents the random censoring time independent of Y_{ij}, a censoring indicator δ_{ij} is observed:

$$\delta_{ij} = I(Y_{ij} \le C_{ij}) = \begin{cases} 1, \text{if } Y_{ij} \le C_{ij}, \\ 0, \text{if } Y_{ij} > C_{ij}. \end{cases}$$

Let x_{ij} be a vector of covariates associated with T_{ij}. The survival function of T can be modelled by a two-component finite mixture as:

$$S(t_{ij}, x_{ij}) = pS_1(t_{ij}, x_{ij}) + (1-p)S_2(t_{ij}, x_{ij}), \tag{1}$$

and the corresponding probability density function of T is:

$$f(t_{ij}, x_{ij}) = pf_1(t_{ij}, x_{ij}) + (1-p)f_2(t_{ij}, x_{ij}), \tag{2}$$

where p denotes the proportion of subjects or observations in the acute phase, $S_g(t_{ij}, x_{ij})$ and $f_g(t_{ij}, x_{ij})$ are the conditional survival function and conditional density function of the g^{th} component (g = 1, 2), respectively. With the concomitant information x_{ij}, effects of demographic characteristics and co-morbidities in the acute and stable phases of infection can be determined. Moreover, if the second component $S_2(t_{ij}, x_{ij}) = 1$, it reduces to the long-term survivor model [11]. Survival mixture models with random effects are next constructed in both proportional hazards and AFT settings.

Under the proportional hazards assumption, the conditional hazard function for the g^{th} component is given by

$$h_g(t_{ij}, x_{ij}) = h_{g0}(t_{ij}) \exp(\eta_g(x_{ij})), \tag{3}$$

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

$$h_{g0}(t_{ij}) = \lambda_g \gamma_g t_{ij}^{\gamma_g - 1}, \tag{4}$$

where $\lambda_g, \gamma_g > 0$ are unknown parameters.

If a Weibull AFT model is assumed, the conditional hazard function for the g^{th} component is given by

$$h_g(t_{ij}, x_{ij}) = \lambda_g \gamma_g t_{ij}^{\gamma_g - 1} \exp(\gamma_g \eta_g(x_{ij})), \tag{5}$$

which may be considered as a Weibull distribution with scale $\lambda_g \exp(\gamma_g \eta_g(x_{ij}))$ and shape parameter γ_g. Same as the Weibull proportional hazards model, covariates affect the scale but not the shape parameter in model (5). Alternatively, a log-logistic AFT model may be defined by the conditional hazard function:

$$h_g(t_{ij}, x_{ij}) = \frac{\gamma_g t_{ij}^{\gamma_g - 1} \exp(\lambda_g + \gamma_g \eta_g(x_{ij}))}{1 + \exp(\lambda_g + \gamma_g \eta_g(x_{ij})) t_{ij}^{\gamma_g}}. \tag{6}$$

Again, $-\infty < \lambda_g < \infty$, and $\gamma_g > 0$ are unknown parameters.

For both proportional hazards and AFT settings, an unobserved random term can be introduced in each conditional hazard function to explain the variability shared by the

clustered observations, following the generalised linear mixed models formulation [16]. Specifically,

$$\eta_g(x_{ij}) = x_{ij}^T \beta_g + U_{gi}, \tag{7}$$

where β_g is the vector of regression coefficients. Without loss of generality, the random cluster effects U_{gi} are taken to be i.i.d. $N(0,\theta_g)$. Based on this formulation, the vector of unknown parameters is $\psi = (p, \beta_1^T, \beta_2^T, u_1^T, u_2^T, \lambda_1, \lambda_2, \gamma_1, \gamma_2)$ where $u_1^T = [U_{11}, U_{12}, \ldots, U_{1M}]$ and $u_2^T = [U_{21}, U_{22}, \ldots, U_{2M}]$. One approach for parameter estimation is by commencing with the best linear unbiased predictor (BLUP) at the initial step and extends to obtain residual maximum quasi-likelihood (REMQL) estimators for the variance component parameters [17]. For given initial values of θ_g, the BLUP estimator of ψ maximizes $l = l_1 + l_2$, where

$$l_1 = \sum_{i=1}^{M} \sum_{j=1}^{n_i} \left[\delta_{ij} \log f(t_{ij}, x_{ij}) + (1 - \delta_{ij}) \log S(t_{ij}, x_{ij}) \right],$$

$$l_2 = -\tfrac{1}{2}[M \log 2\pi\theta_1 + (1/\theta_1) u_1^T u_1] - \tfrac{1}{2}[M \log 2\pi\theta_2 + (1/\theta_2) u_2^T u_2]. \tag{8}$$

Here, l_1 represents the partial log-likelihood of recurrent times conditional on u_1 and u_2, whereas l_2 is the logarithm of the joint probability density function of u_1 and u_2, with u_1 and u_2 being independent. The BLUP estimate of ψ is obtained as a solution of the equation $\partial l / \partial \psi = 0$, which can be solved via an EM algorithm [18]. The REMQL estimates of the variance components θ_1 and θ_2 are then obtained by maximizing the restricted log-likelihood function. Details of the EM estimation procedure are given in Appendix A, and derivations for the REMQL estimates and asymptotic variances are provided in Appendix B.

3. Software Implementation

To implement the modelling methodology and estimation procedure described in Section 2, a graphical user interface (GUI) is developed and coded using the Microsoft visual C++

scientific language, with adaptations taken from numerical library subroutines [19]. This objective-orientated interface is systematically designed for visualization and manipulation of survival data sets. The user-defined components comprise dialogue windows featuring data input and listing, model specification and post-modelling graphical assessments. The program main panel is shown in Figure 1. It allows users to interactively select the model setting or course of action pertinent to the data set in hand. The GUI has the potential to be upgraded into an independent package by incorporating further modelling extensions outlined in Section 5.

[Figure 1]

3.1. Data input and variable selection

The raw data set can be imported into the main panel from any directory after invoking the GUI program. Corrections and additional data entries can be made in the displayed area resembling a window-based spreadsheet. As shown in Figure 2, the variable selection window then prompts the user to define the survival time and choose appropriate covariates, as well as the censoring indicator and the variable describing the random effects or clustering of observations.

[Figure 2]

3.2. Model specification and parameter estimation

Specification of the survival mixture model is initiated by clicking "model" in the main panel. Figure 3 displays the model setting menu along with general information about the data set in the lower section of the dialogue window. Univariate normal is the default random effect type permitted in the current version, but Weibull proportional hazards, Weibull AFT or log-logistic AFT can be chosen as the preferred model type. Meanwhile, initial values of the model parameters are designed to be read from a separate text file which can be browsed and selected from any folder. Another option is the specification of the maximum number of

iterative steps and the associated convergence criterion. Different combinations are available. Suppose $\boldsymbol{I}$ is the pre-determined maximum number of iterations, r_{ψ} is the convergence criterion for ψ; and r_{θ} for the updating of θ_g. Successive iterations are carried out according to the EM algorithm given in Appendix A, until $|\hat{\psi}^{(k+1)} - \hat{\psi}^{(k)}| < r_{\psi}$ and $|\hat{\theta}^{(k+1)} - \hat{\theta}^{(k)}| < r_{\theta}$ or the specified $\boldsymbol{I}$ is attained, so that further iterations cannot improve the accuracy of the regression coefficients. Figure 4 shows an example output after completion of the model fitting process.

[Figure 3 and Figure 4]

3.3. Plotting and model assessment

Plotting facility is built into this interface for assisting with model assessment. Both survival and hazard functions can be plotted against a selected covariate for either mixture component, together with basic editing features for graph title, legend, tick marks and scaling factor for both x- and y-axes; see Figure 5. In addition, the adequacy of the survival mixture models can be assessed graphically by the Cox-Snell residual plot. The Cox-Snell residuals are defined as:

$$e_{ij} = -\log \hat{S}(t_{ij}, x_{ij}),\ i = 1,2,\cdots,M,\ j = 1,2,\cdots,n_i, \tag{9}$$

where $\hat{S}(t_{ij}, x_{ij})$ is the estimated survival function evaluated at parameter estimates $\hat{\psi}$. The Kaplan-Meier estimate, $\hat{K}(e_{ij})$, of the survival function of these residuals are computed, and values of $\log\{-\log \hat{K}(e_{ij})\}$ are then plotted against $\log(e_{ij})$.

[Figure 5]

4. Application to recurrent urinary tract infections

Urinary tract infection (UTI) is one of the most common bacterial infections in elderly women aged 60 years and over, and one in four of these women will develop a recurrence [20].

Various risk factors predispose women of different age groups to recurrence [21]. A retrospective cohort study on recurrent UTI was conducted among elderly women in residential aged-care facilities [22]. Eligibility criteria for the subjects were defined to be female residents aged 60 years or above with an institutionalisation period of at least six months. A total of 201 subjects satisfying the selection criteria were recruited from six aged-care institutions in Perth, Western Australia.

It was found that $M = 93$ of the 201 women experienced an index UTI episode during the two years follow-up period. For this subgroup of women, the outcome variable was taken to be the duration between successive UTI episodes. There are altogether $N = 285$ observations. One third of the cohort had no recurrence during the study period, while the maximum number of recurrent UTI was 17. The average age of the cohort was 85.8 (SD 8.4) years and 32 (34%) of them had a history of prior UTI. The mean recurrence time was 241 (SE 19.6) days.

[Table 1]

With covariates age and history of prior UTI taken at baseline, results from fitting the survival mixture models are presented in Table 1. The results are comparable between the three models. It appears that the hazard rate of recurrent UTI is significantly associated with the subject's history of prior UTI during the acute phase. Moreover, the acute phase proportion is estimated to be 74-82%. For all models, the random subject effects are significant in both acute and stable phases, implying that heterogeneity in UTI recurrence can be attributed to the differences between individual women. The identification of the pertinent risk factor, namely prior history of UTI, after accounting for inter-subject variation, provides useful information on how the recurrent infection in the acute phase is affected. The estimated cumulative hazard function of the first mixture component based on the Weibull AFT survival mixture model is

shown in Figure 6, where age is set at its median value. The increased hazard of UTI recurrence is evident if the woman has experienced UTI before. Indeed, the mean recurrence times were, respectively, 148 (SE 16.3) days and 338 (SE 32.4) days for women with and without a prior history of UTI.

[Figure 6]

Finally, an inspection of the Cox-Snell residual plots found that the Weibull AFT survival mixture model provides the best fit to the recurrent UTI data among the three models. As shown in Figure 7, the residuals from the fitted Weibull AFT survival mixture model generally follow a straight line with unit slope, indicating little departure from the assumed model.

[Figure 7]

5. Discussion

We have proposed an integrated and flexible approach of modelling recurrent infections, in which the observations are correlated and their corresponding survival distribution is composed of two overlapping phases in time. Estimation of parameters is implemented via an EM algorithm to facilitate model fitting. The application on recurrent UTI of elderly women demonstrates how random effects can be adjusted within the survival mixture modelling framework. Significant individual variations are evident at both acute and stable phases, while a pertinent risk factor affecting the recurrent times is identified. The survival mixture methodology developed thus enable practitioners to focus on substantive issues and to draw valid conclusions inferred from such clustered and heterogeneous survival data.

A two-component survival mixture model is considered reasonable for recurrent infections. In modelling the time to death following major cardiac surgery, for example, the risk of death may be characterized by three merging phases. The first (early) phase is the period immediately following surgery in which the risk of dying is relatively high. The second (constant) phase refers to the subsequent period in which the hazard rate is essentially constant. This second phase then merges with the third (late) phase in which the risk of death starts to increase [2]. Under such circumstances, survival mixture models with three or more components can be defined analogously.

Other generalisations are also plausible, including the relaxation of the independence assumption for the random components by allowing a bivariate normal distribution between the two random effect terms. In addition to Weibull and log-logistic distributions, the methodology described in Section 2 can be readily modified to accommodate other parametric or semi-parametric baseline hazard functions, thus giving an advanced and flexible framework to model clustered and heterogeneous survival data. Finally, the mixing proportion p may be specified as a function of covariates, say, via a logistic transform of p. However, if the conditional survival functions and p are both expressed in terms of the same set of covariates x_{ij}, identifiability problems may arise when the proportion of censoring observations is large [5].

The GUI implementation of the survival mixture modelling procedure has three advantages. Firstly, it is a stand-alone window based software and can be run in practically any personal computer that has a Windows XP or later operating system. Secondly, the GUI is integrated in the sense that data input, variable selection, model fitting, parameter estimation, post-modelling assessment and graphical output, are entirely accommodated within the package,

while allowing the user the flexibility to perform specific actions interactively depending on the nature of the data set in hand. Thirdly, the GUI platform can be readily modified by introducing additional dialogue windows and menus to implement model extensions. Further enhancement of the GUI incorporating the aforementioned generalisations will be considered in future research.

Acknowledgements

This research is supported by the Australian Research Council Discovery Grant (Project ID DP0559204) and the Research Grants Council of Hong Kong. A copy of the graphical user interface program is available from the corresponding author.

References

1. A.Y.C. Kuk, C.H. Chen, A mixture model combining logistic regression with proportional hazards regression, *Biometrika* 79 (1992) 531-541.
2. G.J. McLachlan, D.C. McGiffin, On the role of finite mixture models in survival analysis, *Stat. Method. Med. Res.* 3 (1994) 211-226.
3. R. De Angelis, R. Capocaccia, T. Hakulinen, B. Sodeman, A. Verdecchia, Mixture models for cancer survival analysis: application to population-based data with covariates, *Stat. Med.* 18 (1999) 441-454.
4. N. Phillips, A. Coldman, M.L. McBride, Estimating cancer prevalence using mixture models for cancer survival, *Stat. Med.* 21 (2002) 1257-1270.
5. S.K. Ng, G.J. McLachlan, K.K.W. Yau, A.H. Lee, Modelling the distribution of ischaemic stroke-specific survival time using an EM-based mixture approach with random effects adjustment, *Stat. Med.* 23 (2004), 2729 – 2744.
6. L.J. Wei, The accelerated failure time model: a useful alternative to the Cox regression in survival analysis (with discussion), *Stat. Med.* 11 (1992) 1871-1879.
7. N. Keiding, P.K. Andersen, J.P. Klein, The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates, *Stat. Med.* 16 (1997) 215-224.
8. P. Hougaard, Fundamentals of survival data, *Biometrics* 55 (1999) 13-22.
9. P. Hougaard, *Analysis of Multivariate Survival Data*, Springer, 2000.
10. T. Yamaguchi, Y. Ohashi, Y. Matsuyama, Proportional hazards models with random effects to examine centre effects in multicentre cancer clinical trials, *Stat. Method. Med. Res.* 11 (2002) 221-236.
11. K.K.W. Yau, S.K. Ng, Long-term survivor mixture model with random effects: Application to a multicentre clinical trial of carcinoma, *Stat. Med.* 20 (2001) 1591-1607.

12. F. Cooner, S. Banerjee, A.M. McBean, Modelling geographically referenced survival data with a cure fraction, *Stat. Method. Med. Res.* 15 (2006) 307-324.
13. J.P. Klein, C. Pelz, M. Zhang, Modeling random effects for censored data by multivariate normal regression model, *Biometrics* 55 (1999) 497-506.
14. W. Pan, Using frailties in the accelerated failure time model, *Lifetime Data Analy.* 7 (2001) 55-64.
15. P. Lambert, D. Collett, A. Kimber, R. Johnson, Parametric accelerated failure time models with random effects and an application to kidney transplant survival, *Stat. Med.* 23 (2004) 3177-3192.
16. C.A. McGilchrist, Estimation in generalised mixed models, *J. Royal Stat. Soc. Ser. B* 56 (1994) 61-69.
17. C.A. McGilchrist, K.K.W. Yau, The derivation of BLUP, ML, REML estimation methods for generalised linear mixed models, *Comm. Stat.-Theory Meth.* 24 (1995) 2963-2980.
18. G.J. Mclachlan, T. Krishnan, *The EM Algorithm and Extensions*, Wiley, 1997.
19. W.H. Press, W.T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*, *2nd edition*, Cambridge University Press, 1992.
20. J. Harper, Managing UTI in adults, *Practitioner* 244 (2000) 464-471.
21. A.V. Franco, Recurrent urinary tract infections, *Best Pract. Res. Clin. Obstet. Gynaecol.* 19 (2005) 861-873.
22. L. Xiang, A.H. Lee, K.K.W. Yau, G.J. McLachlan, A score test for zero-inflation in correlated count data, *Stat. Med.* 25 (2006) 1660-1671.
23. J.J. Moré, B.S. Garbow, K.E. Hillstrom. *User Guide for MINPACK-1; ANL-80-74*, Argonne National Laboratory, Chicago, 1980.

Appendix A. EM algorithm for BLUP estimation

An EM algorithm is developed for the iterative computation of the BLUP estimate for ψ. In order to pose the estimation procedure as an incomplete-data problem, an unobservable random vector $Z_{ij} = (z_{ij1}, z_{ij2})^T$ is introduced, indicating whether the observation t_{ij} belongs to the first or second component. The complete log-likelihood function is given by:

$$\begin{aligned} l_C(\psi) = \sum_{i=1}^{M}\sum_{j=1}^{n_i}\{z_{ij1}[\log p &+ \delta_{ij}\log f_1(t_{ij},x_{ij}) + (1-\delta_{ij})\log S_1(t_{ij},x_{ij})] \\ &+ z_{ij2}[\log(1-p) + \delta_{ij}\log f_2(t_{ij},x_{ij}) + (1-\delta_{ij})\log S_2(t_{ij},x_{ij})]\} \\ -\tfrac{1}{2}[\sum_{g=1}^{2}(M\log(2\pi\theta_g) &+ \frac{u_g^T u_g}{\theta_g})]. \end{aligned}$$

The EM estimation procedure involves the following steps.

(1) Initialization: Set initial values θ_0 (=1 say) for θ_1 and θ_2. Initial values of other parameters are obtained by fitting a two-component AFT survival mixture model without random effects.

(2) E-step: Calculate the conditional expectation of $l_C(\psi)$, given the observed data T and the current fit $\psi^{(k)}$ of ψ. For the k^{th} iterative step, k = 1, 2, …., compute the current estimated posterior probability that t_{ij} belongs to the first component:

$$\tau_{ij}^{(k)} = \frac{p^{(k)} f_1^{(k)}(t_{ij},x_{ij})^{\delta_{ij}} S_1^{(k)}(t_{ij},x_{ij})^{1-\delta_{ij}}}{p^{(k)} f_1^{(k)}(t_{ij},x_{ij})^{\delta_{ij}} S_1^{(k)}(t_{ij},x_{ij})^{1-\delta_{ij}} + (1-p^{(k)}) f_2^{(k)}(t_{ij},x_{ij})^{\delta_{ij}} S_2^{(k)}(t_{ij},x_{ij})^{1-\delta_{ij}}}$$

(3) M-step: Update the (k+1)$^{\text{th}}$ estimate $\psi^{(k+1)}$ of ψ by maximizing the Q-function

$$\begin{aligned} Q(\psi,\psi^{(k)}) = \sum_{i=1}^{M}\sum_{j=1}^{n_i}\{\tau_{ij}^{(k)} \ [\log p &+ \delta_{ij}\log f_1(t_{ij},x_{ij}) + (1-\delta_{ij})\log S_1(t_{ij},x_{ij})] \\ &+ (1-\tau_{ij}^{(k)})[\log(1-p) + \delta_{ij}\log f_2(t_{ij},x_{ij}) + (1-\delta_{ij})\log S_2(t_{ij},x_{ij})]\} \\ -\tfrac{1}{2}[\sum_{g=1}^{2}(M\log(2\pi\theta_g) &+ \frac{u_g^T u_g}{\theta_g})]. \end{aligned}$$

with respect to ψ , which involves solving a set of non-linear equations and the MINPACK routine HYBRD1 is used for this purpose [23].

For Weibull proportional hazards mixture model, these equations are (g = 1, 2):

$$\begin{cases} \sum_{i=1}^{M}\sum_{j=1}^{n_i}(\tau_{ij}^{(k)})^{(2-g)}(1-\tau_{ij}^{(k)})^{(g-1)}[\delta_{ij}+\log S_g]=0; \\ \sum_{i=1}^{M}\sum_{j=1}^{n_i}(\tau_{ij}^{(k)})^{(2-g)}(1-\tau_{ij}^{(k)})^{(g-1)}[\delta_{ij}+(\delta_{ij}+\log S_g)\gamma_g\log t_{ij}]=0; \\ \sum_{i=1}^{M}\sum_{j=1}^{n_i}(\tau_{ij}^{(k)})^{(2-g)}(1-\tau_{ij}^{(k)})^{(g-1)}[\delta_{ij}+\log S_g]x_{ij}^{T}=0; \\ \sum_{j=1}^{n_i}(\tau_{ij}^{(k)})^{(2-g)}(1-\tau_{ij}^{(k)})^{(g-1)}[\delta_{ij}+\log S_g]-\frac{U_g}{\theta_g}=0. \end{cases}$$

For Weibull AFT mixture model, these equations are (g = 1, 2):

$$\begin{cases} \sum_{i=1}^{M}\sum_{j=1}^{n_i}(\tau_{ij}^{(k)})^{(2-g)}(1-\tau_{ij}^{(k)})^{(g-1)}[\delta_{ij}+\log S_g]=0; \\ \sum_{i=1}^{M}\sum_{j=1}^{n_i}(\tau_{ij}^{(k)})^{(2-g)}(1-\tau_{ij}^{(k)})^{(g-1)}[\delta_{ij}+(\delta_{ij}+\log S_g)(\log t_{ij}+\eta_g)\gamma_g]=0; \\ \sum_{i=1}^{M}\sum_{j=1}^{n_i}(\tau_{ij}^{(k)})^{(2-g)}(1-\tau_{ij}^{(k)})^{(g-1)}[\delta_{ij}+\log S_g]x_{ij}^{T}\gamma_g=0; \\ \sum_{j=1}^{n_i}(\tau_{ij}^{(k)})^{(2-g)}(1-\tau_{ij}^{(k)})^{(g-1)}[\delta_{ij}+\log S_g]\gamma_g-\frac{U_g}{\theta_g}=0. \end{cases}$$

For log-logistic AFT mixture model, these equations are (g = 1, 2):

$$\begin{cases} \sum_{i=1}^{M}\sum_{j=1}^{n_i}(\tau_{ij}^{(k)})^{(2-g)}(1-\tau_{ij}^{(k)})^{(g-1)}[\delta_{ij}S_g\gamma_g-h_g t_{ij}]=0; \\ \sum_{i=1}^{M}\sum_{j=1}^{n_i}(\tau_{ij}^{(k)})^{(2-g)}(1-\tau_{ij}^{(k)})^{(g-1)}[\delta_{ij}+(\delta_{ij}S_g\gamma_g-h_g t_{ij})(\log t_{ij}+\eta_g)]=0; \\ \sum_{i=1}^{M}\sum_{j=1}^{n_i}(\tau_{ij}^{(k)})^{(2-g)}(1-\tau_{ij}^{(k)})^{(g-1)}[\delta_{ij}S_g\gamma_g-h_g t_{ij}]x_{ij}^{T}=0; \\ \sum_{j=1}^{n_i}(\tau_{ij}^{(k)})^{(2-g)}(1-\tau_{ij}^{(k)})^{(g-1)}[\delta_{ij}S_g\gamma_g-h_g t_{ij}]-\frac{U_g}{\theta_g}=0. \end{cases}$$

Estimation for p has the closed-form equation:

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

(4) Iteration: Perform the E-step and M-step iteratively until the sequence $\hat{\psi}^{(k+1)}$ meets a convergence criterion, $|\hat{\psi}^{(k+1)} - \hat{\psi}^{(k)}| < r$ for some specified small *r*.

Appendix B. Estimation of variance components and asymptotic variances

The REMQL estimates of the variance components and the asymptotic variances of the fixed effect parameters are obtained in the manner of Ng *et al* [5]. Let Ω be the negative second derivative of the BLUP type log-likelihood (8) with respect to $(p, \beta_1, \beta_2, u_1, u_2)$. The matrix Ω has dimension $K \times K$, where $K = 1+2(v+M)$, v is the number of covariates and M is the number of clusters. Denote $\Delta = (\Delta_{ij})$ the inverse matrix of Ω, and be partitioned conformally to $p \mid \beta_1 \mid \beta_2 \mid u_1 \mid u_2$. The REMQL estimates of θ_1 and θ_2 are obtained as:

$$\begin{cases} \hat{\theta}_1 = \frac{1}{M}(tr\Delta_{44} + \hat{u}_1^T \hat{u}_1) \\ \hat{\theta}_2 = \frac{1}{M}(tr\Delta_{55} + \hat{u}_2^T \hat{u}_2) \end{cases}$$

with variance-covariance matrix of $(\hat{\theta}_1, \hat{\theta}_2)$ given by:

$$2\begin{pmatrix} \theta_1^{-2}(M - 2\theta_1^{-1}tr\Delta_{44}) + \theta_1^{-4}tr(\Delta_{44}^2) & \theta_1^{-2}\theta_2^{-2}tr(\Delta_{45}\Delta_{54}) \\ \theta_1^{-2}\theta_2^{-2}tr(\Delta_{45}\Delta_{54}) & \theta_2^{-2}(M - 2\theta_2^{-1}tr\Delta_{55}) + \theta_2^{-4}tr(\Delta_{55}^2) \end{pmatrix}^{-1}.$$

Furthermore, the asymptotic variance-covariance matrix of $(\hat{p}, \hat{\beta}_1, \hat{\beta}_2)$ is given by:

$$\begin{pmatrix} \Delta_{11} & \Delta_{12} & \Delta_{13} \\ \Delta_{21} & \Delta_{22} & \Delta_{23} \\ \Delta_{31} & \Delta_{32} & \Delta_{33} \end{pmatrix}.$$

*** Conflict of Interest Statement**

conflict of interest – none declared.

Table(s)

Table 1. Parameter estimates (standard error) from fitting two-component survival mixture models with random effects to the recurrent UTI data.

	Weibull proportional hazards model		Weibull AFT model		Log-logistic AFT model	
	1st component	2nd component	1st component	2nd component	1st component	2nd component
p	0.818*	0.182	0.744*	0.256	0.737*	0.263
	(0.036)		(0.037)		(0.036)	
θ	0.525*	0.521*	0.514*	0.377*	0.618*	1.175*
	(0.215)	(0.141)	(0.206)	(0.177)	(0.280)	(0.388)
γ	1.146	2.401	1.317	2.421	1.717	4.227
log λ	-6.159*	-11.610*	-7.321*	-12.361*	-9.103*	-19.419*
	(1.398)	(2.659)	(1.596)	(2.546)	(2.156)	(5.151)
Age (β_1)	-0.010	-0.044	-0.007	0.020	-0.008	0.0151
	(0.016)	(0.036)	(0.014)	(0.012)	(0.014)	(0.014)
Prior UTI (β_2)	0.973*	-0.030	0.834*	0.165	0.942*	0.320
	(0.276)	(0.356)	(0.237)	(0.194)	(0.252)	(0.226)

* p-value < 0.05

Figure(s)

Figure 1. Main panel of the graphical user interface.

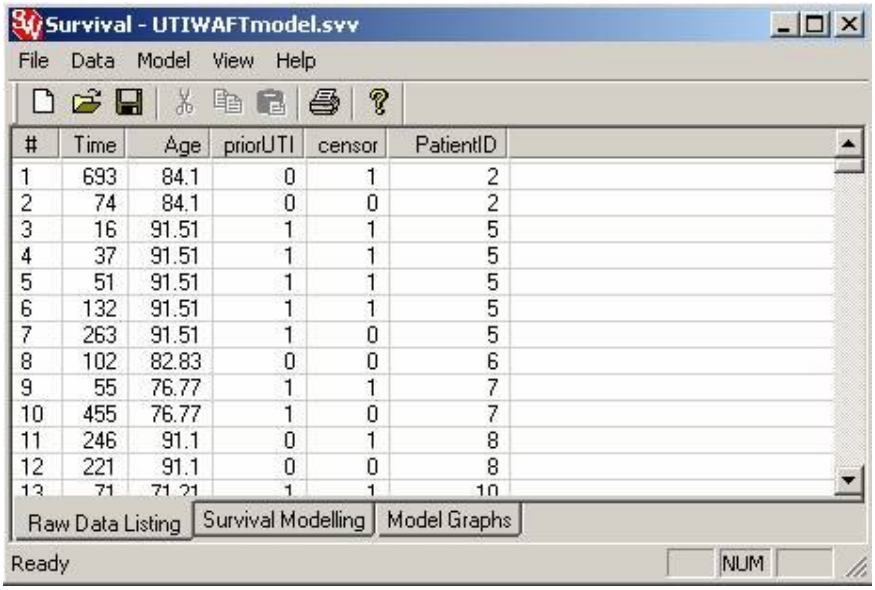


Figure 2. Variable selection window of the graphical user interface.

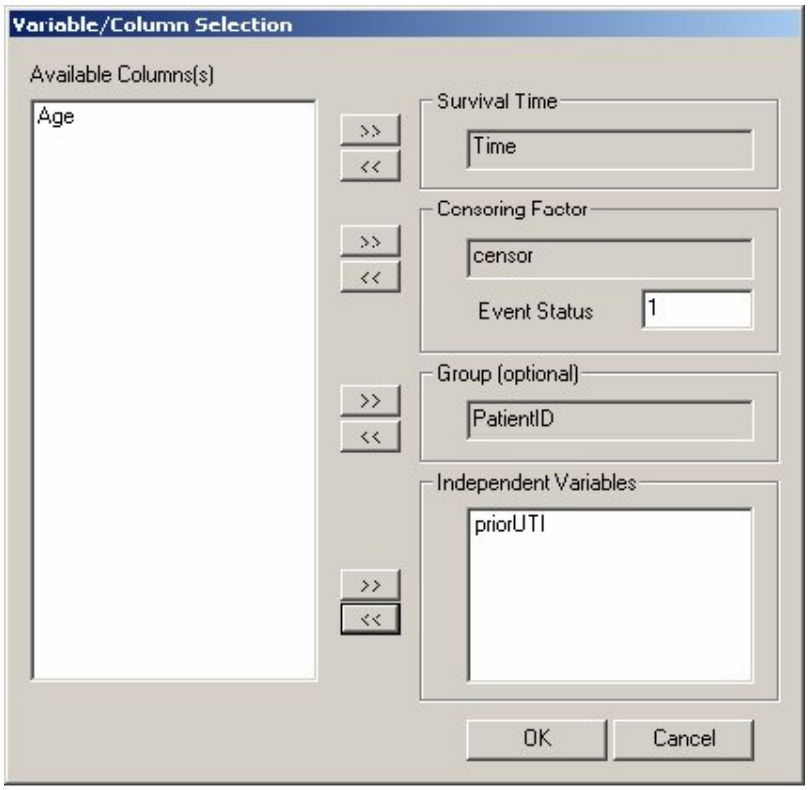


Figure 3. Model specification menu of the graphical user interface.

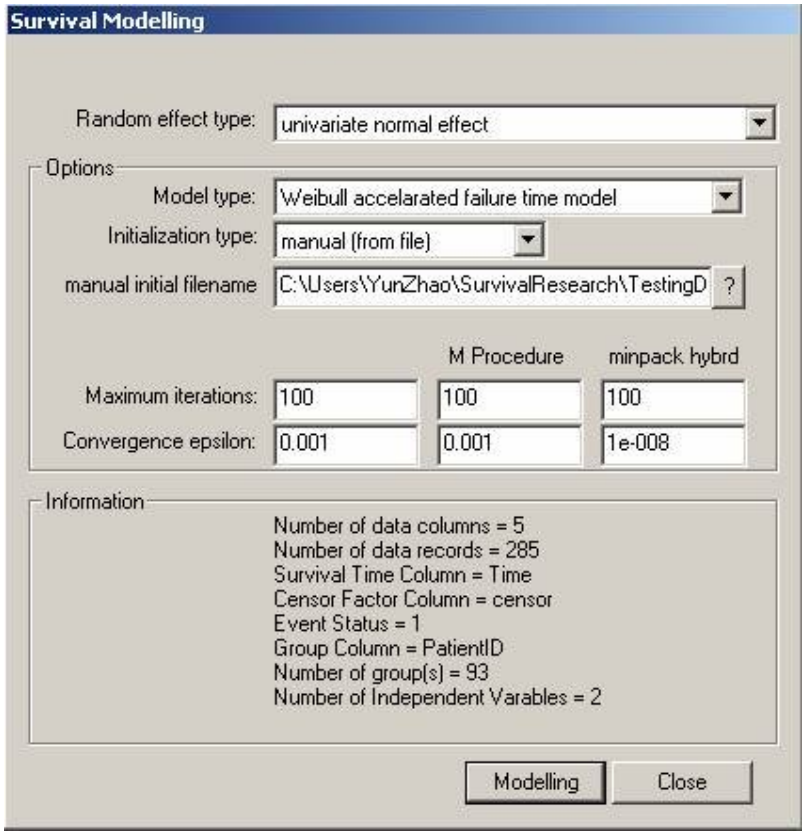


Figure 4. Example output after completion of Weibull AFT model fitting.

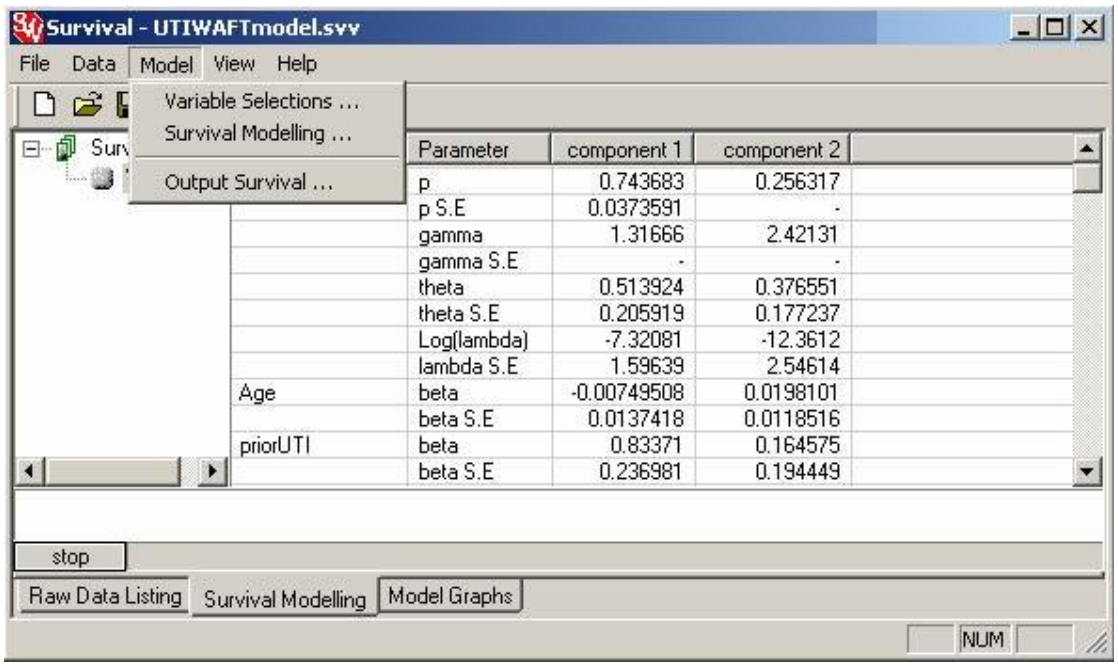


Figure 5. Plotting facility of the graphical user interface.

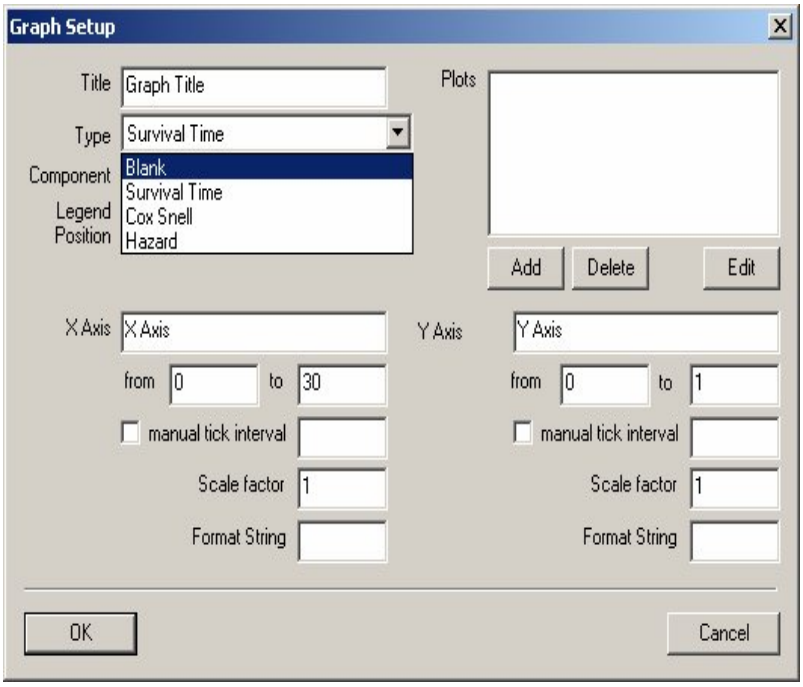


Figure 6. Hazard function of the first mixture component with respect to history of prior UTI, based on the Weibull AFT survival mixture model, for the recurrent UTI data.

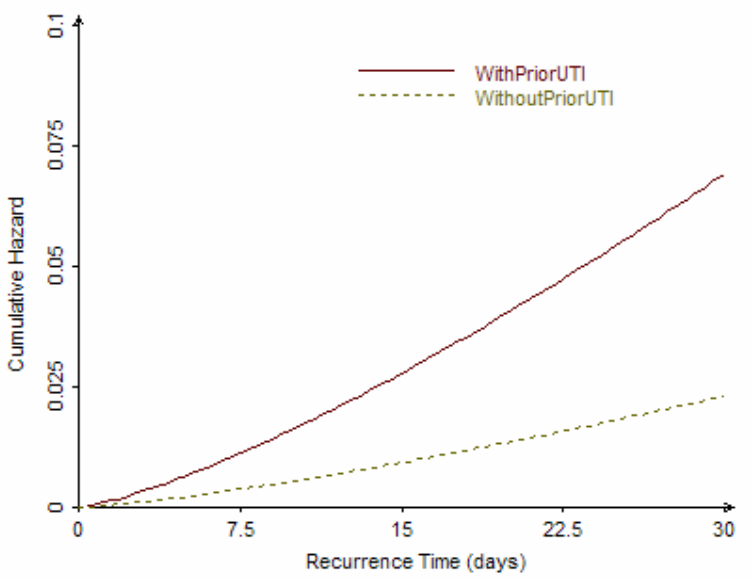
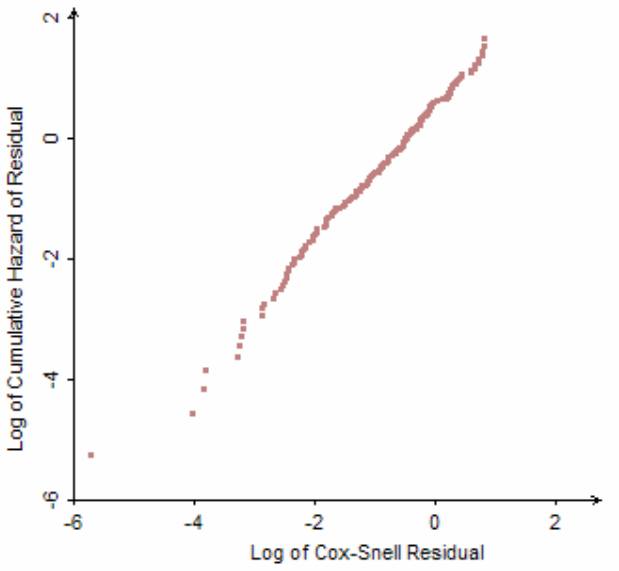


Figure 7. Cox-Snell residual plot for Weibull AFT survival mixture model fitted to the recurrent UTI data.



Author Biography

Professor Andy H. Lee received a BMath degree in 1982 from the University of Waterloo, Canada, a MMath degree in 1983 from the same university, and a PhD in 1988 from the Australian National University. He is currently Professor of Biostatistics at the School of Public Health, Curtin University of Technology, Australia. His multidisciplinary research includes statistical computing, nutritional epidemiology and chronic disease modelling.

Dr Yun Zhao received a BSc degree in applied mathematics from Hefei University of Technology, China, in 1982, a MSc degree in 2000 from Curtin University of Technology, Australia and a PhD (statistics) in 2004 from the same university. Since 2005, she has been a Research Fellow at the School of Public Health, Curtin University of Technology. Her research interests focus on mixture modelling and computer interface.

Professor Kelvin K.W. Yau was born in Hong Kong, in 1961. He received a BSc degree from the Chinese University of Hong Kong, in 1984, a MStats degree from the University of New South Wales, Australia, in 1993 and a PhD degree from the Australian National University, in 1996. Since 1995, he has been with the City University of Hong Kong, where he is Professor in Statistics. His research interests focus on biomedical data analysis and medical statistics.

Dr Shu-Kay Ng obtained his BSc degree from the University of Hong Kong and MScSt degree from the University of Queensland. After receiving his PhD degree in 1999, he was awarded research fellowship by the Australian Research Council. He was recently appointed as Senior Lecturer at Griffith University, where he provides biostatistical consultancy to medical staff. His current research projects are in the fields of machine learning, neural networks and survival analysis.