Journal of Statistical Research *ISSN 0256 - 422 X*
2005, Vol. 39, No. 2, pp. 135-153
Bangladesh

AN APPROXIMATE METHOD FOR A FRAILTY MODEL IN THE PRESENCE OF AN IMMUNE PROPORTION [1]

B. Hasan
National Cancer Institute of Canada-Clinical Trials Group
Queen's University, Kingston, Ontario, K7L 3N6, Canada.
Email: bhasan@ctg.queensu.ca

R.S. Singh
Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario, N1G 2W1, Canada.
Email: rssingh@uoguelph.ca

H. Pesotan
Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario, N1G 2W1, Canada.
Email: hpesotan@uougelph.ca

SUMMARY

This article considers an extension of the existing survival model with an immune proportion known as a *latent data* (LD) model. Random effects are introduced in this LD model. A generalized linear mixed model using a penalized quasi likelihood approach for the parameter estimates is proposed. The model enables the prediction of the random effect and retains the proportional hazard property of the LD model. Application of the method is carried out on two real data sets. A simulation study is conducted to evaluate the model's performance. Two different types of censoring are considered. The results show that the estimates have relatively small bias in all cases and the method works equally well in both the random and fixed censoring cases.

Keywords and phrases: Surviving fraction; Frailty model; Simulation study; Latent variable; Penalized quasi likelihood; Generalized linear mixed model; Censoring proportion.

AMS Classification: Place Classification here. Leave as is, if there is no classification

1 Introduction

An immune, or a long-term survivor, is defined as an object which is not subject to the event of interest (Maller and Zhou, 1996). Hence, the object is immortal with regard to the event of interest under study. Consequently, the lifetime of the immune object will be infinite, at least theoretically, under the event of interest being investigated. These immune observations form an immune proportion in the data and are usually visible in the Kaplan-Meier plot of the survival function. Another possible indication of their presence is the existence of some large values of censored data. The immune proportion is also known as the *surviving proportion/fraction* or *cure proportion/fraction*. Modeling the survival data using a "regular" survival analysis (which ignores immunes) in the presence of immunes might lead to drawing misleading conclusions from the study.

[1]The research is based on a part of the first author's Ph.D. thesis, and is supported, in part, by the Natural Sciences and Engineering Research Council of Canada

Recently, Yakovlev et al. (1993) introduced a new model to deal with survival data with an immune proportion. Chen et al.(1999) have studied this model in a Bayesian context. This approach proceeds by introducing a latent data variable N, the number of clonogens(carcinogenic cells which will propagate tumors in detectable forms), which is assumed to have a Poisson distribution. Observe that the variable N is not observable and will be referred to as a *latent data* variable. The survival function of the patients under this model is given below:

$$S_p(t) = e^{-\theta} e^{\theta S(t)} = e^{-\theta F(t)}, \tag{1.1}$$

where $S(t)$ is a proper survival function and $F(t) = 1 - S(t)$. The corresponding hazard function is given by:

$$h_p(t) = \theta f(t), \tag{1.2}$$

where θ is the mean of the Poisson distribution specified for N and $f(t) = F'(t)$ is a probability distribution function (pdf) of the *progression times* for carcinogenic cells. This model implies that immune proportion consists of patients who have $N = 0$.

This model has more biological meaning than the corresponding classical model discussed in Maller and Zhou (1996), particularly in modeling survival data for cancer diseases, as pointed out by Yakovlev et al.(1993) and Chen et al.(1999). We will refer to this model as a *latent data* (LD) model or a *latent variable* model.

In some circumstances, there is more than one measurement taken from the same subject, leading to what is known as correlated data. This type of data has a heterogeneity factor built in. Heterogeneity could also occur as a result of natural group existence. For example, a centre effect exists if multiple centres are used in clinical trials, where more than one patient is treated in each of the centres. Heterogeneity could result in a misleading analysis if it is not incorporated in the model. The possible impact of excluding the heterogeneity factor from the model when it exists is discussed in Section 2.

In survival analysis, the model dealing with heterogeneity is known as a *frailty model*. There exist several approaches to handle heterogeneity when survival data contain immune observations. The common approach is by introducing random effects into the survival function of the mixture distribution approach. This approach, for instance, can be found in Yau and Ng (2001). They used the *generalized linear mixed model* (GLMM) approach developed earlier by McGilchrist (1994) for parameter estimates of their model. In addition, there is a compound Poisson distribution approach introduced by Aalen (1992). The compound distribution approach assumes that the frailty random variable follows a compound Poisson distribution, that allows for a point mass at zero corresponding to the cured fraction. The survivor function can then be found by integrating out the frailty random variable through the use of a Laplace transform.

The aim of this article is to provide an alternative model for handling heterogeneity when an immune proportion exists. It is achieved by extending the model given in (1.1). The extension proceeds by introducing a random effect for the objects under study, to take into account the heterogeneity of the objects. The random effect in the model acts multiplicatively on the hazard function, as commonly found in the frailty model.

In addition, a simple estimation method for the model is proposed. It is shown that the random effect can be handled by incorporating it into the Poisson generalized linear model, which is one part of the likelihood in Yakovlev's model. Because of the existence of random components, this part becomes a Poisson GLMM. The random effects can be assumed to follow a certain type of distribution. In this article, the random effect u_i discussed later in Section 2 is assumed to follow a common normal distribution. However, since the estimation method for the Poisson GLMM is based on an approximate method known as the *Penalized Quasi Likelihood* (PQL), only the first and the second moment assumptions for the random effects are needed.

This article is organized as follows: Section 2 presents the extension of Yakovlev's latent variable model to include random effects. Likelihood for this extended model along with an estimation procedure are discussed. Applications of the methods to real data sets are discussed in section 3. Section 4 provides a simulation study to assess the model. Section 5 mentions some directions for possible further research.

2 Frailty Model with an Immune Proportion

Frailty is a concept in survival analysis which resembles random effects in linear or generalized mixed linear models. It represents unobserved heterogeneity across subjects. This unobserved heterogeneity may arise from measures of multivariate or correlated failure times. At the individual level, it comes from multiple failure time measurements, such as in sequences of asthmatic attacks, infection episodes, tumor diagnosis, tumor recurrences or bleeding incidents (Prentice et al., 1981). The correlated failure times could also occur at the group level where cluster structure exists, such as animals nested within a litter, children nested within a family, or patients nested within clinical centres in multi-centre clinical trials. In this paper we will develop a frailty model which is applicable to multi-centre clinical trial framework.

Frailty in a survival analysis model is typically introduced through the hazard function. Given the basic hazard function $h_0(t)$, the individual hazard $h(t|z)$ conditioned on the frailty Z is assumed to take the following form

$$h(t|Z) = Z\ h_0(t). \tag{2.1}$$

Here Z is a positive random variable describing heterogeneity of an individual, with mean, $E(Z)$, equal to 1 for identifiability reasons. If Z in (2.1) is individual specific, then the model is known as a $frailty$ model. If Z is group specific, then the model is known as a $shared\ frailty$ model.

Lancaster (1979, 1990) provides an argument for the use of frailty, or error term, in a regression model of survival analysis. He explains that the population hazard is a bent-down version of a basic hazard function. In other words, this hazard function falls faster or increases slower than the basic hazard function (with no frailty). This argument is similar to the "selection effects" of Aalen (1998).

2.1 Approximate Method for Frailty Model in the Presence of an Immune Proportion

There are two commonly used methods to handle random effects in the frailty models with an immune proportion, the marginal and the conditional approaches. The first approach focuses on the change in the population-average hazard function of the population when the frailty random variable is integrated out. The second method, which avoids integrating out the random effects, is more flexible in handling high-dimensional random effects. This method does not emphasize on the form of the hazard function, since in this approach the random effects are not integrated out. Rather, it focuses on the prediction of group specific effects shared within observations of each individual, centre or family. The predicted values of the random effects will give information about the effects of individuals or groups on the population hazard.

The development of the frailty model for survival data with an immune proportion proposed in this article proceeds by extending the approach of Yakovlev et al.(1993) and Chen at al.(1999) and Chen and Ibrahim (2001) within the multi-centre clinical trials framework. More specifically, the model formulation assumes that there are n clinical centres treating cancer patients. Suppose that in the i^{th} centre there are l_i patients treated. In each centre time zero is the time when the study begins. Patients will receive treatment randomly based on a standard protocol and then are monitored until the end of the study. From the beginning to the end of the study, a survival time

or a censoring time will be observed. A survival time is recorded if a patient dies within the study period. A censoring time will be observed if a patient is alive till the end of the study period. Provided that the study period is long enough, we will expect that some of the patients will die from cancer. We assume that some patients are cured by the end of the study. It is also assumed that due to variation in patient treatment from any one centre to another during the period of study that a centre effect is present, and it is important to introduce this effect into the model used to analyze the data obtained from such a multi-centre clinical trial. This source of variation will be treated as a random effect, i.e. we assume that the centres involved in the study are a random sample from a population of centres. Below, we construct the likelihood for our proposed frailty model with an immune proportion. A similar likelihood for a model without random effects can be found in Yakovlev et al.(1993) and Chen et al.(1999) and Chen and Ibrahim (2001).

In the ensuing development we refer to the models (1.1) and (1.2) presented in the introduction. Moreover, the parameter θ and the random variable N referred to here are those presented in the Section 1. The likelihood for the data for the LD model is given as follows:

$$L(\lambda,\theta|\{(t_i,d_i)\}) = \left[\prod_{i=1}^{n}\{S(t_i|\lambda)\}^{N_i-d_i}\{N_i f(t_i|\lambda)\}^{d_i}\right] exp\left\{\sum_{i=1}^{n} N_i \log(\theta) - \sum_{i=1}^{n}\log(N_i!) - n\theta\right\}, \quad (2.2)$$

where t_i is the observed lifetime, $f(t|\lambda)$ is a proper probability distribution function, $S(t|\lambda)$ is a proper survival function and d_i is a censoring indicator. Details discussion of this model can be found in Yakovlev and Tsodikov (1996) and Chen et al. (1999) among others.

We assume that the type of censoring in this study is either a type I or a random censoring. The structure of the likelihood for the data resembles the likelihood in (2.2), except that now a random effect Z_i is introduced into the mean of the Poisson random variable N for each centre. Therefore, conditional on the random variable Z_i, the mean of the Poisson distribution becomes $Z_i\theta$.

Sometimes it is more convenient to work with a transformed version of a random effect Z_i than to consider it in its original form. In this article the log transformation form of the random effect, namely $u_i = \log Z_i$ is adopted. This transformation enables a direct inclusion of the random effects into the Poisson log-linear model.

Assuming u_i to be iid with pdf $g(u|\sigma^*)$, the likelihood function can be expressed as:

$$\begin{aligned} L(\lambda,\theta,\sigma^*|u_i,N_{ij},\{(t_i,d_i)\}) &= \prod_{i=1}^{n}\prod_{j_i=1}^{l_i} S(t_{ij}|\lambda)^{N_{ij}-d_{ij}}(N_{ij}f(t_{ij}|\lambda))^{d_{ij}} \\ exp\{\sum_{i=1}^{n}(\sum_{j_i=1}^{l_i}(N_{ij}(\log\theta+u_i) - \log(N_{ij}!) &- (\log\theta+u_i)) + \log g(u_i))\} \\ &= L_1(\lambda|t)L_2(N|u)L_3(u), \end{aligned} \quad (2.3)$$

where σ^{*2} is the variance of the random effect u_i.

The second part of the likelihood in (2.3), namely L_2 and L_3, is an extended likelihood (Pawitan, 2001) which has the form of a Poisson regression with random effects u_i's. If the log link $\eta_{ij} = \log\theta_{ij}$ is taken then the conditional mean of the Poisson random variables can be expressed as $\exp(\eta + u_i)$. When covariates are included, they enter the model through the link function by placing $\eta = \boldsymbol{x'\beta}$. The augmented likelihood function can then be written as:

$$L(\lambda,\beta,\sigma^*|\{(t_i,d_i)\}) = \prod_{i=1}^{n}\prod_{j_i=1}^{l_i}\{S(y_{ij}|\lambda)\}^{N_{ij}-d_{ij}}\{N_{ij}f(y_{ij}|\lambda)\}^{d_{ij}}$$

$$exp\left(\sum_{i=1}^{n}\left[\sum_{j_i=1}^{l_i}\left\{N_{ij}(\boldsymbol{x}'_{ij}\boldsymbol{\beta}+u_i) - \log(N_{ij}!) - (\boldsymbol{x}'_{ij}\boldsymbol{\beta}+u_i)\right\} + \log g(u_i)\right]\right). \quad (2.4)$$

The hazard function corresponding to the likelihood in (2.3) is

$$h_p(t_{ij}|u_i) = exp(u_i)\ \theta_{ij} f(t_{ij}), \tag{2.5}$$

where θ_{ij} is the mean of Poisson random variable N_{ij}, and f is a proper pdf specified for the progression time. Observe that the hazard function (2.5) with the frailty $Z_i = exp(u_i)$ removed from it is exactly the hazard function (1.2) studied in Yakovlev et al.(1993). Unlike the traditional frailty model where assumptions are made on the random variable Z_i's, here, assumptions are imposed directly on the transformed form of the random variable u_i. That is, u_i is assumed to have mean equal to 0, and variance σ^{*2}. Models which have this hazard form are known as shared frailty hazard models as discussed in the previous subsection, where the multiplicative random effect Z_i is known as the frailty. Expression (2.5) reflects the contribution of the immunes and the heterogeneity in the hazard function. The hazard function of the survival data with immunes has no less than one peak (Yakovlev and Tsodikov, 1996). It should be emphasized that although the model (2.5) is developed here in the context of a multi-centre clinical trial framework it can also be used in other contexts, for example, a situation in which the n centres are replaced by n individuals and multiple observations are made on each individual. This model can be extended further to a multi-level random effects model such as a study in which multiple observations are observed on an individual in multi-centre clinical trials.

The second part of the likelihood in (2.3) has the form of a Poisson log-linear regression with random effects u_i's, where the responses N_{ij} are latent data and as such are not observed. The estimation procedure can follow the procedure for the regular cure rate model as in Chen et al.(1999). The only difference is that the Poisson regression has random effects in the model (2.5), as opposed to a regular Poisson regression in the model (1.2). The Poisson regression with random effects is solved within the framework of GLMM's, as discussed in the next section.

There are some recent closely related developments for the model with a hazard function as found in (2.5). Ibrahim et al. (2001) discussed this model in the context of bivariate multivariate cure rate, where the event time consists of two different readings from the same person, namely t_{i1} (the time to first infection) and t_{i2} (the time to second infection). They then specified N_k, k=1,2 to follow the Poisson distribution with mean $Z\theta_k$. The differences between Ibrahim et al.'s model and the model developed in this article lies primarily in the fact that their model assumes different Poisson means for the time to event t_{i1} and t_{i2}. In this article, this assumption is not necessary. The contribution of multiple events in each object, or the group effects, is handled through the inclusion of random effects (u) in the model. In addition, Ibrahim et al. proposed a Bayesian approach for their model but the approach used in this article is the GLMM approach within the frequentist framework.

When covariates are introduced into the model, the linear predictor of the link function for the Poisson generalized linear model becomes $\exp(\eta) = \theta = \exp(X\beta)$. The hazard function for the model can then be written as

$$h_p(t_{ij}) = \ \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + u_i) f(t_{ij}).$$

The random effect u has mean 0 and variance σ^{*2}. This formulation guarantees that the random effect $Z_i = exp(u_i)$, which acts multiplicatively on the hazard function, is positive.

2.2 Estimation Procedure

In the likelihood functions (2.3) and (2.4), the parameters that have to be estimated are the parametric vector $\boldsymbol{\lambda}$, the vector $\boldsymbol{\beta}$ of regression parameters and σ^{*2} the variance of the random effects u_i's. The vector $\boldsymbol{\lambda}$ appears when a choice of distribution is made on the response variable which are the progression times. The vector $\boldsymbol{\beta}$ appears when a choice of covariates is made in

the model. In this article we impose the lognormal distribution $f(t|\boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = (\mu, \sigma)$, on the progression times.

The process we will use in doing this estimation is a two step method known as the expectation maximization (EM) algorithm. We refer to Tanner (1996) for a good discussion about this method.

We start the EM algorithm by specifying initial values $(N_{ij})_0$, μ_0 and σ_0. Note that once $(N_{ij})_0$ have been specified, it implies that $\boldsymbol{\beta}_0$ and σ_0^{*2} are also specified. At the $(k+1)^{th}$ expectation step ($k = 0, 1, 2, \cdots$), $N_{ij}^{(k)}$ is replaced by its expected value

$$\left(N_{ij}^{(k+1)}|u_i\right) = E\left(N_{ij}^{(k)}|u_i\right) \quad \equiv \quad \theta_{ij}^{(k)} S(s_i, \mu^{(k)}) + d_i,$$

where $\theta_{ij} = (\boldsymbol{x}'_{ij}\boldsymbol{\beta} + u_i)$, d_{ij} is the censoring indicator of the j^{th}-patient in the ith centre and $S(.)$ is the survival function. Therefore, the $(k+1)^{th}$ expectation step of the EM algorithm has the following form:

$$E\{l(\theta, \mu, \sigma|\{s_i, d_i\}, u_i, \sigma^*, \theta^{(k)}, \mu^{(k)}, \sigma^{(k)})\} = \sum_{i=1}^{n}\sum_{j_i=1}^{l_i}\{N_{ij}^{(k+1)} - d_{ij}\}\log\Psi\left(\frac{s_{ij}-\mu}{\sigma}\right) - r\log\sigma -$$

$$\frac{1}{2\sigma^2}\sum_{i,j\in D}(s_{ij}-\mu)^2 + \sum_{i=1}^{n}\left[\sum_{j_i=1}^{l_i}\{N_{ij}^{k+1}(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + u_i) - \log(N_{ij}^{k+1}!) - (\boldsymbol{x}'_{ij}\boldsymbol{\beta} + u_i)\} + \log g(u_i)\right].$$

The maximization step consists of two parts: (i) estimating the parameters μ and σ and (ii) estimating the regression parametric vector $\boldsymbol{\beta}$ and the variance σ^{*2} of the random effects u_i. For (i) we use the Newton-Raphson algorithm (see Hasan et al., 2005) and for (ii) we use the penalized quasi likelihood (PQL) method. For a discussion of the PQL method, see Schall (1991) and Breslow and Clayton (1993). Venables and Ripley (2002) have recently written a computer program on the PQL method. We continue the EM algorithm and stop when a pre-set convergence criterion has been met.

As discussed earlier, the expressions in (2.3) and (2.4) reflect two components in the model which can be separately estimated. The first part of the likelihood dealing with the estimation of the parameters in $f(t)$ was discussed for instance in Chen and Ibrahim (2001). When the lognormal distribution is specified for $f(t)$, the estimation procedure is discussed in Hasan et al.(2005).

2.3 Inferences and Hypothesis Tests

As commonly known, standard errors are not available for the estimates produced through the EM algorithm. Therefore, calculating the standard errors of the estimates requires that the latent variables are eliminated. For the survival model with an immune proportion, Yakovlev and Tsodikov (1996) have shown that the latent data N can be integrated out so that the score function and its variance can be derived from the the likelihood. Therefore, inferences can be made based on the likelihood method. For the cure rate frailty model discussed here however, the standard likelihood model is not applicable because of the existence of random effects.

One way to obtain the standard deviations of the estimates is by bootstrapping. This method is straightforward and particularly useful when there is no direct method available to obtain the information matrix of the estimates such as in the EM algorithm. One issue in the use of the bootstrap method is the number of bootstrap simulations needed to derive the standard deviation estimates. To have a very large number of bootstrap replicates is ideal, but is often not feasible due to computational restrictions. This is particularly true for the method developed in this article, where intensive calculations are performed to do the estimation. However, for the purpose of obtaining standard deviations, the number of bootstrap replication, R, can be as few as 50 (Efron and Thibshirani, 1993). In this article, after trying several choices of R (i.e.,

$R = 50, 100, 200,$ and $R = 400$), it is found that $R = 100$ gives a bootstrap standard deviation just as good as $R = 400$ does. Therefore, the number of bootstrap replications, $R = 100$, is chosen.

3 Applications

In this section, the method developed in Section 2 will be applied to real data sets. There are two data sets considered. The first one is the carcinoma data found in Kalbfleisch and Prentice (2003). This data set is suitable for the method developed since it involves a number of institutions in which patients were treated. Therefore, it is a multi-centre clinical trial where the centre can be accounted for as a random effect. Another data set is the bladder cancer data listed in Wei et al.(1989). This data set has multiple records of time to reoccurrences of bladder cancer on each person. Since the measurements within each person are not independent, the frailty model can be applied since a person specific random effect is present. There are two algorithms used to fit Schall (1991)'s PQL proposal for GLMM. The first algorithm is the *glmmPQL* written by Venables and Ripley (2002) and the second one is *reglm* written by Schmidt (http://www.statsci.org/s/reglm.html). Both of them were written as $S - Plus$ 2000 functions and are claimed to have been derived from Schall's proposal by their writers. The estimation for the variance component σ^* employs restricted maximum likelihood (REML) in both algorithms.

Note that the estimates for immune proportions are derived from the relationship $1 - p = exp(-exp(\eta))$, where η is the intercept of the regression parameters if no covariates are included in the model. Throughout this section and the rest of this article therefore, the inferences for the immune proportions in the event of no covariates in the model are carried out through the inferences for the intercepts of the regression parameters. However, the immune proportions will still be reported to give a clear view of their presence, even though they are not tested for significance.

3.1 Carcinoma Clinical Trial

The data provided in Kalbfleisch and Prentice (2003) is a subset of results from a clinical trial studied by the Radiation Therapy Oncology Group in the United States. The data consists of survival records of patients with squamous carcinoma from three sites in the oropharynx, collected with the participation of six institutions. Patients in the study were randomly assigned to one of two treatment groups, namely radiation therapy alone and radiation therapy together with a chemotherapeutic agent. The main interest is in comparing these two treatments with respect to the survival time of the patients. In addition, there were various covariates recorded in the study, which would be expected to relate to the patients' survival. Six covariates were specified in the study, namely sex, tumor stage classification (T-stage), lymph node stage classification (N-stage), age, the degree of differentiation of the tumor (grade), and the general condition of the functional capacity of the patient at the time of diagnosis. Hence, in addition to comparing the treatments, investigating to what extent risk factors among the covariates related to the patient's survival were also of interest.

The treatment protocol was applied to the patients during a 90 day period. When the treatment was done, patients received medical care by the participating institutions. No restrictions, except those specifically required by the study, were placed on the past 90 day care. Variability as a result of differences in the institutional care post the 90 days of treatment is therefore inevitable. It is then plausible to consider the institution effect in the analysis of the data. In Kalbfleisch and Prentice (2003), the institution effect was considered as a fixed effect. That is, inferences from the analysis may be valid only for those institutions which participated in the clinical trial. A random effect model may be adopted to give further insights into the variability of the institution

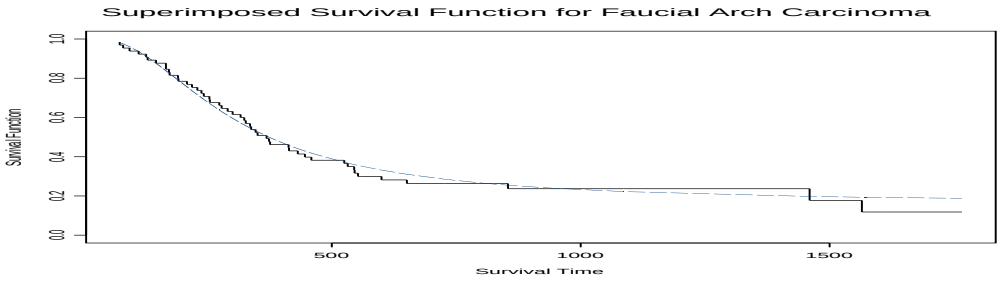

Figure 1: Survival Function Estimated by Kaplan-Meier and Latent Variable Methods

effect. In that case, the participating institutions are considered as a random sample from the institutions of interest.

In this article, for illustration purposes, only a subset of the data will be used, namely the faucial site subset. The only covariate included is the T-stage. The main interest is to see whether less severe tumor stage increases the cure probability of patients. For this purpose, T-stage is reclassified so that only two categories are included, namely the massive tumor with extension to adjoining tissue and the non-massive tumor. In addition, it is assumed that there exists an unobservable random effect u in the data.

There were 65 patients recorded for the subset of the data chosen. They are from 6 participating institutions. Out of the 65 patients, 50 died of the disease and the rest were either still alive at the end of the follow-up period or lost to follow-up. Patients were lost to follow-up if they moved or transferred to an institution not participating in the study. The total number of censored observations were 15 (23.1 percent). The data is fitted using the long-term frailty model, $h_p(t_{ij}) = exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + u_i)f(t_{ij})$, proposed in the previous section which we refer to below as the *frailty latent variable* model.

It is assumed that the progression time follows a lognormal distribution, and estimation of the parameters is derived from the EM algorithm approach described in the previous section. To estimate the fixed and random components as well as the variance components, a PQL method (Breslow and Clayton, 1993) based on an algorithm proposed by Schall (1991) is used. This estimation is carried out on each iteration until the estimates converge. The algorithm applied is the $S-Plus$ 2000 function $glmmPQL$ of Venables and Ripley(2002) which is used for this particular data set. The bootstrap method (Efron and Thibshirani, 1993) with R = 100 (one hundred bootstrap replications) is used for estimating the standard errors of each of the parameters in the model.

The graph of the survival function is given in Figure 1. The figure shows lines estimated by Kaplan-Meier method(solid line) and the latent variable method using the lognormal distribution (dashed line). They are very close.

The first four rows of Table 1 give the parameter estimates for the model when the covariate T-stage is excluded. The bootstrap standard deviation for the estimates are also included. The values

Table 1: Parameter Estimates of the Models for Carcinoma Data

	Frailty Latent Variable Model	Latent Variable Model
intercept	0.561 (0.110)	0.561(0.101)
σ^*	1.369e-07(0.178)	
μ	6.134(0.086)	6.134(0.088)
σ	0.796(0.069)	0.796(0.076)
intercept	0.417(0.156)	0.417(0.155)
T-stage	0.346(0.276)	0.346(0.281)
σ^*	2.809e-07(0.199)	
μ	6.134(0.085)	6.134(0.071)
σ	0.796(0.074)	0.796(0.077)

in brackets are standard deviations calculated from the bootstrap replications. For comparative purposes, the estimates obtained by a similar model without random effects and their bootstrap standard deviation are also included.

It can be seen from the second row of the table that the estimate for variance component of the random effect σ^* is effectively 0, an indication that there is not any significant institution effect present in the data. This observation is supported by a relatively large value of the bootstrap standard error of the variance component estimate. The estimate for the intercepts and their bootstrap standard errors for both models are similar. The same is true for the estimate of the mean μ and the standard deviation σ of the lognormal distribution. All these estimates are significantly different from 0. The estimates of the cured proportion in both of the models are similar, about 17.3 percent, compared to 15.5 percent for the cure proportion estimated by the mixture model (fitted using $GFCURE\ S-Plus$ 2000 function) with a Weibull distribution.

When the covariate T-stage is included in the model the estimates are reported in the last five rows of Table 1. The results are similar to the model without covariates. The cure proportion for patients with small primary tumor is 21.9 percent, much higher than that of the patients with severe primary tumor (11.7 percent). However, based on a Wald-test, this difference is not statistically significant, i.e there is not enough evidence to show that a severe tumor condition in T-stage reduces the proportion of cure in the data.

3.2 Bladder Cancer Data Example

In this second example a bladder cancer data set listed in Wei et al.(1989)is used to illustrate the use of the frailty latent variable method of analysis. The experiment was conducted by Veterans Administration Cooperative Urological Research Group. All patients entering the study had superficial bladder tumors. The tumors were removed and then the patients were randomly assigned to one of three treatments, namely placebo, thiotepa and pyridoxine. Multiple reoccurrences of tumors were observed on many of the patients during the study. Upon observing the reoccurrences, the tumors were removed at each visit. The interest is to evaluate the effectiveness of treatment (thiotepa) against placebo. The data set is available in $S-Plus$ 2000 listed as *bladder* data. There are several variables recorded in the data set, namely: *id* referring to the patient ID, *rx* referring to the treatment (placebo =1 and thiotepa = 2), *number* referring to the initial number of tumors, *size* referring to the size (cm) of largest initial tumor, *start* referring to entry time into

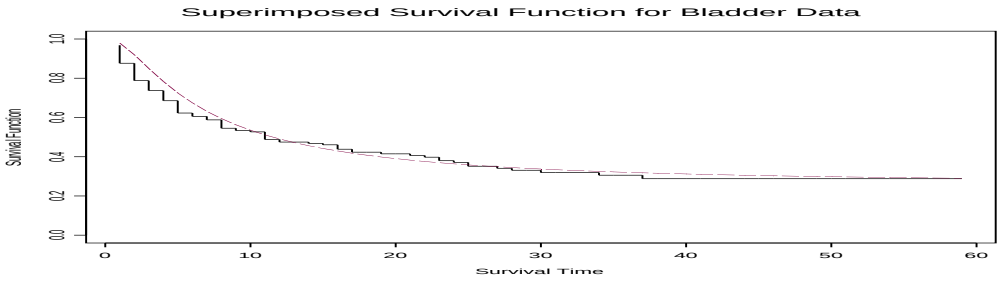

Figure 2: Survival Function Estimated by Kaplan-Meier and Frailty Latent Variable Model

the study or time of last occurrence, *stop* referring to reoccurrence or censoring time and *enum* referring to which reoccurrences out of four occurrences (1 to 4).

This data set is suitable for the frailty latent variable method, since there are some patients for whom more than one survival time or censor time is recorded. In this sub-section we assume that the progression times follow a lognormal distribution. In addition, as all the patients entered this study had their tumor removed earlier, there was a good chance of the existence of longterm survivor in this study. To assess the significance of the estimates, bootstrap standard errors based on 100 bootstrap replications are generated.

There were 85 patients with superficial bladder tumors. Of these patients, 47 were randomized into the placebo group, and 38 were randomized into the thiotepa group. For the purpose of applying the frailty latent variable method, the data is slightly modified by defining the time to occur or reoccur as the difference between *stop* and *start*. If this time is 0 it is removed since then there is not really any occurrence or reoccurrence. Following this setting, for the 85 patients there were 190 records for the time to occurrence or reoccurrence. Out of the 190 records, 78 of the times were censored. Since there are multiple records for most of the patients, the patient specific random effects are included.

The Kaplan-Meier plot of the survival time of the patients can be found in Figure 2. Along with the Kaplan-Meier curve, the survival time estimates from the latent variable model and frailty latent variable model are superimposed. Note that no covariates are included in the models. Both figures are close.

It can be seen from Figure 2 that models with and without frailty agree. This indicates that there is not much change in the estimates with the inclusion of the random effect. In fact, the variance component estimate σ^* of the random effect is relatively small (0.026) and the bootstrap standard error is equal to 0.145. The cure proportion given by the latent variable model is 0.265 while the estimate for the model with frailty is 0.268. The first four rows of Table 2 list all the estimates in the models and their standard deviations. Note that as in the previous example, the number of bootstrap replications used to estimate standard deviations is 100. From the table, the intercept is significantly different from 0 (estimated value 0.276 with bootstrap standard error 0.073). Other estimates of the parameters of lognormal distribution, namely the mean and the

Table 2: Parameter Estimates of the Models for Bladder Data

	Frailty Latent Variable Model	Latent Variable Model
intercept	0.276(0.073)	0.285(0.075)
σ^*	0.026(0.145)	
μ	2.373(0.062)	2.373(0.071)
σ	1.101(0.053)	1.101(0.057)
intercept	0.104(0.283)	0.123(0.205)
treat	-0.328(0.267)	-0.326(0.211)
number	0.135(0.062)	0.131(0.049)
size	-0.012(0.086)	-0.015(0.072)
σ^*	0.027(0.194)	
μ	2.373(0.069)	2.373(0.071)
σ	1.101(0.051)	1.101(0.050)

standard deviation (μ and σ) are also significant.

A further investigation to determine whether the cure proportion in the data is related to the treatment and /or other factors such as number and size of tumors is needed. For this purpose, a model which includes covariates *treat*, *number* and *size* is fitted. The parameter estimates and their bootstrap standard deviations are listed in final seven rows of Table 2. It can be seen from the second half of Table 2 that the data provide some weak evidence that thiotepa treatment increases the cured proportion. Applying the treatment to the patients increases the proportion of cured patients in the study up to 11.9 %. However, a Wald-test for the treatment effect gives a p-value of 0.0612 in the model without frailty and 0.1094 in the model with frailty. This finding is similar to the one obtained by Wei et al. (1989) when analyzing the same data set. On the other hand, there is strong evidence that the number of tumors reduces the proportion of cured patients in the data. The estimate for the coefficient of the number of tumors is 0.1345 with bootstrap standard error equal to 0.06202 in the model with frailty and 0.04949 in the model without frailty. A one unit increase in the number of tumors will reduce the cure proportion by up to 4.93 %. The score test to test the hypothesis for the effect of the number of tumors gives a p-value of 0.015 for the model with frailty and 0.003 for the model without frailty. The estimate of size of the tumor is a small negative number with a large standard error. This indicates that the tumor's size does not have any significant effect on the proportion of the cured patients in the study.

4 Simulation Study

A simulation study to evaluate the performance of the frailty model and its proposed estimation method is conducted by adopting a multi-centre clinical trial setting, where centres are treated as random effect. Various combinations of parameter values that relate to the various values of censoring proportions in the susceptible population and various amounts of variability in the random effects are applied. The focus of the investigation is to determine the accuracy of the method in predicting the cure proportion. Hence, the fixed effect parameters in the Poisson GLMM only consists of an intercept. Consequently, there are no covariates included in the model. In addition, two types of censoring patterns namely the random censoring and fixed censoring are

considered. Finally, a comparison will be made as to whether the developed model/method works well with these two types of censoring patterns.

4.1 Simulation Methods

The failure time function for the susceptibles sub-population derived from the survival function in (1.1) can be expressed as:

$$f_s(t) = \frac{exp\{-exp(\eta)F(t|\lambda)\}}{1 - exp\{-exp(\eta)\}}\ exp(\eta)f(t|\lambda), \tag{4.1}$$

where $f(t|\lambda)$ is a univariate distribution specified for the progression time of the carcinogenic cells, $\eta = \mathbf{x}\boldsymbol{\beta} + u$ is the linear predictor that connects the immune proportion in the population with some parametric vector $\boldsymbol{\beta}$ in the model and u is a random effect which follows a Gaussian distribution with mean 0 and standard deviation σ^*.

In all the simulation sets, the susceptibles sub-population failure times are generated from (4.1), and the failure times of the immune sub-population is taken to be infinity. A lognormal distribution is specified for $f(t|\lambda)$; where $\lambda = (\mu, \sigma)$ are the parameters in the distribution. A set of values (-1, -1/2, 0, 1/2, 1) is assigned for β. Due to the nature of the model from which the simulated data is generated, these parameter values are not independent of the censoring proportions of the susceptibles in the generated data. Jointly with the parameter values used in the censoring distribution, the β values determine the censoring fraction in the simulated data.

When a lognormal distribution with $\mu = \log 1000$ and $\sigma = 1$ is specified for the censor distribution and β = (-1, -1/2, 0, 1/2, 1) is used, the corresponding censoring fractions for the susceptibles are 28.5%, 26.8%, 24.3%, 20.4% and 14.9%. For $\mu = \log 2000$, the corresponding censoring fractions are 14.5%, 13.5%, 11.8%, 9.35% and 6.3%. Note that since the lifetimes of the immune sub-population are arbitrarily assigned as censors, the censoring fraction in the generated data is higher than these numbers. The total proportion of censored values can be easily determined by taking into account the contribution of the immune proportion.

In addition to examining the model in the presence of random censoring, the model is also examined in the presence of fixed censoring values, i.e., type I censoring. There are two fixed censoring values L chosen, namely L= 700 and L= 1200. As for random censoring, the censoring fraction in the susceptibles sub-population cannot be easily determined. Hence, the same type of simulation as for the random censoring case is used. When β = (-1, -1/2, 0, 1/2, 1) and L=700, the censoring fractions are 32.6%, 29.8%, 26.1%, 19.8% and 12.2%. For L=1200, the censoring fractions in the model are 16.3%, 14.6%, 12.3%, 8.8% and 4.8%.

Apart from the parameter β and the parameters μ and σ in the censoring distribution, parameters for the random effects u (the centre effect and its distribution function) also need to be determined. A normal distribution with mean 0 and two parameter values for the standard deviation σ^* namely $\sigma^* = 1/3$ and σ^*=1 are used. These two values determine the amount of variability introduced in the simulated data. The larger the value of σ^* the greater the variability introduced in the data. σ^*=1/3 gives a light amount of variability in the data, while σ^*=1 gives a high amount of variability in the data.

There are 10 centres in the simulated data and each centre has 15 patients. Therefore, the number of random samples n to be generated is 150, which are randomly placed in the immune or susceptibles group. The placement is done by generating binary random numbers B from the Bernoulli distribution with probability $1-p = exp(-exp(\eta))$, where $\eta = \beta + u$. If B=1 the sample belongs to the immune sub-population otherwise it belongs to the susceptibles sub-population. To generate lifetime distribution of susceptibles $f_s(t)$ given by (4.1) a rejection/acceptance method is adopted (Tanner, 1996). This method works well since the functional form of $f_s(t)$ is known. Once a random sample of lifetimes for the susceptibles sub-population has been generated, the

observed lifetimes can be determined by taking the minimum of the lifetime or the censor time. In addition, the censor indicator and the centre number are recorded.

4.2 Simulation Results and Discussion

There are five different values of β, two different values of σ^*, two different values of either the logmean, μ, or L for the censor distribution and two different types of censors, namely the fixed and random censors. Therefore, forty different simulations are made. The results are shown in Tables 3, 4, 5 and 6. Each table consists of ten simulation sets corresponding to five different values of β and two different values of either the logmean, μ (for random censoring), or L (for fixed censoring). The first column of the tables contains the parameter names followed by their true values in the second column. The third and the fourth columns are the means of the parameter estimates obtained from the 100 simulations for two different values (estimates-1 and estimates-2) of the either μ's or the L's. The values in brackets are the standard deviations calculated from the 100 estimates in the simulations. In addition the tables also report the censoring fraction values for the susceptibles.

From Tables 3 and 4, the results for the simulations with random censoring, there are small biases in the estimates for β. The bias is slightly more noticeable when more variability is introduced in the random effects ($\sigma^* = 1$ as opposed to $\sigma^* = 1/3$), which implies that more variability is introduced in the data. In the data with the higher censoring fractions in the susceptibles sub-population, where the censoring distribution has $\mu = \log 1000$, the parameter estimate for β = -1 is slightly improved when σ^*= 1 is changed to σ^*= 1/3. For the smaller censoring fraction in the susceptibles sub-population (the random censor distribution $\mu = \log 2000$), the two estimates for β have relatively small biases and are relatively close to each other. As β increases, which corresponds to a decrease in the censoring fraction, the biases decrease in all cases where small variability is introduced in the data. For the simulated data with higher variability, the biases increase slightly as β increases. However, these biases are relatively small particularly when the transformed version of β, the $1 - p$ are considered.

The estimates for the parameters in the lognormal distribution, namely μ and σ have relatively small biases in all cases regardless of the variability and the values of the censor distribution means. All the estimates also have small standard deviations indicating less variability of the estimated values in the simulation.

The estimates for the variance component σ^* of the random effects show small biases in most cases. These estimates behave differently for $\beta \leq 0$ compared to $\beta > 0$. In the former case, when $\mu = \log 2000$, the bias of the estimate of σ^* is relatively smaller than when μ= $\log 1000$. For $\beta > 0$, the biases are larger when $\mu = \log 2000$. This pattern is consistent for both assumed values of σ^*.

Tables 5 and 6 show simulation results for the model when fixed censoring is used. As in the case of random censoring, the estimates for parameters β, μ, σ and $1 - p$ show small biases in all cases regardless of the amount of variability introduced in the data. These biases are slightly smaller than those seen in the random censoring simulations. The estimates when the censoring fraction is smaller, i.e. $L = 700$, are as good as the ones when censoring fraction is greater, i.e. $L = 1200$. The standard deviations in all of these estimates are also similar to the corresponding estimates for the random censoring case in Tables 3 and 4. In addition, despite their similarities, the estimates for the variance component σ^* of the random effects show slightly smaller biases in the random censoring case than the ones for the fixed censoring, for most of the cases.

Overall, the results obtained from the simulations show that the model and the estimation procedures are good, i.e. the estimates show very small biases. The results also show that the estimates in the model are robust to various amount of censoring proportions in the susceptibles sub-population. This feature is important and desirable since in real data, censoring observations resulting from the immune sub-population are indistinguishable from those resulting from

Table 3: Parameter Estimates, Censor Distribution $\mu = \log 1000$ (estimates-1) and $\mu = \log 2000$ (estimates-2), SD of Random Effects $\sigma^*= 1$

Set-1

censor-prop		28.5 %	14.5 %
Parameter	True Value	estimates-1 (SD)	estimates-2 (SD)
β	-1	-1.133*(0.353)*	-1.038*(0.296)*
$1-p$	0.692	0.725	0.702
σ	1	0.977*(0.093)*	0.977*(0.085)*
μ	6.215	6.093*(0.166)*	6.082*(0.155)*
σ^*	1	0.848*(0.347)*	0.915*(0.345)*

Set-2

censor-prop		26.8 %	13.5 %
Parameter	True Value	estimates-1 (SD)	estimates-2 (SD)
β	-1/2	-0.538*(0.302)*	-0.541*(0.279)*
$1-p$	0.545	0.558	0.559
σ	1	0.989*(0.101)*	0.993*(0.072)*
μ	6.215	6.082*(0.167)*	6.087*(0.119)*
σ^*	1	0.851*(0.317)*	0.878*(0.311)*

Set-3

censor-prop		24.3 %	11.8 %
Parameter	True Value	estimates-1 (SD)	estimates-2 (SD)
β	0	-0.113*(0.299)*	-0.122*(0.263)*
$1-p$	0.368	0.409	0.413
σ	1	0.992*(0.081)*	0.998*(0.076)*
μ	6.215	6.108*(0.163)*	6.098*(0.127)*
σ^*	1	0.905*(0.262)*	0.890*(0.262)*

Set-4

censor-prop		20.4 %	9.35 %
Parameter	True Value	estimates-1 (SD)	estimates-2 (SD)
β	1/2	0.442*(0.231)*	0.399*(0.270)*
$1-p$	0.192	0.211	0.225
σ	1	1.019*(0.081)*	1.018*(0.064)*
μ	6.215	6.125*(0.164)*	6.121*(0.159)*
σ^*	1	0.822*(0.211)*	0.839*(0.199)*

Set-5

censor-prop		14.9 %	6.3 %
Parameter	True Value	estimates-1 (SD)	estimates-2 (SD)
β	1	0.885*(0.184)*	0.901*(0.195)*
$1-p$	0.066	0.089	0.085
σ	1	1.049*(0.081)*	1.041*(0.068)*
μ	6.215	6.226*(0.144)*	6.181*(0.131)*
σ^*	1	0.863*(0.217)*	0.808*(0.203)*

Table 4: Parameter Estimates, Censor Distribution $\mu = \log 1000$ (estimates-1) and $\mu = \log 2000$ (estimates-2), SD of Random Effects $\sigma^*= 1/3$

Set-1

censor-prop		28.5 %	14.5 %
Parameter	True Value	estimates-1 (SD)	estimates-2 (SD)
β	-1	-1.015*(0.188)*	-0.985*(0.209)*
$1-p$	0.692	0.696	0.688
σ	1	1.012*(0.119)*	0.985*(0.093)*
μ	6.215	6.224*(0.118)*	6.180*(0.122)*
σ^*	1/3	0.271*(0.277)*	0.291*(0.242)*

Set-2

censor-prop		26.8 %	13.5 %
Parameter	True Value	estimates-1 (SD)	estimates-2 (SD)
β	-1/2	-0.486*(0.151)*	-0.497*(0.147)*
$1-p$	0.545	0.541	0.544
σ	1	0.983*(0.077)*	0.985*(0.081)*
μ	6.215	6.188*(0.107)*	6.201*(0.105)*
σ^*	1/3	0.219*(0.209)*	0.244*(0.186)*

Set-3

censor-prop		24.3 %	11.8 %
Parameter	True Value	estimates-1 (SD)	estimates-2 (SD)
β	0	-0.028*(0.121)*	0.001*(0.137)*
$1-p$	0.368	0.409	0.413
σ	1	1.005*(0.079)*	0.997*(0.070)*
μ	6.215	6.216*(0.089)*	6.185*(0.080)*
σ^*	1/3	0.273*(0.188)*	0.301*(0.180)*

Set-4

censor-prop		20.4 %	9.35 %
Parameter	True Value	estimates-1 (SD)	estimates-2 (SD)
β	1/2	0.497*(0.110)*	0.489*(0.115)*
$1-p$	0.192	0.193	0.196
σ	1	1.013*(0.057)*	1.005*(0.059)*
μ	6.215	6.203*(0.089)*	6.201*(0.077)*
σ^*	1/3	0.299*(0.158)*	0.269*(0.180)*

Set-5

censor-prop		14.9 %	6.3 %
Parameter	True Value	estimates-1 (SD)	estimates-2 (SD)
β	1	1.005*(0.089)*	1.010*(0.091)*
$1-p$	0.066	0.065	0.064
σ	1	0.996*(0.059)*	0.996*(0.060)*
μ	6.215	6.214*(0.080)*	6.201*(0.083)*
σ^*	1/3	0.329*(0.217)*	0.302*(0.136)*

Table 5: Parameter Estimates, Censor Value $L = 700$ (estimates-1) and $L = 1200$ (estimates-2), SD of Random Effects $\sigma^*= 1$

Set-1

censor-prop		32.6 %	16.3 %
Parameter	True Value	estimates-1 (SD)	estimates-2 (SD)
β	-1	-1.102*(0.322)*	-0.984*(0.330)*
$1-p$	0.692	0.710	0.680
σ	1	0.973*(0.145)*	0.955*(0.110)*
μ	6.215	6.072*(0.179)*	6.126*(0.162)*
σ^*	1	0.956*(0.373)*	0.826*(0.325)*

Set-2

censor-prop		29.8 %	14.6 %
Parameter	True Value	estimates-1 (SD)	estimates-2 (SD)
β	-1/2	-0.592*(0.291)*	-0.560*(0.323)*
$1-p$	0.545	0.570	0.559
σ	1	0.988*(0.127)*	1.008*(0.101)*
μ	6.215	6.114*(0.190)*	6.090*(0.135)*
σ^*	1	0.876 *(0.294)*	0.897*(0.273)*

Set-3

censor-prop		26.1 %	12.3 %
Parameter	True Value	estimates-1 (SD)	estimates-2 (SD)
β	0	-0.028*(0.201)*	-0.043*(0.259)*
$1-p$	0.368	0.378	0.384
σ	1	1.002*(0.089)*	1.010*(0.084)*
μ	6.215	1.002*(0.089)*	6.121*(0.136)*
σ^*	1	0.826*(0.229)*	0.864*(0.259)*

Set-4

censor-prop		19.8 %	8.8 %
Parameter	True Value	estimates-1 (SD)	estimates-2 (SD)
β	1/2	0.426*(0.188)*	0.422*(0.202)*
$1-p$	0.192	0.220	0.221
σ	1	1.017*(0.085)*	1.019*(0.074)*
μ	6.215	6.197*(0.135)*	6.171*(0.122)*
σ^*	1	0.813*(0.264)*	0.827*(0.246)*

Set-5

censor-prop		12.2 %	4.8 %
Parameter	True Value	estimates-1 (SD)	estimates-2 (SD)
β	1	0.934*(0.159)*	0.873*(0.173)*
$1-p$	0.066	0.083	0.097
σ	1	1.071*(0.083)*	1.081*(0.079)*
μ	6.215	6.267*(0.160)*	6.266*(0.130)*
σ^*	1	0.812*(0.238)*	0.844*(0.230)*

Table 6: Parameter Estimates, Censor Value $L = 700$ (estimates-1) and $L = 1200$ (estimates-2), SD of Random Effects $\sigma^*= 1/3$

Set-1

censor-prop		32.6 %	16.3 %
Parameter	True Value	estimates-1 (SD)	estimates-2 (SD)
β	-1	-1.049*(0.184)*	-0.984*(0.157)*
$1-p$	0.692	0.702	0.691
σ	1	0.989*(0.132)*	0.996*(0.103)*
μ	6.215	6.208*(0.129)*	6.199*(0.108)*
σ^*	1/3	0.301*(0.325)*	0.256*(0.229)*

Set-2

censor-prop		29.8 %	14.6 %
Parameter	True Value	estimates-1 (SD)	estimates-2 (SD)
β	-1/2	-0.506*(0.116)*	-0.484*(0.135)*
y $1-p$	0.545	0.547	0.544
σ	1	1.008*(0.089)*	1.007*(0.079)*
μ	6.215	6.1968*(0.096)*	6.190*(0.092)*
σ^*	1/3	0.263*(0.218)*	0.246*(0.215)*

Set-3

censor-prop		26.1 %	12.3 %
Parameter	True Value	estimates-1 (SD)	estimates-2 (SD)
β	0	0.000*(0.108)*	0.025*(0.146)*
$1-p$	0.368	0.368	0.359
σ	1	1.004*(0.079)*	0.989*(0.075)*
μ	6.215	6.203*(0.096)*	6.192*(0.090)*
σ^*	1/3	0.290*(0.187)*	0.276*(0.176)*

Set-4

censor-prop		19.8 %	8.8 %
Parameter	True Value	estimates-1 (SD)	estimates-2 (SD)
β	1/2	0.494*(0.093)*	0.500*(0.104)*
$1-p$	0.192	0.196	0.194
σ	1	0.993*(0.074)*	0.988*(0.060)*
μ	6.215	6.222*(0.091)*	6.212*(0.073)*
σ^*	1/3	0.290*(0.160)*	0.289*(0.155)*

Set-5

censor-prop		12.2 %	4.8 %
Parameter	True Value	estimates-1 (SD)	estimates-2 (SD)
β	1	1.009*(0.071)*	1.011*(0.080)*
$1-p$	0.066	0.065	0.065
σ	1	1.005*(0.066)*	0.995*(0.054)*
μ	6.215	6.217*(0.081)*	6.204*(0.075)*
σ^*	1/3	0.286*(0.144)*	0.275*(0.154)*

the susceptibles sub-population. This is particularly true if the censoring observations from the susceptibles sub-population are large numbers.

5 Further Discussion

In this article, an extension of the model proposed originally by Yakovlev et al.(1993) to include frailty has been studied. Use of this model to real data sets has been successful. A GLMM using a PQL approach is proposed for parameter estimation. This proposal is naturally appealing because of its simplicity and ready applicability utilizing existing statistical software. Moreover, the new mode l retains the proportional hazard property of Yakovlev's model. This property is desirable in most cases in survival analysis.

Application of the method to two real data sets shows that the survival functions of the model without covariates are close to the ones generated by the Kaplan-Meier method. When covariates are included, the estimates in the model are close to the estimates in the model without frailty. This similarity comes as no surprise, since the variance component of the frailty random effects are close to 0 for both data sets. This implies that there is not really any frailty effects in both data sets.

A simulation study conducted to evaluate estimates in the model confirmed that the estimates have relatively small bias. The method works equally well in both the random and fixed censoring cases. One problem to be faced is the difficulty in computing standard deviations for the estimates in the model. This can be handled by using the bootstrap method.

The method for survival data with an immune proportion and frailty developed in this article is readily extended to the situation where more than one level of random effects exists in the data. An alternative method for estimation and prediction for the frailty model can be achieved by integrating out the random effect. This method is well developed for small dimension of random effects but is restrictive for large dimension of random effects. The advantage of this method lies in the fact that the inferences can be built on a likelihood base. Therefore, it is an interesting alternative for future considerations.

References

[1] Aalen, O. (1992). Modelling heterogeneity in survival analysis by the compound poisson distribution. *Annals of Applied Probability* **2**, 951–972.

[2] Aalen, O. (1998). Frailty models. *In* Statistical Analysis of Medical Data *(Eds B.S. Everitt and G. Dunn)* pages 59–74.

[3] Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

[4] Chen, M. and Ibrahim, J. (2001). Maximum likelihood methods for cure rate models with missing covariates. *Biometrics* **57**, 43–52.

[5] Chen, M., Ibrahim, J. and Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction. *Journal of American Statistical Association* **94**, 909–919.

[6] Efron, B. and Tibshirani, R. (1993). *The Jacknife, the Bootstrap and Other Resampling Plans.* Chapman and Hall, New York.

[7] Hasan, B., Singh, R. and Pesotan, H. (2005). On the use of the lognormal model for survival data with surviving fraction. Technical Report 2005-307, Department of Mathematics and Statistics, University of Guelph, Guelph, ON, Canada.

[8] Ibrahim, J., Chen, M. and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer, New York.

[9] Kalbfleisch, J. and Prentice, R. (2003). *The Statistical Analysis of Failure Time Data (2nd Edition)*. Wiley, New York.

[10] Lancaster, T. (1979). Econometric methods for the duration of unemployment. *Econometrica* **47**, 939–956.

[11] Lancaster, T. (1990). *The Econometric Analysis of Transition Data*. Cambridge University Press, Cambridge.

[12] Maller, R. and Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*. Wiley, New York.

[13] McGilchrist, C. (1994). Estimation in generalized linear mixed model. *Journal of the Royal Statistical Society B* **56**, 61–69.

[14] Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, New York.

[15] Prentice, R., Williams, B. and Peterson, A. (1981). On the regression analysis of multivariate failure time data. *Biometrika* **68**, 373–379.

[16] Schall, R. (1991). Estimation in generalized linear model with random effects. *Biometrika* **78**, 719–727.

[17] Tanner, M. (1996). *Tools for Statistical Inference (3rd Edition)*. Springer, Berlin.

[18] Venables, W. and Ripley, B. (2002). *Modern Applied Statistics with S (4th Edition)*. Springer, Berlin.

[19] Wei, L., Lin, D. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distribution. *Journal of the American Statistical Association* **84**, 1065–1073.

[20] Yakovlev, A., Asselain, B., Bardou, V., Forquet, A., Hoang, T., Rochefodiere, A. and Tsodikov, A. (1993). A simple stochastic model of tumor reoccurence and its application to data on premenopausal breast cancer. *In* Biometrie et Analyse de Donnees Spatio-Temporelles*, (Eds B.Asselain, M.Boniface, C.Duby, C.Lopez, J.P. Masson, and J. Tranchefort)* **12**, 66–82.

[21] Yakovlev, A. and Tsodikov, A. (1996). *Stochastic Models of Tumor Latency and their Biostatistical Applications*. World Scientific, Singapore.

[22] Yau, K. and Ng, A. (2001). Longterm survivor mixture model with random effects: application to a multi-centre clinical trial of carcinoma. *Statistic in Medicine* **20**, 1591–1607.