# Survival Mixture Modelling of Recurrent Infections

Andy H. Lee[1] Yun Zhao[1] Kelvin K.W. Yau[2] S.K. Ng[3]

[1]Department of Epidemiology and Biostatistics, School of Public Health, Curtin University of Technology, GPO Box U 1987, Perth, WA, 6845, AUSTRALIA.

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

[2]Department of Management Sciences, City University of Hong Kong, Tat Chee Avenue, Kowloon Tong, HONG KONG.

E-mail: mskyau@cityu.edu.hk

[3]School of Medicine, Logan Campus L03, Griffith University, University Drive, Meadowbrook, QLD, 4131, AUSTRALIA.

E-mail: s.ng@griffith.edu.au

**Keywords:** Accelerated failure time, Mixture model, Random effects, Recurrent infections

Recurrent infections data are commonly encountered in biomedical applications, 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 dependency of the recurrent observations, random effects are incorporated within the conditional hazard function. 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 correlated and heterogeneous survival data.

## 1 Introduction

Survival mixture models are often used to model heterogeneous failure time data in medical research (McLachlan and McGiffin (1994), De Angelis et al. (1999), Phillips et al. (2002)). Recently, a two-component Weibull survival mixture model was proposed by Ng et al. (2004) to analyse ischaemic stroke-specific survival time, in which patients are grouped into acute and chronic phases after the index stroke event. These 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. 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 of Wei (1992), 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.

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 dependency of recurrent observations has not been addressed satisfactorily in the literature. Survival frailty models are mainly 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. Therefore, this paper aims to present a unified and flexible approach of modelling recurrent infections data by a finite mixture of survival distributions incorporating random effects.

## 2 Two-component survival mixture models with random effects

Let $Y_{ij}$ denote the $j$th recurrent time ($j$ = 1, 2, …, $n_i$) within individual $i$ ($i$ = 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 $\delta_{ij}$ 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 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 $g^{\text{th}}$ component ($g$ = 1, 2), respectively. With the concomitant information $x_{ij}$, effects of covariates 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 of Yau and Ng (2001).

Under the proportional hazards assumption, the conditional hazard function for the $g^{\text{th}}$ 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 $g^{\text{th}}$ component is

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

which may be considered as a Weibull distribution with scale $\lambda_g \exp(\gamma_g \eta_g(x_{ij}))$ and shape parameter $\gamma_g$. 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, $-\infty < \lambda_g < \infty, \text{and } \gamma_g > 0$ are unknown parameters.

For both proportional hazards and AFT settings, an unobserved random effect term can be introduced in each conditional hazard function to explain the variability shared by the recurrent observations, in the manner of Wang et al. (2007). Specifically,

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

where $\beta_g$ is the vector of regression coefficients. Without loss of generality, the random subject 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 (Wang et al. (2007)). For given initial values of $\theta_g$, the BLUP estimator of $\psi$ 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 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 $\psi$ is obtained as a solution of the equation $\partial l / \partial \psi = 0$, which can be solved via an EM algorithm (Ng et al. (2004), Wang et al. (2007)). The REMQL estimates of

the variance components $\theta_1$ and $\theta_2$ are then obtained by maximizing the restricted log-likelihood function. Details of the EM estimation procedure and derivations for the REMQL estimates and asymptotic variances are omitted for brevity, but are available upon request.

## 3 Software and model assessment

To implement the modelling methodology and estimation procedure described in Section 2, a graphical user interface is developed and coded using the Microsoft visual C++ scientific language, with adaptations taken from numerical library subroutines of Press and Vetterling (1992). 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 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 $\log(e_{ij})$.

## 4 Simulation study

A small scale simulation study is conducted to investigate the properties of the REMQL estimators under the two-component Weibull AFT survival mixture model (5) in finite sample settings. Following the simulation design for the Weibull proportional hazards model of Ng et al. (2004), we assume $M$ = 20 subjects with $n_i$ = 25 observations each. A single continuous covariate $X_{ij}$ ($i$ =1,…,20, $j$ = 1,…,25) is generated from a standard normal distribution. Each individual has a probability $p$ or (1–$p$) belonging to the first or second component, respectively. Hence, for the $g^{\text{th}}$ ($g$ = 1, 2) component, the survival time $t_{ij}$ is generated based on the Weibull conditional probability density function $f_g(t_{ij}, X_{ij})$, with $\eta_g(X_{ij}) = X_{ij}^{\ T}\beta_g + U_{gi}$, where $U_{gi} \sim N(0, \theta_g).$ For the realization of censorship, let $C$ be the fixed censoring time. Instead of observing the survival time of interest $t_{ij}$, we observe ($Y_{ij}$, $\delta_{ij}$), where $Y_{ij}$ = min($t_{ij}$, $C$) and

$$\delta_{ij} = \begin{cases} 1, \text{if } t_{ij} \le C, \\ 0, \text{if } t_{ij} > C. \end{cases}$$

Without loss of generality, $C$ is fixed to be 1000 and 500 replications are performed. The true values for parameters ($p$, $\lambda$, $\gamma$, $\beta$) of the first and second mixture component are set as (0.1, 0.05, 1.5, 0.5) and (0.9, 0.01, 0.5, -0.5), respectively. Two settings are considered: ($\theta_1$, $\theta_2$) = (1, 1) and ($\theta_1$, $\theta_2$) = (0.5, 0.5).

Results are presented in Table 1. In general, parameters $p$, $\theta_1$ and $\beta_1$ are well estimated, but the biases for the estimates of $\beta_2$ when $\boldsymbol{\theta}$ = 1 are observed within the range of ±18%, whereas $\theta_2$ is moderately biased. The observed discrepancies appear to be reasonable in terms of mean squared error (MSE) criterion and the sampling standard error (SE) of the estimates.

Table 1: Simulation results for Weibull AFT model.

| Simulation 1 | True value | Mean | MSE | Sampling SE | Bias |
|---|---|---|---|---|---|
| $p$ | 0.1 | 0.105 | 0.022 | 0.022 | 0.005 |
| $\theta_1$ | 1 | 0.954 | 0.214 | 0.209 | -0.046 |
| $\theta_2$ | 1 | 1.155 | 0.241 | 0.185 | 0.155 |
| $\beta_1$ | 0.5 | 0.548 | 0.182 | 0.176 | 0.048 |
| $\beta_2$ | -0.5 | -0.686 | 0.256 | 0.178 | -0.186 |
| Simulation 2 | True value | Mean | MSE | Sampling SE | Bias |
| $p$ | 0.1 | 0.104 | 0.018 | 0.017 | 0.004 |
| $\theta_1$ | 0.5 | 0.431 | 0.119 | 0.097 | -0.069 |
| $\theta_2$ | 0.5 | 0.606 | 0.129 | 0.074 | 0.106 |
| $\beta_1$ | 0.5 | 0.521 | 0.103 | 0.101 | 0.021 |
| $\beta_2$ | -0.5 | -0.548 | 0.120 | 0.110 | -0.048 |

## 5 Application to recurrent urinary tract infections

Urinary tract infection (UTI) is a common bacterial infection in elderly women aged 60 years and above, and one in four of these women will develop a recurrence (Franco (2005)). A retrospective cohort study on recurrent UTI was conducted among elderly women in residential aged-care facilities (Xiang et al. (2006)). 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.

With covariates age and history of prior UTI taken at baseline, results from fitting the survival mixture models are presented in Table 2. 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.

Table 2: 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) | |
| $\theta$ | 0.525* | 0.521* | 0.514* | 0.377* | 0.618* | 1.175* |
| | (0.215) | (0.141) | (0.206) | (0.177) | (0.280) | (0.388) |
| $\gamma$ | 1.146 | 2.401 | 1.317 | 2.421 | 1.717 | 4.227 |
| log $\lambda$ | -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 ($\beta_1$) | -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 ($\beta_2$) | 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

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 1, 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.

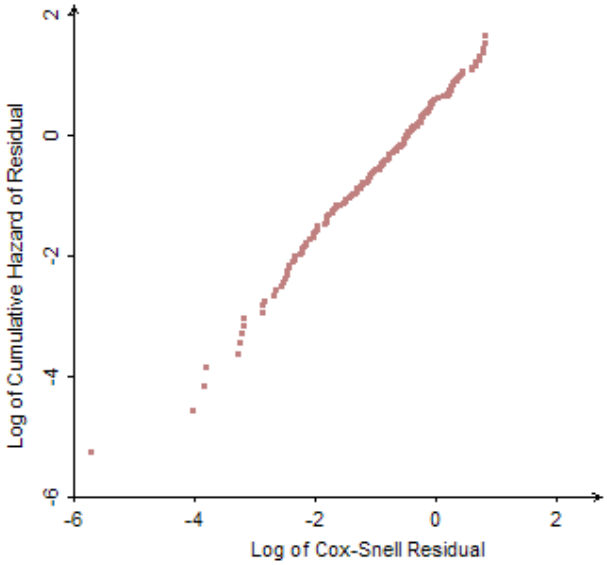


Figure 1: Cox-Snell residual plot for Weibull AFT model fitted to the recurrent UTI data.

## 6 Conclusion

An integrated and flexible approach of modelling recurrent infections is proposed, 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. A small scale simulation study found that the REMQL estimators for the Weibull AFT model behave reasonably well in finite sampling conditions. Moreover, 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.


## Acknowledgement

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 upon request.


## References


De Angelis, R., Capocaccia, R., Hakulinen, T., Sodeman, B. and Verdecchia, A. (1999). Mixture models for cancer survival analysis: application to population-based data with covariates. *Statistics in Medicine*, 18: 441-454.

Franco, A.V. (2005). Recurrent urinary tract infections. *Best Practice & Research. Clinical Obstetrics & Gynaecology*, 19: 861-873.

McLachlan, G.J. and McGiffin, D.C. (1994). On the role of finite mixture models in survival analysis. *Statistical Methods in Medical Research*, 3: 211-226.

Ng, S.K., McLachlan, G.J., Yau, K.K.W. and Lee, A.H. (2004). Modelling the distribution of ischaemic stroke-specific survival time using an EM-based mixture approach with random effects adjustment. *Statistics in Medicine*, 23: 2729-2744.

Phillips, N., Coldman, A. and McBride, M.L. (2002). Estimating cancer prevalence using mixture models for cancer survival. *Statistics in Medicine*, 21: 1257-1270.

Press, W.H. and Vetterling, W.T. (1992). *Numerical Recipes in C: The Art of Scientific Computing, 2nd edition*. Cambridge University Press.

Wang, K., Yau, K.K.W., Lee, A.H. and McLachlan, G.J. (2007). Multilevel survival modeling of recurrent urinary tract infections. *Computer Methods and Programs in Biomedicine*, 87: 225-229.

Wei, L.J. (1992). The accelerated failure time model: a useful alternative to the Cox regression in survival analysis (with discussion). *Statistics in Medicine*, 11: 1871-1879.

Xiang, L., Lee, A.H., Yau, K.K.W. and McLachlan, G.J. (2006). A score test for zero-inflation in correlated count data. *Statistics in Medicine*, 25: 1660-1671.

Yau, K.K.W. and Ng, S.K. (2001). Long-term survivor mixture model with random effects: Application to a multicentre clinical trial of carcinoma. *Statistics in Medicine*, 20: 1591-1607.