Comparison of different estimation procedures for proportional hazard models with random effects

José Abrahantes, Catherine Legrand, Tomasz Burzykowski, Paul Janssen, Vincent Ducrocq, Luc Duchateau

▶ **To cite this version:**

José Abrahantes, Catherine Legrand, Tomasz Burzykowski, Paul Janssen, Vincent Ducrocq, et al.. Comparison of different estimation procedures for proportional hazard models with random effects. Computational Statistics and Data Analysis, 2007, 51, pp.3913-3930. hal-02658252

HAL Id: hal-02658252
https://hal.inrae.fr/hal-02658252

Submitted on 30 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 3913–3930
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Comparison of different estimation procedures for proportional hazards model with random effects

José Cortiñas Abrahantes[a,*], Catherine Legrand[b], Tomasz Burzykowski[a], Paul Janssen[a], Vincent Ducrocq[c], Luc Duchateau[d]

[a] *Center for Statistics, Hasselt University, Agoralaan D, B3590 Diepenbeek, Belgium*
[b] *European Organization for Research and Treatment of Cancer (EORTC), B1200 Brussels, Belgium*
[c] *Station de Génétique Quantitative et Appliquée, Institut National de la Recherche Agronomique, 78352 Jouy en Josas, France*
[d] *Department of Physiology, Biochemistry and Biometrics, Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, B9820 Merelbeke, Belgium*

Received 11 January 2005; received in revised form 15 March 2006; accepted 29 March 2006
Available online 24 April 2006

Abstract

Proportional hazards models with multivariate random effects (frailties) acting multiplicatively on the baseline hazard are a topic of intensive research. Several estimation procedures have been proposed to deal with this type of models. Four procedures used to fit these models are compared in two real-life datasets and in a simulation study. The performance of the four methods is investigated in terms of the bias of point estimates, their empirical variability and the bias of the estimation of the variability.
© 2006 Elsevier B.V. All rights reserved.

Keywords: Frailty model; Restricted or residual maximum likelihood; Penalized partial likelihood; Laplace approximation; Multivariate failure-time data

1. Introduction

In applied sciences, one is often confronted with *correlated data*. This generic term embraces a multitude of data structures, such as multivariate observations, clustered data, repeated measurements, longitudinal data, and spatially correlated data. Instances of this type of research can be encountered in virtually every empirical branch of science.

We will focus on clustered failure-time data. A way to model this type of data is to use a proportional hazards model conditional on random effects introduced to allow for correlation between the observations from the same cluster. First proposals for such a modelling strategy concentrated on the univariate mixed effects model (also called shared frailty model), which only includes a univariate random effect in the model.

However, shared frailty models have some limitations. For instance, they force the unobserved factors (frailties) to be the same for all failure times within the cluster (Xue and Brookmeyer, 1996). This may not always be desirable. Another drawback is that in most cases, a univariate frailty can only induce positive association within the cluster

* Corresponding author. Tel.: +32 11 268215; fax: +32 11 268299.
E-mail address: jose.cortinas@uhasselt.be (J.C. Abrahantes).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.03.009

(Xue and Brookmeyer, 1996). Clearly, there are some situations in which the failure times within the same cluster may be negatively associated.

To avoid these limitations, models with multivariate, correlated random effects have been proposed. The main problem with the use of such models is parameters estimation. To this aim, various estimation approaches have been proposed.

For instance, McGilchrist and Aisbett (1991) and McGilchrist (1993, 1994) used a penalized likelihood approach. Sastry (1997) proposed the EM algorithm for a nested frailty model assuming gamma-distributed frailties. Other variants of the EM algorithm were developed by Xue and Brookmeyer (1996), Vaida and Xu (2000), Ripatti et al. (2002), and Cortiñas and Burzykowski (2005). The various implementations of the algorithm differ in the methods used to compute conditional expectations of random effects at the E-step. In particular, Xue and Brookmeyer (1996) considered the use of numerical integration, Vaida and Xu (2000) applied Monte Carlo Markov Chain (MCMC) sampling, Ripatti et al. (2002) used rejection sampling, while Cortiñas and Burzykowski (2005) applied the Laplace approximation. Maples et al. (2002) combined the EM algorithm with the Newton–Raphson method to estimate a two-level model by maximizing an empirical version of the partial likelihood. Xue (1998, 2001) developed an alternative fitting method based on estimating equations derived from a Poisson regression formulation. The formulation was also used by Ha and Lee (2003) to propose an approach based on a hierarchical likelihood, and by Ma et al. (2003) to develop a method based on fitting random effects Poisson models. Bayesian approaches were considered by, e.g., Ducrocq and Casella (1996), Gustafson (1997), Sinha and Dey (1997), Sargent (1998), and Xue and Ding (1999). Ripatti and Palmgren (2000) proposed estimation based on a penalized partial likelihood obtained by applying the Laplace approximation to the marginal likelihood function.

There are different advantages and drawbacks related to the different approaches. For instance, the EM algorithm implementation developed by Sastry (1997) allows to analyze survival data clustered at two levels. However, it assumes a particular form of the model (basically, shared frailty), with a piecewise baseline hazard and gamma-distributed frailites. The EM algorithm implementation developed by Cortiñas and Burzykowski (2005) allows for an arbitrary baseline hazard and can be applied to various distributions of random effects. It is also less computationally intensive than the implementations considered by Xue and Brookmeyer (1996), Vaida and Xu (2000), and Ripatti et al. (2002). However, due to the asymptotic nature of the Laplace approximation, the method requires that the cluster size is large enough. Maples et al. (2002) suggest that in their combined approach a quicker convergence rate, related to the use of the Newton–Raphson method, can be obtained. They report, however, bias in the estimates of standard deviations and covariances of the random effects. Ha et al. (2001) and Ha and Lee (2003) indicate that in some extreme cases the hierarchical likelihood approach might be biased.

Information on the relative merits of the various approaches is scarce. The objective of this paper is to, at least partially, fill this gap. To this aim, a simulation study is conducted. We consider the methods developed by McGilchrist and Aisbett (1991), Ducrocq and Casella (1996), Ripatti and Palmgren (2000) and Cortiñas and Burzykowski (2005). The main reasons for this choice were the differences in the nature of the methods, software availability, and feasibility of conducting a simulation study.

The paper is organized as follows. Section 2 briefly recalls the proportional hazards model with random effects. In Section 3 the four estimation methods are reviewed. In Section 4 the estimation methods are applied to two case studies. Section 5 describes the simulation study. The results are presented in Section 6. A short discussion of the results in Section 7 concludes the paper.

2. The proportional hazards model with random effects

We consider clustered failure-time data with N clusters. The failure-time random variable corresponding to subject j ($j = 1, \ldots, n_i$) from cluster i ($i = 1, \ldots, N$) will be denoted by Y_{ij}. It is assumed that observations of Y_{ij} can be right-censored. Thus, for subject j in cluster i we observe $T_{ij} = \min\left(C_{ij}, Y_{ij}\right)$, where C_{ij} is a censoring time (random variable), independent of Y_{ij}. Additionally, a censoring indicator δ_{ij} is observed, with δ_{ij} equal to 1 if $T_{ij} = Y_{ij}$, and 0 otherwise.

As we mentioned in the introduction the univariate shared frailty model was the first proposal to handle clustered failure-times data. It can be written as

$$\lambda\left(t_{ij} \mid \beta, \omega_i\right) = \lambda_0\left(t_{ij}\right) \omega_i \exp\left(x_{ij}^{\mathrm{T}} \beta\right), \tag{1}$$

where t_{ij} are the realizations of the random variable T_{ij}, $\lambda_0(.)$ is the baseline hazard function, β is a vector of fixed-effects corresponding to a vector of covariates x_{ij}, and cluster-specific random effects ω_i are assumed to be independent, identically distributed random variables with a common density function $f(\omega_i; \theta)$, where θ is the parameter quantifying the variability of frailties. One of the most common distribution assumed for the frailties is the gamma distribution (Clayton, 1978; Vaupel et al., 1979; Oakes, 1982; Hougaard, 2000). The main reason is that in this case it is easy to derive closed form expressions of marginal survival, density and hazard function. In the case of a parametric hazard, and if the random effects are gamma distributed, an analytic expression for the likelihood can be derived. On the other hand, if the hazard is unspecified, then the EM algorithm with closed form expression for the conditional expectation of the frailties can be used. It is worth noting that model (1) can be rewritten in the following form:

$$\lambda\left(t_{ij}|\beta, b_{i0}\right) = \lambda_0\left(t_{ij}\right)\exp\left(x_{ij}^{\mathrm{T}}\beta + b_{i0}\right), \tag{2}$$

where $b_{i0} = \ln\omega_i$. In what follows, we will distinguish between "frailties" ω_i and "random effects" b_{i0}.

We will consider the following extension of model (2):

$$\lambda\left(t_{ij}|\beta, b_i\right) = \lambda_0\left(t_{ij}\right)\exp\left(x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}}b_i\right), \tag{3}$$

where $\lambda_0(.)$ and β are the baseline hazard function and the vector of fixed effects, respectively, and b_i is a d-dimensional vector of random effects associated with a vector of covariates z_{ij}. We will assume that the random effects $b_i^{\mathrm{T}} = (b_{i0}, b_{i1}, \ldots, b_{id})$ are normally distributed with mean 0 and variance-covariance matrix $D = D(\theta)$. To simplify formulas, we will also use the baseline cumulative hazard defined as

$$\Lambda_0\left(t_{ij}\right) = \int_0^{t_{ij}} \lambda_0(u)\,\mathrm{d}u.$$

Model (3) can be seen as a linear mixed-effects model on the log-hazard scale. The estimation of the parameters β and θ from the observed data T_{ij} is our main interest. Denote by $g(.)$ and $S(.)$ the density and survival functions of Y_{ij}, respectively. Using the well-known relationships (see, e.g., Klein and Moeschberger, 1997) $g(t) = \lambda(t)S(t)$ and $S(t) = \exp\left\{-\int_0^t \lambda(u)\,\mathrm{d}u\right\}$, the likelihood for the observed data

$$\prod_{i=1}^{N}\prod_{j=1}^{n_i}\left\{g\left(t_{ij}\right)^{\delta_{ij}} S\left(t_{ij}\right)^{1-\delta_{ij}}\right\}$$

can be written as

$$\prod_{i=1}^{N}\prod_{j=1}^{n_i}\left\{\lambda\left(t_{ij}\right)^{\delta_{ij}} S(t_{ij})\right\} = \prod_{i=1}^{N}\prod_{j=1}^{n_i}\left\{\lambda\left(t_{ij}\right)^{\delta_{ij}}\mathrm{e}^{-\int_0^t \lambda(u)\,\mathrm{d}u}\right\}.$$

It follows that, assuming model (3) and the conditional independence of the observations within a cluster given b_i, one might write the (conditional) log-likelihood for the observed data as

$$l^{\mathrm{C}}(\beta, \lambda_0, b) = \sum_{i=1}^{N} l_i^{\mathrm{C}}(\beta, \lambda_0, b_i), \tag{4}$$

where

$$l_i^{\mathrm{C}}(\beta, \lambda_0, b_i) = \sum_{j=1}^{n_i}\left[\delta_{ij}\left\{\ln\lambda_0\left(t_{ij}\right) + x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}}b_i\right\} - \Lambda_0\left(t_{ij}\right)\exp\left(x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}}b_i\right)\right] \tag{5}$$

is the (conditional) log-likelihood for the observed data in the ith cluster, and b denotes the vector resulting from "stacking" vectors b_i for all clusters. The (marginal) likelihood of the observed data for all clusters can then be expressed as

$$L^{\mathrm{M}}(\beta, \theta, \lambda_0) = \int\prod_{i=1}^{N} L_i^{\mathrm{A}}(\beta, \theta, \lambda_0, b_i)\,\mathrm{d}b_i, \tag{6}$$

where

$$L_i^{\mathrm{A}}(\beta, \theta, \lambda_0, b_i) = f(b_i; \theta)\, \mathrm{e}^{l_i^{\mathrm{C}}(\beta, \lambda_0, b_i)} \tag{7}$$

and $f(b_i; \theta)$ is the density function of b_i. Note that (7) can be treated as the likelihood of the "augmented" data for cluster i, treating b_i as additional observations. Consequently,

$$L^{\mathrm{A}}(\beta, \theta, \lambda_0, b) = \prod_{i=1}^{N} L_i^{\mathrm{A}}(\beta, \theta, \lambda_0, b_i) \tag{8}$$

is the likelihood of the "augmented" data for all clusters.

One might consider using directly the likelihood function (6) in the inference on β and θ. There are, however, two major problems with using it for this purpose. First, it depends on the baseline hazard function $\lambda_0(.)$. Unless a parametric form of the hazard can be assumed, the usefulness of (6) is limited. Second, the integral in (6) will usually be multidimensional, unless a very simple model is considered, and, in general, will not be available in a closed form.

Several estimation approaches have been proposed to circumvent these problems. In the following section some of the methods are reviewed.

3. Estimation methods

In this section the approaches proposed by McGilchrist and Aisbett (1991), Ducrocq and Casella (1996), Ripatti and Palmgren (2000) and Cortiñas and Burzykowski (2005) are reviewed. In what follows, we will assume that b_i are normally distributed with mean 0 and variance-covariance matrix

$$D(\theta) = \begin{pmatrix} \theta_0 & 0 & \dots & 0 \\ 0 & \theta_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \theta_d \end{pmatrix}.$$

3.1. REML estimation method

McGilchrist and Aisbett (1991) used the penalized likelihood approach to estimate the fixed effects and the residual maximum likelihood (REML) to estimate the variance components of the random effects. Their method consists of finding the best linear unbiased predictors (BLUP) of the fixed and random components in a first stage to use them next to find REML estimates of the variance-covariance parameters.

3.1.1. BLUP and REML estimators

The estimation procedure is a generalization of the results developed by Schall (1991). In order to find the BLUP it is necessary to maximize the sum of two components. The first component is the partial log-likelihood of failure times treating the random effects as fixed:

$$l_1 = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \delta_{ij} \left\{ x_{ij}^{\mathrm{T}} \beta + z_{ij}^{\mathrm{T}} b_i - \ln \sum_{t_{kl} \geqslant t_{ij}} \exp\left(x_{kl}^{\mathrm{T}} \beta + z_{kl}^{\mathrm{T}} b_k \right) \right\}. \tag{9}$$

It can be seen that Eq. (9) is just the partial likelihood corresponding to Eq. (5). The second component is related to the distribution associated to the random effects

$$l_2 = -\frac{1}{2} \sum_{g=0}^{d} \left(N \ln 2\pi\theta_g + \sum_{i=1}^{N} \frac{b_{ig}^2}{\theta_g} \right). \tag{10}$$

The algorithm iterates between two steps. First, given an estimate for the variance-covariance matrix $D(\theta)$, one iteration is performed to update the estimates of the parameters β and b_i. Second, based on the updated values for β and b_i, the

REML estimator of $D(\theta)$ is used. Once $D(\theta)$ is estimated and updated, the process starts all over again. The details are as follows.

Given values $\beta^{(p)}$ and $b_i^{(p)}$ of the fixed and the random effects, obtained at the pth iteration, the Newton–Raphson iterative procedure is used for maximizing $l_1 + l_2$ to obtain BLUP estimators $\beta^{(p+1)}$ and $b_i^{(p+1)}$. Let $\eta_{ij} = x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}} b_i$ and $\eta = (\eta_1^{\mathrm{T}}, \eta_2^{\mathrm{T}}, \ldots, \eta_N^{\mathrm{T}})$, where $\eta_i = \left(\eta_{i1}, \eta_{i2}, \ldots, \eta_{in_i}\right)^{\mathrm{T}}$. In matrix form, $\eta = X\beta + Zb$, where X and Z are design matrices for the fixed and the random effects, respectively. The random effects are of the form $b = \left(b_0^{\mathrm{T}}, b_1^{\mathrm{T}}, \ldots, b_d^{\mathrm{T}}\right)^{\mathrm{T}}$, where $b_g = \left(b_{1g}, b_{2g}, \ldots, b_{Ng}\right)^{\mathrm{T}}$. The Newton–Raphson procedure is carried out as follows:

$$\begin{pmatrix} \beta^{(p+1)} \\ b^{(p+1)} \end{pmatrix} = \begin{pmatrix} \beta^{(p)} \\ b^{(p)} \end{pmatrix} - A^{-1} \begin{pmatrix} 0 \\ \left\{D(\theta)^{(p)}\right\}^{-1} b^{(p)} \end{pmatrix} + A^{-1} \begin{pmatrix} X^{\mathrm{T}} \\ Z^{\mathrm{T}} \end{pmatrix} \frac{\partial l_1}{\partial \eta}, \tag{11}$$

where

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} X^{\mathrm{T}} \\ Z^{\mathrm{T}} \end{pmatrix} \left[\frac{-\partial^2 l_1}{\partial \eta \partial \eta^{\mathrm{T}}}\right] (X \;\; Z) + \begin{pmatrix} 0_X & 0_Z \\ 0_Z & \left\{D(\theta)^{(p)}\right\}^{-1} \otimes I_N \end{pmatrix},$$

$(p+1)$ and (p) indicate the iterations of the algorithm, $\otimes$ denotes the Kronecker product and I_N is the identity matrix of dimension $N \times N$. The dimension of the zero matrices 0_X and 0_Z depends on the dimension of the vectors β and b, respectively. Matrix 0_X is a square matrix, while 0_Z has the same number of rows as 0_X, but its number of columns depends on the dimension of b. The inverse of A will be denoted by M and can be expressed as

$$\begin{aligned} M &= \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \\ &= \begin{pmatrix} A_{11}^{-1} + A_{11}^{-1} A_{12} S_{A_{11}}^{-1} A_{21} A_{11}^{-1} & -A_{11}^{-1} A_{12} S_{A_{11}}^{-1} \\ -S_{A_{11}}^{-1} A_{21} A_{11}^{-1} & S_{A_{11}}^{-1} \end{pmatrix}, \end{aligned}$$

where $S_{A_{11}} = \left(A_{22} - A_{21} A_{11}^{-1} A_{12}\right)$.

Given the estimates of β and b, the REML estimator of θ_g is

$$\theta_g^{(p+1)} = \frac{b_g^{(p+1)^{\mathrm{T}}} b_g^{(p+1)}}{N - \theta_g^{(p)^{-1}} \operatorname{tr}\left(M_{22(g)}\right)}, \tag{12}$$

where $\operatorname{tr}\left(M_{22(g)}\right)$ indicates the sum of elements of the diagonal of the submatrix of M_{22} related to the component g of the random effects. The algorithm alternates between (11) and (12) to estimate the values of the parameters.

3.1.2. Variance estimation

The elements needed in the estimation of the variance-covariance matrices for the estimated $\hat{\beta}$ and $\hat{\theta}$ are computed in the iterative procedure. The variance-covariance matrix for the estimate of β is given by M_{11}. The asymptotic variance of $\hat{\theta}_g$ is given by

$$2\hat{\theta}_g^2 \left\{N - 2\hat{\theta}_g^{-1} \operatorname{tr}\left(M_{22(g)}\right) + \hat{\theta}_g^{-2} \operatorname{tr}\left(M_{22(g)}^2\right)\right\}^{-1}. \tag{13}$$

3.2. Approximate marginal likelihood method

Using the derivation of a penalized likelihood solution obtained by Breslow and Clayton (1993) for the generalized linear mixed model assuming Gaussian random effects, Ripatti and Palmgren (2000) presented a parallel approximation for model (3).

3.2.1. Penalized partial likelihood

Ripatti and Palmgren (2000) approximate the marginal likelihood (6) using the Laplace approximation. For Gaussian random effects the marginal likelihood (6) can be rewritten as

$$L^{\mathrm{M}}(\beta, \theta, \lambda_0) = c|D(\theta)|^{-N/2} \int \mathrm{e}^{-\kappa(b)}\, \mathrm{d}b,$$

where

$$\kappa(b) = \sum_{i=1}^{N} \left[\sum_{j=1}^{n_i} \left\{ \delta_{ij} \left(\ln \lambda_0 \left(t_{ij}\right) + x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}} b_i \right) - \Lambda_0 \left(t_{ij}\right) \exp\left(x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}} b_i \right) \right\} - \frac{1}{2} b_i^{\mathrm{T}} \{D(\theta)\}^{-1} b_i \right]. \quad (14)$$

Let κ', κ'' denote the first and the second order partial derivatives of κ with respect to b. Ignoring a multiplicative constant, the approximation of the logarithm of the marginal likelihood takes the form:

$$l^{\mathrm{M}}(\beta, \theta, \lambda_0) \approx -\frac{N}{2} \ln |D(\theta)| - \frac{1}{2} \ln |\kappa''(\tilde{b})| - \kappa(\tilde{b}), \quad (15)$$

with $\tilde{b} = \tilde{b}(\beta, \theta)$ the solution to $\kappa'(\tilde{b}) = 0$.

Ripatti and Palmgren (2000) show that, for fixed θ, the values $\hat{\beta}(\theta)$ and $\hat{b}(\theta)$, which maximize the penalized log-likelihood (14), also maximize the penalized partial log-likelihood

$$l^{\mathrm{PPL}}(\beta, \theta, \lambda_0, b) = \sum_{i=1}^{N} \left[\sum_{j=1}^{n_i} \delta_{ij} \left\{ \left(x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}} b_i \right) - \ln \sum_{t_{kl} \geqslant t_{ij}} \exp\left(x_{kl}^{\mathrm{T}}\beta + z_{kl}^{\mathrm{T}} b_k \right) \right\} - \frac{1}{2} b_i^{\mathrm{T}} D(\theta)^{-1} b_i \right]. \quad (16)$$

Note that the penalized partial log-likelihood is just the sum of the elements on Eqs. (9) and (10) containing the parameters of interest (β and b). The estimating equations for $\beta(\theta)$ and $b(\theta)$, for a given θ, are of the form:

$$\sum_{i=1}^{N} \sum_{j=1}^{n_i} \delta_{ij} \left\{ x_{ij} - \frac{x_{ij} \exp\left(x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}} b_i \right)}{\sum_{t_{kl} \geqslant t_{ij}} \exp\left(x_{kl}^{\mathrm{T}}\beta + z_{kl}^{\mathrm{T}} b_k \right)} \right\} = 0, \quad (17)$$

$$\sum_{i=1}^{N} \left[\sum_{j=1}^{n_i} \delta_{ij} \left\{ z_{ij} - \frac{z_{ij} \exp\left(x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}} b_i \right)}{\sum_{t_{kl} \geqslant t_{ij}} \exp\left(x_{kl}^{\mathrm{T}}\beta + z_{kl}^{\mathrm{T}} b_k \right)} \right\} - \{D(\theta)\}^{-1} b_i \right] = 0, \quad (18)$$

where the product of a scalar by a vector results in multiplying each elements of the vector by the scalar.

Ripatti and Palmgren (2000) propose to find $\hat{\beta}(\theta)$ and $\hat{b}(\theta)$ by alternating between solving the Eqs. (17) and (18). Once $\hat{\beta}(\theta)$ and $\hat{b}(\theta)$ are computed, θ is updated by maximizing the approximate profile log-likelihood derived from (15):

$$l^{\mathrm{M}}(\beta, \theta, \lambda_0) \approx -\frac{N}{2} |\ln\ D(\theta)| - \frac{1}{2} \ln |\kappa''(\hat{b})| - \frac{1}{2} \hat{b}^{\mathrm{T}} \{D(\theta)\}^{-1} \hat{b}. \quad (19)$$

Ripatti and Palmgren (2000)propose to use $\kappa''_{\mathrm{PPL}}(b) = (\partial^2 l^{\mathrm{PPL}})/(\partial b \partial b^{\mathrm{T}})$ instead of $\kappa''(b)$, given its better empirical performance. An estimating equation for θ can be obtained after differentiation of (19) and some simplifications. In the particular case of a diagonal $D(\theta)$ the solution of the estimating equation takes the following simple form:

$$\hat{\theta}_g = \frac{\hat{b}_g^{\mathrm{T}} \hat{b}_g + \mathrm{tr}\left\{ \kappa''_{\mathrm{PPL}}(\hat{b})^{-1}_{(g)} \right\}}{N}, \quad (20)$$

where $\mathrm{tr}\left\{ \kappa''_{\mathrm{PPL}}(\hat{b})^{-1}_{(g)} \right\}$ indicates the sum of the elements of the diagonal of the submatrix of $\kappa''_{\mathrm{PPL}}(\hat{b})^{-1}$ associated with the gth component of the random effects.

3.2.2. Variance estimation

Estimates of the variance-covariance matrix of the fixed effects can be obtained using standard Cox regression with the estimated random effects as an offset. In order to estimate the variance-covariance matrix of $\hat{\theta}$ it is necessary to differentiate (19) twice with respect to θ and take the expectation with respect to b. Under the assumed diagonal form of $D(\theta)$, the variance for $\hat{\theta}$ is given by

$$\mathrm{var}(\hat{\theta}_g) = 2\hat{\theta}_g^2 \left[N + \frac{1}{\hat{\theta}_g^2} \mathrm{tr}\left\{ \kappa''_{\mathrm{PPL}}(\hat{b})^{-1}_{(g)} \kappa''_{\mathrm{PPL}}(\hat{b})^{-1}_{(g)} \right\} - \frac{2}{\hat{\theta}_g} \mathrm{tr}\left\{ \kappa''_{\mathrm{PPL}}(\hat{b})^{-1}_{(g)} \right\} \right]^{-1}. \tag{21}$$

3.3. Bayesian estimation approach

Ducrocq and Casella (1996) have proposed a Bayesian approach to estimate the parameters of the distribution of the random effects. In this approach the variance components related to the distribution of the random effects are estimated from their marginal posterior distribution after integrating out β and b. As this integration cannot be performed analytically, the Laplace approximation is used.

3.3.1. Laplace approximation of the marginal posterior distribution

Applying the Bayes theorem, the joint posterior density for model (3) is proportional to

$$L^{\mathrm{B}}(\beta, b, \theta \mid y) \propto L(y \mid \beta, b) \times \pi_0(b \mid \theta) \times \pi_0(\beta) \times \pi_0(\theta). \tag{22}$$

In this expression, the first factor is the partial likelihood (see (9)), while $\pi_0(b \mid \theta)$ is the joint normal density (see (10)).

Ducrocq and Casella (1996) assume a flat prior for θ and β

$$\pi_0(\theta) \propto 1 \quad \text{and} \quad \pi_0(\beta) \propto 1.$$

Therefore, the log joint posterior density is given by the sum of Eqs. (9) and (10). It is interesting to note that the term "posterior density" is in fact used here for convenience, acknowledging that it is obtained using the partial likelihood and not the full likelihood.

According to the Bayesian principle, estimation of the vector of variance components θ of the random effects should be based on its marginal posterior distribution after integrating out the nuisance parameters β and b:

$$L^{\mathrm{P}}(\theta \mid y) = \int L^{\mathrm{B}}(\beta, b, \theta \mid y)\, \mathrm{d}\beta\, \mathrm{d}b. \tag{23}$$

As this integration cannot be performed analytically, Ducrocq and Casella (1996) propose to approximate this marginal posterior density using the Laplace approximation. More precisely, for any given value θ^* of θ, they show that

$$L^{\mathrm{P}}\left(\theta^* \mid y\right) \approx \int \exp\left\{ l^{\mathrm{B}}\left(\hat{\Psi}_{\theta^*} \mid y, \theta^*\right) - \frac{1}{2}\left(\Psi - \hat{\Psi}_{\theta^*}\right)^{\mathrm{T}} H_{\theta^*} \left(\Psi - \hat{\Psi}_{\theta^*}\right) \right\} \mathrm{d}\beta\, \mathrm{d}b, \tag{24}$$

where H_{θ^*} is the negative Hessian matrix of the joint posterior distribution computed at the maximum $\hat{\Psi}_{\theta^*} = \left(\hat{\beta}_{\theta^*}, \hat{b}_{\theta^*}\right)$ of $l^{\mathrm{B}}\left(\beta, b \mid y, \theta = \theta^*\right)$. Recognizing under the integral sign of this last equation the kernel of a multivariate normal density with mean $\hat{\Psi}_{\theta^*}$ and variance H_{θ^*}, Ducrocq and Casella derive the following approximation of the marginal posterior density for any θ^* of θ:

$$l^{\mathrm{P}}\left(\theta^* \mid y\right) \approx \text{constant} + l^{\mathrm{B}}\left(\hat{\Psi}_{\theta^*} \mid y, \theta^*\right) - \tfrac{1}{2} \ln \left| H_{\theta^*} \right|. \tag{25}$$

The method of the simplex (Nelder and Mead, 1965) is used, in an upper level of iterations, to select the value of θ which maximizes this approximate log marginal posterior distribution. This value is taken as the point estimate for θ. Note that, instead of the mode, one could also use another point estimate, given that the whole posterior distribution of θ is available.

For any fixed value θ^* of θ, the log joint posterior density is maximized using a limited memory quasi-Newton method (Liu and Nocedal, 1989) to obtain point estimates $\hat{\beta}_{\theta^*}$ and $\hat{b}_{\theta^*}$ of β and b. The negative Hessian matrix H_{θ^*} is

then computed at this maximum. Based on $\hat{\Psi}_{\theta^*}$ and H_{θ^*}, the approximate log marginal posterior density at θ^*, obtained from formula (25), is computed.

3.3.2. Variance estimation

As for the Ripatti and Palmgren (2000) approach, estimates of the variance of the fixed effects β are easily obtained using standard Cox regression with the estimated random effects as an offset.

Estimates of the standard error of $\hat{\theta}$, as well as other point estimates of the distribution of $\hat{\theta}$, can be derived from the knowledge of the full marginal posterior density. To avoid repeated computations of (25), and in particular of the negative Hessian matrix H for many different values of θ, Ducrocq and Casella (1996) propose to summarize the general characteristics of the distribution (25) through the computation of its first three moments by numerical integration based on the Gauss–Hermite quadrature. It is worth mentioning that, in general, this distribution appears to be substantially skewed. It follows that computing the standard deviation of the marginal posterior distribution of θ and using it as standard error of the parameter can lead to overestimation. It is important to note that if we are interested to test whether $\theta > 0$, we can use the whole distribution to compute the confidence interval, without relying on asymptotic theory, what can be seen as an advantage of this approach.

3.4. The EM algorithm with the Laplace approximation

Cortiñas and Burzykowski (2005) developed an estimation method based on the use of the Laplace approximation at the E-step in the EM algorithm.

3.4.1. The E-step

In the E-step the expectation of the logarithm of likelihood (8), conditional on the observed data and on the current values $\beta^{(p)}$, $\theta^{(p)}$ and $\lambda_0^{(p)}(.)$ of parameters β, θ and $\lambda_0(.)$, respectively, is computed. The expectation, denoted by $Q(\beta, \theta, \lambda_0)$, can be written as

$$Q\left(\beta, \theta, \lambda_0\right) = Q_1\left(\beta, \lambda_0\right) + Q_2(\theta), \tag{26}$$

where

$$Q_1\left(\beta, \lambda_0\right) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left[\delta_{ij} \left\{\ln\, \lambda_0\left(t_{ij}\right) + x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}} E\left(b_i\right)\right\} - \Lambda_0\left(t_{ij}\right) \exp\left\{x_{ij}^{\mathrm{T}}\beta + \ln E\left(\mathrm{e}^{z_{ij}^{\mathrm{T}} b_i}\right)\right\}\right] \tag{27}$$

and

$$Q_2(\theta) = -\frac{1}{2} \sum_{g=1}^{d} \left\{N \ln\left(2\pi\theta_g\right) + \sum_{i=1}^{N} \frac{E\left(b_{ig}^2\right)}{\theta_g}\right\}, \tag{28}$$

with $E(.)$ denoting the expected values. To simplify the notation, the dependence of the expected values in (27) and (28) on the observed data and $\beta^{(p)}$, $\theta^{(p)}$ and $\lambda_0^{(p)}(.)$ has been suppressed. Note that $Q_1(\beta, \lambda_0)$ is just the conditional log-likelihood (4), where the random effects b_i are replaced by their expectations. It is important to note that the expectations in Eqs. (27) and (28) will not be available in a closed form. The conditional expectations that need to be computed involve integrals of the form

$$E\left\{g\left(b_i\right)\right\} = \frac{\int g\left(b_i\right) \mathrm{e}^{l_i^{\mathrm{C}}\left(\beta^{(p)}, \lambda_0^{(p)}, b_i\right) + \ln f\left(b_i, \theta^{(p)}\right)} \mathrm{d}b_i}{\int \mathrm{e}^{l_i^{\mathrm{C}}\left(\beta^{(p)}, \lambda_0^{(p)}, b_i\right) + \ln f\left(b_i, \theta^{(p)}\right)} \mathrm{d}b_i}. \tag{29}$$

Cortiñas and Burzykowski (2005) propose to use the Laplace formula to compute these expectations. Using the formula results in the approximations

$$E\left\{g\left(b_i\right)\right\} \approx g\left(\widetilde{b}_i\right), \tag{30}$$

where $\widetilde{b_i}$ is an isolated global minimum of

$$k\left(b_i\right) = -\frac{1}{n_i}\left\{l_i^{\mathrm{C}}\left(\beta^{(p)}, \lambda_0^{(p)}, b_i\right) + \ln f\left(b_i, \theta^{(p)}\right)\right\}. \tag{31}$$

The set of initial values for β and $\lambda_0(.)$ are obtained using the Cox regression without random effects. The initial values for θ can be specified by taking $D(\theta)$ equal to, e.g., the identity matrix.

3.4.2. The M-step

In the M-step new estimates $\beta^{(p+1)}$ and $\theta^{(p+1)}$ are found by maximizing the functions Q_1 and Q_2, respectively. To estimate β the profile likelihood approach is used. Assuming no ties, in order to keep notation simple, the value of the baseline hazard which maximizes Q_1 is

$$\lambda_m^{(p+1)} = \frac{1}{\sum_{t_{kl} \geqslant t_m} \exp\left\{x_{kl}^{\mathrm{T}}\beta^{(p)} + z_{kl}^{\mathrm{T}} b_k\right\}}, \tag{32}$$

where $\lambda_m = \lambda_0\left(t_m\right)$ and t_m $(m = 1, \ldots, r)$ are the distinct uncensored failure times. Substituting (32) into Q_1 gives the following profile-likelihood for β:

$$Q_1'(\beta) = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\delta_{ij}\left[x_{ij}^{\mathrm{T}}\beta - \ln\sum_{t_{kl} \geqslant t_{ij}} \exp\left\{x_{kl}^{\mathrm{T}}\beta + \ln E\left(\mathrm{e}^{z_{kl}^{\mathrm{T}} b_k}\right)\right\}\right]. \tag{33}$$

The form of (33) resembles that of the partial log-likelihood for the Cox proportional hazards model with offsets $\ln E\left(\mathrm{e}^{z_{ij}^{\mathrm{T}} b_i}\right)$. New value $\beta^{(p+1)}$ of β is obtained by maximizing Q_1' using standard software for the Cox model, as in the method proposed by Ripatti and Palmgren (2000).

Given that the density of the random effects b_i belongs to the exponential family, the estimation of $D(\theta)$ is generally straightforward. Hence, maximizing Q_2 leads to the estimator

$$\hat{D}(\theta) = \frac{1}{N}\sum_{i=1}^{N} E\left(b_i b_i^{\mathrm{T}}\right). \tag{34}$$

3.4.3. Variance estimation

The variance-covariance matrix of the solution $\left(\hat{\beta}, \hat{\lambda}_0, \hat{\theta}\right)$ obtained from the EM algorithm can be estimated using the inverse of the observed information matrix computed from the formula proposed by Louis (1982):

$$I\left(\beta, \lambda_0, \theta\right) = \left[E\left\{-l^{A''}\left(\beta, \lambda_0, \theta\right)\right\} - E\left\{l^{A'}\left(\beta, \lambda_0, \theta\right) l^{A'}\left(\beta, \lambda_0, \theta\right)^{\mathrm{T}}\right\}\right], \tag{35}$$

where $l^{A'}$ and $l^{A''}$ are the first and the second derivatives with respect to $(\beta, \lambda_0, \theta)$ of the logarithm of the "augmented" likelihood (8).

In order to compute standard error for the fixed parameters and the variance component of the random effects, it would be necessary to invert $I\left(\beta, \lambda_0, \theta\right)$. The dimension of the matrix $I\left(\beta, \lambda_0, \theta\right)$ can be very large, since it depends on $\lambda_0(.)$ and hence on the number of distinct uncensored failure times. Cortiñas and Burzykowski (2005) proposed to estimate the standard error of the parameter of interest by inverting only the relevant blocks of the matrix, corresponding to β and θ.

4. Case studies

In this section two case studies will be presented and analyzed using each of the four estimation methods described in the previous section. The first study was described by Duchateau et al. (2002) and analyzed later also by Cortiñas and Burzykowski (2005). The second case study was presented by Legrand et al. (2005).

Table 1
Results of the analysis of the disease free survival data of the patients included in the breast cancer trial (standard error in parentheses)

Method	β	σ_0^2	σ_1^2
McGilchrist	0.162 (0.068)	0.050 (0.032)	9×10^{-5} (0.016)
EM-Laplace	0.162 (0.068)	0.054 (0.026)	8×10^{-5} (0.013)
Ripatti	0.162 (0.069)	0.051 (0.028)	9×10^{-5} (0.014)
Ducrocq	0.162 (0.068)	0.052 (0.064)	1×10^{-5} (0.100)

4.1. Analysis of survival data in a breast cancer clinical trial

In this case study we will use data on disease free survival time of patients from an European Organization for the Research and Treatment of Cancer (EORTC) early breast cancer clinical trial comparing peri-operative chemotherapy with surgery alone (Clahsen et al., 1996). The trial included 15 centers, with the following number of patients per center: 6, 19, 25, 39, 48, 53, 54, 60, 78, 184, 185, 206, 311, 622, 902. Duchateau et al. (2002) used this trial to study the between-center variability in the baseline hazard. To this aim, they applied a shared frailty model with a gamma-distributed frailty to model disease free survival. As a result, they estimated baseline hazard (assumed constant) and the hazard ratio for the surgery-alone treatment to equal 0.07 and 1.16, respectively. The variance of the frailty distribution was estimated to be equal to 0.092.

The data were re-analyzed by Cortiñas and Burzykowski (2005), allowing for the variation in both the baseline hazard and the treatment effect using the following model:

$$\lambda\left(t_{ij}|\beta, b_{i0}, b_{i1}\right) = \lambda_0\left(t_{ij}\right) \mathrm{e}^{b_{i0}+x_{ij}(\beta+b_{i1})}, \tag{36}$$

where

$$\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \sim N_2\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}\right\}. \tag{37}$$

Cortiñas and Burzykowski (2005) obtained similar estimated values for the baseline hazard, hazard ratio and variance associated to the random center effects to those obtained by Duchateau et al. (2002). The correlation between random effects b_{i0} and b_{i1} was estimated to equal 0.37.

Here we will re-analyze the data using model (36) but ignoring the correlation between random effects ($\sigma_{01} = 0$), given its small estimated value obtained by Cortiñas and Burzykowski (2005). The model will be fitted to the data using each of the four estimation methods reviewed in Section 3.

The results of the four estimation methods are shown in Table 1. All the methods produce similar point estimates, consistently yielding a very small estimate (almost 0) for the variance of the random treatment effects. Some differences in standard errors of the estimated variance parameters can be seen (most notably for the Bayesian approach of Ducrocq and Cassella), however. The difference observed in the estimated standard errors may be attributable to the fact that for the Bayesian approach the estimated error is based on the posterior distribution of the variance estimates, while the other three methods use the asymptotic normality of the estimates. One would expected similar results if the posterior distribution were symmetric, which is not the case here (data not shown).

4.2. Analysis of disease free interval data in a bladder cancer clinical trial

Bladder cancer is a common urological malignancy and about 70–80% of all bladder cancers are superficial (stage Ta-T1). Standard treatment typically consists of transurethral resection (TUR) conducted with the aim of removing all tumors. However, a high proportion of patients experience recurrence or progression to muscle invasive disease, even after complete resection. Therefore, randomized phase III trials have been conducted over the last decades to investigate the use of prophylactic treatment following TUR. The objective of such treatment is both to remove residual, unresectable lesions and to prevent recurrence after complete resection.

Table 2
Results of the analysis of disease free interval of patients included in the bladder cancer trials (standard error in parentheses)

Method	β	σ_0^2	σ_1^2
McGilchrist	−0.100 (0.092)	0.107 (0.041)	0.106 (0.058)
EM-Laplace	−0.104 (0.061)	0.100 (0.024)	0.109 (0.026)
Ripatti	−0.100 (0.088)	0.106 (0.032)	0.100 (0.039)
Ducrocq	−0.101 (0.092)	0.108 (0.053)	0.109 (0.089)

In this case study, we consider the individual patient data of 2649 eligible bladder cancer patients randomized by 63 European centers in seven consecutive phase III randomized clinical trials conducted by the Genito-Urinary Group of the EORTC (EORTC 30781, Newling et al., 1995; 30782, Bouffioux et al., 1992, 1995; 30791, Kurth et al., 1984; 30831, 30832, Witjes et al., 1998; 30845 and 30863, Oosterlinck et al., 1993). All these patients had Ta-T1 bladder cancer, approximately half with primary bladder cancer and half with recurrent disease. Within the context of these trials, patients in each of the participating centers were treated with or without further intravesical treatment after TUR. The major baseline characteristics were, in general, well balanced over these two groups, with slightly more patients with multiple tumors and patients with Ta disease in the intravesical treatment group.

Considering the 35 centers which accrued at least 20 patients, our analysis is based on 2292 patients, 1004 (43.8%) without and 1445 patients (54.5%) with the further intravesical treatment. The number of patients per center varied from 21 to 249, with a median of 52 and mean of 65. The analyzed endpoint is disease free interval (time from randomization to the date of the first bladder recurrence, censoring the patients without recurrence at the date of last available follow up cytoscopy).

The data were analyzed using model (36), but assuming no correlation between random effects ($\sigma_{01} = 0$). The model was fitted to the data using each of the four estimation methods reviewed in Section 3.

The results obtained for the four estimation methods are shown in Table 2. As for the first case study, all the methods produce similar point estimates. However, more remarkable differences in standard errors of the estimated values can be seen. For instance, the EM algorithm proposed by Cortiñas and Burzykowski (2005) produces smallest standard errors. This might be due to the fact that Cortiñas and Burzykowski (2005) do not invert the whole information matrix given by (35) (see Section 3.4.3). On the other hand, the Bayesian approach of Ducrocq and Casella (1996) yields the highest values of standard errors. This can be again attributed to the use of the full posterior, rather than an asymptotic, distribution (see previous section).

5. Simulation study

The analyzes of two case studies presented in the previous section suggest that, while the point estimates obtained by the four different methods are quite similar, there may be differences in the standard error of the estimates. In order to investigate in more detail the performance of the different estimation methods, a simulation study is conducted using a setting similar to the second case study.

5.1. Simulation model and parameter settings

The data were generated using the following proportional hazards model:

$$\lambda\left(t_{ij}|\beta, b_{i0}, b_{i1}\right) = \lambda_0\left(t_{ij}\right) \mathrm{e}^{b_{i0}+x_{ij}^{\mathrm{T}}(\beta+b_{i1})}, \tag{38}$$

with

$$\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \sim N_2\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{pmatrix}\right\}. \tag{39}$$

Table 3
Interpretation of σ_0^2

σ_0^2	Median time-to-event for patients with $x_{ij} = 0$
0.04	(6.5–12.5 years)
0.08	(5.7–14.4 years)
0.4	(3.2–25.5 years)

Table 4
Interpretation of σ_1^2

σ_0^2	σ_1^2	Hazard ratio	Median time-to-event for patients with $x_{ij} = 1$
0.04	0.08	(1.26–3.21)	(2.6–7.9 years)
0.08	0.04	(1.45–2.80)	(2.6–7.9 years)
0.08	0.08	(1.26–3.21)	(2.4–8.7 years)
0.4	0.8	(0.46–8.77)	(0.8–27.1 years)

Model (38) corresponds to the setting of a multicenter clinical trial, in which heterogeneity appears both in the center-specific baseline hazards, as well as associated to the covariate. Parameters of the model were chosen to mimic data available in a real bladder cancer clinical trial database (Royston et al., 2004). In this simulation study we considered the similar distribution of patients over centers as in the real bladder dataset, namely 2323 patients accrued by 37 centers. Distribution of patients over centers was as follows: 21, 23, 23, 25, 26, 30, 30, 32, 34, 34, 34, 35, 35, 35, 37, 39, 41, 42, 42, 43, 52, 52, 53, 56, 61, 63, 66, 72, 85, 86, 91, 104, 116, 120, 155, 183, 247.

Two simulation settings were considered. In the first one moderately censored data (around 40%) were generated, while in the second one highly censored data (around 60%) were simulated. In both settings we simulated data using different combinations of values of σ_0^2 and σ_1^2, namely 0.04/0.08, 0.08/0.04, 0.08/0.08 and 0.4/0.8. A constant baseline hazard of $\lambda_0(t) \equiv 0.077$ was used. In order to have a better idea of the interpretation of the values of σ_0^2, σ_1^2 and λ_0, one might consider the spread of the median time-to-event from center to center. In our simulation b_{i0} and b_{i1} are normally distributed, thus the resulting distribution function for the median time-to-event is log-normal with parameters $\ln(\ln 2) - \ln \lambda_0 - \beta x_{ij}$ and $\sigma_0^2 + \sigma_1^2 x_{ij}^2$. For patients with $x_{ij} = 0$, the distribution does neither depend on β, nor on σ_1^2. For such patients, Table 3 presents, for each particular value of σ_0^2, the interval containing the median time-to-event of 90% of the centers.

In the next sections we describe details of both settings.

5.1.1. Moderate censoring setting

In this setting we consider a covariate x_{ij}, which divides the population in two groups: 30% of the patients have $x_{ij} = 0$ and 70% have $x_{ij} = 1$. The covariate can be seen as corresponding to, e.g., a prognostic index. The parameter β was set equal to 0.7, which corresponds to the estimated value for the real dataset. Given β and λ_0, we can compute the median time-to-event for a model without the random effects, which equals 9 in the group defined by $x_{ij} = 0$ and 4.5 in the group defined by $x_{ij} = 1$. Table 4 presents, for each particular value of σ_0^2 and σ_1^2, the intervals containing the hazard ratio of the effect of x_{ij} for 90% of the centers. Also, a corresponding interval for the median time-to-event in the group of patients with $x_{ij} = 1$ is given.

For each parameter setting, 250 datasets were generated in the following way. First, $N = 37$ random effects for the overall center effect and $N = 37$ random effects for the center-specific covariate effect were generated according to (39). We considered an accrual period (AP) of 1065 days (about 3 years) and a further follow up period (FP) of 2440 days (about 6.7 years). The actual observation for a patient was the minimum of the time-to-event and the time at risk. The former was generated using an exponential random variable with parameter $\lambda\left(t_{ij}|\beta_i, b_{i0}, b_{i1}\right)$ given by (38). The time at risk for a patient who entered in the study as kth subject at time $kAP/2323$ was defined as

$$\frac{AP(2323 - k)}{2323} + FP.$$

Table 5
Interpretation of σ_1^2

σ_0^2	σ_1^2	Hazard ratio	Median time-to-event for patients with $x_{ij} = 1$
0.04	0.08	(0.52–1.33)	(6.1–19.1 years)
0.08	0.04	(0.60–1.16)	(6.1–19.1 years)
0.08	0.08	(0.52–1.33)	(5.6–20.8 years)
0.4	0.8	(0.19–3.63)	(1.8–65.5 years)

These particular choices of the parameters resulted in approximately 60% of the individuals experiencing the event of interest.

5.1.2. Heavy censoring setting

In this setting, it was assumed that x_{ij} represented the treatment assignment. Hence, an equal split of patients in the two treatment groups $\left(x_{ij} = 0 \text{ and } x_{ij} = 1\right)$ was used. Assuming that the clinical trial takes place in good prognosis bladder cancer patients, and that the experimental treatment leads to 20% increase in the median disease free interval, i.e., from 9 to 10.8 years, we used a baseline hazard of 0.077 as in the "moderate censoring" setting, with β equal to -0.182. Given the values of β, λ_0, σ_0^2 and σ_1^2, we can compute intervals containing the hazard ratio of the effect of the covariate and median time-to-event for 90% of the centers (Table 5).

For each parameter setting, 250 datasets were generated. In this setting we considered an accrual period of 621 days (about 1.7 years) and a further follow up period of 2192 days (about 6 years). The generating mechanism for the data was similar to the one used in "moderate censoring" setting. As a result of the choices of the parameters, approximately 40% of the individuals in the datasets experienced the event of interest.

5.2. Numerical implementation

McGilchrist and Aisbett's approach was implemented using SAS-IML v8.2. The iterative procedure was stopped if the maximum of the relative difference between the fixed effects and variance estimates for two consecutive iterations was smaller than 10^{-3}. The method proposed by Ducrocq and Casella (1996) was implemented with The Survival Kit (Ducrocq and Sölkner, 1994, 1998) (www.boku.ac.at/nuwi/software/softskit.htm), a package of Fortran programs developed in the field of animal genetics to estimate survival models with random effects. The joint estimation of two variance components was not implemented in the original version. Therefore, we used a modified version proposed by Legrand et al. (2005). In this case, the convergence criterion required that the standardized norm of the vector of first derivatives of $l^{\mathrm{B}}\left(\beta, b|y, \theta = \theta^*\right)$ at its maximum had to be less than 10^{-8}. The EM algorithm proposed by Cortiñas and Burzykowski (2005) was implemented using SAS-IML v8.2. The EM algorithm was stopped when the maximum of the absolute changes for the fixed effects, the variance estimates and the log-likelihood was smaller than 10^{-5}. The method proposed by Ripatti and Palmgren (2000) was applied using the S+ functions developed by Therneau (2003). In this case the convergence criterion was the relative change in log-likelihood smaller than 10^{-4}.

5.3. The analysis model and robustness of the estimation methods

The main part of the simulation study was devoted to the investigation of the performance of the four estimation methods under the correctly specified model. That is, the data were generated under model (38), and the same model was used in the analysis.

A limited attempt to investigate the robustness of the estimation methods was also undertaken. All the methods assume proportional hazards, while allowing for an unspecified baseline hazard and normally distributed random effects. Thus, one could consider their performance when, for instance, the proportionality of hazards assumption is violated, or when the distribution of the random effects is not normal. In our investigation we focused on the latter. To this aim, we simulated the data using the proportional hazards model (38) under "moderate censoring," but with (independent) b_{i0} and b_{i1} distributed according to a symmetric Laplace distribution with mean 0 and variances equal to 0.4 and 0.8, respectively. The two Laplace distributions have the same means and variances as the normal distributions, but thicker tails.

6. Results of the simulations

6.1. Moderate censoring setting

Table 6 presents the results of the simulation for the moderate censoring setting. In this setting, none of the methods used in the simulations experienced convergence difficulties. The parameter β was, in general, estimated well by all the methods, with a relative absolute bias less than 4% in any of the considered cases. The bias increased with increasing σ_0^2 and σ_1^2. The bias of the estimates obtained by the Ripatti and Palmgren (2000) approach, in the case of $\sigma_0^2 = 0.04$ and $\sigma_1^2 = 0.08$, was higher than for the other methods. On the other hand, the approach of Ducrocq and Casella (1996) produced the largest bias when $\sigma_0^2 = 0.4$ and $\sigma_1^2 = 0.8$. The variability of estimates of β, measured by the empirical standard error, was similar for all the methods compared. In general, the Ripatti and Palmgren approach produced fixed-effects estimates with the largest empirical standard error. It is important to note, that the model-based estimates produced by Ducrocq and Casella (1996) and McGilchrist and Aisbett (1991) approaches were closer to the empirical standard error than for the other two methods.

The estimates σ_0^2 and σ_1^2 for all the methods were on average comparable. The version of the EM algorithm proposed by Cortiñas and Burzykowski (2005) yielded estimates with, in general, the smallest empirical variability, while Ducrocq and Casella's approach gave estimates with the largest variability. The Ducrocq and Casella approach tended to overestimate the empirical variability, while the Ripatti and Palmgren method and the version of the EM algorithm proposed by Cortiñas and Burzykowski (2005) tended to underestimate it. The model-based standard errors produced by the approach of McGilchrist and Aisbett were in most of the cases closer to the empirical standard error than for the other methods.

Table 7 shows the mean squared error (MSE) for each parameter in each of the settings studied. For the "moderate censoring" one can observe that the MSE of the estimates of β was consistently the smallest for the McGilchrist and Aisbett approach. On the other hand, the EM algorithm proposed by Cortiñas and Burzykowski (2005) tended to produce the smallest MSE for σ_0^2. Overall, however, no pattern can be seen that would indicate that one method is preferable to all other in all circumstances.

Table 6
Moderate censoring setting

Method	$\hat{\beta}$	$\hat{\sigma}_0^2$	$\hat{\sigma}_1^2$
	$\sigma_0^2 = 0.04$ *and* $\sigma_1^2 = 0.08$		
McGilchrist	0.705 (0.074; 0.073)	0.040 (0.023; 0.023)	0.078 (0.035; 0.034)
Ripatti	0.715 (0.119; 0.094)	0.038 (0.023; 0.024)	0.079 (0.026; 0.039)
EM-Laplace	0.702 (0.055; 0.079)	0.044 (0.021; 0.023)	0.083 (0.030; 0.031)
Ducrocq	0.703 (0.081; 0.078)	0.042 (0.033; 0.025)	0.079 (0.047; 0.039)
	$\sigma_0^2 = 0.08$ *and* $\sigma_1^2 = 0.04$		
McGilchrist	0.701 (0.067; 0.068)	0.079 (0.030; 0.027)	0.043 (0.029; 0.028)
Ripatti	0.702 (0.083; 0.095)	0.076 (0.025; 0.030)	0.039 (0.021; 0.030)
EM-Laplace	0.693 (0.055; 0.071)	0.075 (0.022; 0.028)	0.055 (0.024; 0.026)
Ducrocq	0.713 (0.073; 0.079)	0.085 (0.041; 0.039)	0.039 (0.039; 0.031)
	$\sigma_0^2 = 0.08$ *and* $\sigma_1^2 = 0.08$		
McGilchrist	0.703 (0.074; 0.079)	0.084 (0.033; 0.032)	0.075 (0.038; 0.038)
Ripatti	0.702 (0.082; 0.085)	0.077 (0.025; 0.036)	0.077 (0.026; 0.042)
EM-Laplace	0.696 (0.055; 0.085)	0.076 (0.028; 0.031)	0.080 (0.029; 0.032)
Ducrocq	0.706 (0.081; 0.082)	0.084 (0.046; 0.039)	0.082 (0.053; 0.042)
	$\sigma_0^2 = 0.4$ *and* $\sigma_1^2 = 0.8$		
McGilchrist	0.689 (0.160; 0.150)	0.405 (0.122; 0.121)	0.797 (0.224; 0.213)
Ripatti	0.691 (0.122; 0.165)	0.383 (0.101; 0.121)	0.770 (0.195; 0.213)
EM-Laplace	0.698 (0.156; 0.166)	0.375 (0.083; 0.092)	0.765 (0.189; 0.219)
Ducrocq	0.672 (0.161; 0.167)	0.402 (0.162; 0.140)	0.752 (0.218; 0.194)

The mean estimates for 250 simulated datasets for the four different methods. In parentheses: the mean model-based and empirical (first and second number) standard error.

Table 7
Mean squared error for the parameters for the four different methods ($\times 10^{-3}$)

Method	Moderate censoring setting			Heavy censoring setting		
	$\hat{\beta}$	$\hat{\sigma}_0^2$	$\hat{\sigma}_1^2$	$\hat{\beta}$	$\hat{\sigma}_0^2$	$\hat{\sigma}_1^2$
	$\sigma_0^2 = 0.04$ *and* $\sigma_1^2 = 0.08$					
McGilchrist	5.35	0.53	1.16	6.97	0.53	2.81
Ripatti	9.06	0.58	1.52	17.16	0.50	2.12
EM-Laplace	6.25	0.55	0.97	9.03	0.49	0.90
Ducrocq	6.09	0.63	1.52	6.95	0.58	2.30
	$\sigma_0^2 = 0.08$ *and* $\sigma_1^2 = 0.04$					
McGilchrist	4.63	0.73	0.79	7.06	1.16	1.45
Ripatti	9.03	0.92	0.90	26.28	0.84	1.30
EM-Laplace	5.09	0.81	0.90	5.97	0.63	0.76
Ducrocq	6.41	1.55	0.96	6.21	1.23	1.45
	$\sigma_0^2 = 0.08$ *and* $\sigma_1^2 = 0.08$					
McGilchrist	6.25	1.04	1.47	7.06	1.23	2.04
Ripatti	7.23	1.31	1.77	7.97	1.03	2.61
EM-Laplace	7.24	0.98	1.02	8.48	0.97	0.97
Ducrocq	6.76	1.54	1.77	7.06	1.17	2.75
	$\sigma_0^2 = 0.4$ *and* $\sigma_1^2 = 0.8$					
McGilchrist	22.62	14.67	45.38	27.12	14.71	60.91
Ripatti	27.31	14.93	46.27	26.28	13.99	64.73
EM-Laplace	27.56	9.09	49.19	25.07	9.06	41.16
Ducrocq	28.67	19.60	39.94	24.82	14.68	17.73

6.2. Heavy censoring setting

Table 8 shows the results of the simulation for the heavy censoring setting. In this setting, Ripatti and Palmgren's approach and the EM algorithm with the Laplace approximation experienced convergence problems (up to 14 and 17%, respectively). The mean estimated values reported here were based only on the cases when convergence was reached.

Fixed effects β were, in general, well estimated for $\sigma_0^2 = 0.04$ and $\sigma_1^2 = 0.08$ and for $\sigma_0^2 = 0.08$ and $\sigma_1^2 = 0.04$. For these cases, the relative bias was smaller than 9%. The picture was somewhat different when the variances increased. The relative bias for the fixed effects was still smaller than 9% for McGilchrist and Aisbett's approach, but, in general, it reached almost 15% for the other three approaches. Similar to the previous "moderate censoring" setting, the estimates for the Ripatti and Palmgren approach showed, in general, the largest empirical variability, while those for the Ducrocq and Casella method produced, in general, estimates with the smallest variability. Note also, that the model-based standard error for the approach proposed by the Ducrocq and Casella (1996) was the closest to the empirical value.

The estimates of variances of the random effects, similarly to the previous setting, were comparable for all the methods. The McGilchrist and Aisbett method produced model-based standard errors very close to the true variability. The Ripatti and Palmgren approach and the version of the EM algorithm proposed by Cortiñas and Burzykowski (2005) underestimated the empirical variability, while the method of Ducrocq and Casella overestimated it. Of the three last methods, the version of the EM algorithm proposed by Cortiñas and Burzykowski (2005), yielded model-based standard errors closest to the empirical standard error.

From Table 7 one can observe that under the "heavy censoring," the Ducrocq and Cassella approach was producing the estimates of β with the smallest MSE. On the other hand, the EM algorithm proposed by Cortiñas and Burzykowski (2005) tended to produce the smallest MSE for both σ_0^2 and σ_1^2. Thus, similarly to the "moderate censoring" case, no pattern can be seen that would clearly favor one of the methods in all circumstances.

6.3. Robustness of the estimation methods

The results of the simulations investigating the effect of misspecifying the analysis model (assuming a normal distribution for Laplace-distributed random effects) are shown in Table 9. It can be seen that the mean estimated fixed

Table 8
Heavy censoring setting

Method	$\hat{\beta}$	$\hat{\sigma}_0^2$	$\hat{\sigma}_1^2$	Convergence (%)
	$\sigma_0^2 = 0.04$ *and* $\sigma_1^2 = 0.08$			
McGilchrist	−0.173 (0.085; 0.083)	0.039 (0.026; 0.023)	0.081 (0.042; 0.053)	100
Ripatti	−0.166 (0.092; 0.130)	0.044 (0.021; 0.022)	0.080 (0.036; 0.046)	86.4
EM-Laplace	−0.168 (0.067; 0.094)	0.043 (0.015; 0.022)	0.079 (0.026; 0.030)	83.2
Ducrocq	−0.167 (0.084; 0.082)	0.042 (0.031; 0.024)	0.080 (0.031; 0.048)	100
	$\sigma_0^2 = 0.08$ *and* $\sigma_1^2 = 0.04$			
McGilchrist	−0.181 (0.077; 0.084)	0.080 (0.034; 0.034)	0.043 (0.036; 0.038)	100
Ripatti	−0.176 (0.100; 0.162)	0.079 (0.024; 0.029)	0.039 (0.034; 0.036)	94.4
EM-Laplace	−0.176 (0.067; 0.077)	0.078 (0.022; 0.026)	0.056 (0.015; 0.020)	88.6
Ducrocq	−0.171 (0.077; 0.078)	0.082 (0.042; 0.035)	0.042 (0.056; 0.038)	100
	$\sigma_0^2 = 0.08$ *and* $\sigma_1^2 = 0.08$			
McGilchrist	−0.169 (0.083; 0.083)	0.081 (0.036; 0.035)	0.076 (0.045; 0.045)	100
Ripatti	−0.167 (0.086; 0.088)	0.081 (0.024; 0.032)	0.082 (0.037; 0.051)	97.2
EM-Laplace	−0.168 (0.066; 0.091)	0.082 (0.028; 0.031)	0.082 (0.029; 0.031)	94.8
Ducrocq	−0.169 (0.086; 0.083)	0.083 (0.044; 0.034)	0.087 (0.072; 0.052)	100
	$\sigma_0^2 = 0.4$ *and* $\sigma_1^2 = 0.8$			
McGilchrist	−0.197 (0.167; 0.164)	0.392 (0.126; 0.121)	0.763 (0.232; 0.244)	100
Ripatti	−0.156 (0.143; 0.160)	0.392 (0.078; 0.118)	0.765 (0.180; 0.252)	100
EM-Laplace	−0.155 (0.129; 0.156)	0.385 (0.085; 0.094)	0.766 (0.165; 0.200)	100
Ducrocq	−0.160 (0.165; 0.156)	0.406 (0.151; 0.121)	0.767 (0.227; 0.129)	100

The mean estimates for 250 simulated datasets for the four different methods. In parentheses: the mean model-based and empirical (first and second number) standard error.

Table 9
Moderate censoring setting using Laplace-distributed random effects

Method	β	σ_0^2	σ_1^2
McGilchrist	0.692 (0.158; 0.151)	0.376 (0.118; 0.147)	0.736 (0.211; 0.285)
Ripatti	0.689 (0.120; 0.151)	0.376 (0.119; 0.152)	0.769 (0.215; 0.291)
EM-Laplace	0.685 (0.127; 0.162)	0.374 (0.106; 0.151)	0.730 (0.201; 0.277)
Ducrocq	0.703 (0.158; 0.148)	0.406 (0.160; 0.159)	0.727 (0.194; 0.225)

The mean estimates for 250 simulated datasets for the four different methods. In parentheses: the mean model-based and empirical (first and second number) standard error.

effects are close to the true value of 0.7 (relative bias of at most 2.1%) and very similar to those obtained for the moderate censoring setting when the simulation and analysis models coincided (see Section 6.1). The empirical variability and the model-based standard errors are somewhat smaller for the misspecified case. MSE for all the methods is comparable (around 2.3%).

The absolute relative bias for the variance parameters is larger (up to around 9%) for the misspecified analysis model than for the correct model (in general, less than 5%). The empirical variability of the estimates is considerably larger for the former. In general, for all methods the mean model-based standard errors underestimate the empirical ones with the absolute relative bias ranging from 16% to 42% (except for the Ducrocq and Casella approach for σ_0^2, for which it is 1%). For all methods MSE is around 2.3% for σ_0^2 and around 8.5% (except for the Ducrocq and Casella approach, for which it is 5.6%) for σ_1^2.

7. Concluding remarks

Proportional hazards models with multivariate random effects offer several advantages over univariate shared frailty models (Xue and Brookmeyer, 1996), especially when times from the same cluster are negatively associated. The main stumbling block in the use of the former models is estimation methods.

In this paper we compared the performance of four estimation methods, proposed by McGilchrist and Aisbett (1991), Ducrocq and Casella (1996), Ripatti and Palmgren (2000) and Cortiñas and Burzykowski (2005). The main reason for this choice was software availability and feasibility of conducting a simulation study.

Each of the methods seems to offer some advantages and drawbacks. For instance, under "heavy censoring" in the simulation study assuming the same simulation and analysis models, Ripatti and Palmgren's approach, as well as the method proposed by Cortiñas and Burzykowski (2005), experienced convergence problems. The non-convergence rate for the method proposed by Cortiñas and Burzykowski (2005) was somewhat higher. No such problems were seen for the other two methods. The estimates obtained by the modified version of the EM algorithm proposed by Cortiñas and Burzykowski (2005) seemed to express the smallest empirical variability. However, model-based standard errors underestimated the variability. Similar underestimation could be observed for the Ripatti and Palmgren approach. Ducrocq and Casella's method, on the other hand, produced conservative standard errors, especially for the estimated variance components. However, it can be noted that, in general, the distribution of the estimates of θ appears to be substantially skewed. From this point of view, the advantage of Ducrocq and Casella's approach is that it can provide an estimate of the whole distribution, which can be used for testing purposes.

In the simulation study, the point estimates produced by the four methods were, in general, comparable. However, their empirical variability differed. As a result, some differences in the MSE could be observed. For instance, under "moderate censoring" the MSE of the estimates of the fixed effect was the smallest for the McGilchrist and Aisbett approach, while under "heavy censoring" it was the smallest for the Ducrocq and Cassella approach. On the other hand, the EM algorithm proposed by Cortiñas and Burzykowski (2005) tended to produce the smallest MSE for the variance parameters. Overall, no clear pattern favoring one of the methods irrespectively of circumstances could be seen.

A small simulation study was conducted to investigate the robustness of the four methods to a misspecification of the distribution of the random effects. More concretely, the true (Laplace) distribution had thicker tails than the one (normal) assumed in the analysis model. There were no major, consistent differences in robustness of the compared methods. Even with this relatively mild misspecification the relative bias of the estimated variance parameters almost doubled. A considerable increase in the bias of the estimated standard errors was also observed. This warrants a more complete robustness study. This is a topic for future research.

Taking into account the above, it is difficult to formulate a clear recommendation. The choice of the method can depend, e.g., on the aim of the analysis (fixed effects, variance components), amount of censoring present in the data, magnitude of the variability of random effects. The results presented in this paper can offer some guidance in this respect.

Acknowledgments

The authors thank Chris Aisbett for a critical review of the manuscript that substantially improved its content and Richard Sylvester from the EORTC for useful comments. They gratefully acknowledge support from FWO-Vlaanderen Research Project "Sensitivity Analysis for Incomplete and Coarse Data" and Belgian IUAP/PAI network no. P5/24 "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data" of the Belgian Government (Belgian Science Policy). The authors thank the European Organization for Research and Treatment of Cancer for permission to use the data from EORTC trials 10854 (Breast Cancer Group) and 30781, 30782, 30791, 30831, 30832, 30845, and 30863 (Genito-Urinary Tract Cancer Group) for this research. The contents of this publication and methods used are solely the responsibility of the authors and do not necessarily represent the official views of the EORTC.

References

Bouffioux, C., Denis, L., Oosterlinck, W., Viggiano, G., Vergison, B., Keuppens, F., de Pauw, M., Sylvester, R., Cheuvart, B., 1992. Adjuvant chemotherapy of recurrent superficial transitional cell carcinoma: results of a European Organization for Research and Treatment of Cancer randomised trial comparing intravesical instillation of thiotepa, doxorubicin and cisplatin. J. Urology 148, 297–301.

Bouffioux, C., Kurth, K.H., Bono, A., Oosterlinck, W., Kruger, C.B., de Pauw, M., Sylvester, R., 1995. Intravesical adjuvant chemotherapy for superficial transitional cell bladder carcinoma: results of two European Organization for Research and Treatment of Cancer randomized trials with mitomycin C and doxorubicin comparing early versus delayed instillations and short-term versus long-term treatment European Organization for Research and Treatment of Cancer Genitourinary Group. J. Urology 153, 934–941.

Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized linear models. J. Amer. Statist. Assoc. 88, 9–25.
Clahsen, P.C., van de Velde, C.J., Julien, J.P., Floiras, J.L., Delozier, T., Mignolet, F.Y., Sahmoud, T.M., 1996. Improved local control and disease-free survival after preoperative chemotherapy for early-stage breast cancer. A European Organization for Research and Treatment of Cancer breast cancer cooperative group study. J. Clinical Oncology 14, 745–753.
Clayton, D.G., 1978. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika 65, 141–151.
Cortiñas, A.J., Burzykowski, T., 2005. A version of the EM algorithm for proportional hazards model with random effects. Biometrical J. 47, 847–862.
Duchateau, L., Janssen, P., Lindsey, P., Legrand, C., Nguti, R., Sylvester, R., 2002. The shared frailty model and the power for heterogeneity tests in multicenter trials. Comput. Statist. Data Anal. 40, 603–620.
Ducrocq, V., Casella, G., 1996. A Bayesian analysis of mixed survival models. Genet. Sel. Evol. 28, 505–529.
Ducrocq, V., Sölkner, J., 1994. The Survival Kit, a FORTRAN package for the analysis of survival data. Fifth World Congress on Genetics Applied to Livestock Production, vol. 22, Department of Animal of Poultry Science, University of Guelph, Guelph, Ont., Canada, pp. 51–52.
Ducrocq, V., Sölkner, J., 1998. The Survival Kit—V3.0 a package for large analyses of survival data. Sixth World Congress on Genetics Applied to Livestock Production, vol. 27, Animal Genetics and Breeding Unit, University of New England, Armidale, Australia, pp. 447–448.
Gustafson, P., 1997. Large hierarchical Bayesian analysis of multivariate survival data. Biometrics 53, 230–242.
Ha, I.D., Lee, Y., 2003. Estimating frailty models via Poisson hierarchical generalized linear models. J. Comput. Graphical Statist. 12, 663–681.
Ha, I.D., Lee, Y., Song, J., 2001. Hierarchical likelihood approach for frailty models. Biometrika 88, 233–243.
Hougaard, P., 2000. Analysis of Multivariate Survival Data. Springer, New York.
Klein, J.P., Moeschberger, M.L., 1997. Survival Analysis Techniques for Censored and Truncated Data. Springer, New York.
Kurth, K.H., Schröder, F.H., Tunn, U., Ay, R., Pavone-Macaluso, M., Debruyne, F., dePauw, M., Dalesio, O., tenKate, F., 1984. Adjuvant treatment of superficial transitional cell bladder carcinoma: preliminary results of an European Organization for Research and Treatment of Cancer randomized trial comparing doxorubicin hydrochloride, ethoglucid and transurethral resection alone. J. Urology 132, 258–262.
Legrand, C., Ducrocq, V., Janssen, P., Sylvester, R., Duchateau, L., 2005. A Bayesian approach to jointly estimate center and treatment by center heterogeneity in a proportional hazards model. Statist. Medicine 24, 3789–3804.
Liu, D.C., Nocedal, J., 1989. On the limited memory BFGS method for large scale optimization. Math. Programming 45, 503–528.
Louis, T.A., 1982. Finding the observed information matrix when using the EM algorithm. J. Roy. Statist. Soc. (Ser. B) 44, 190–200.
Ma, R., Krewski, D., Burnett, R.T., 2003. Random effects Cox models: a Poisson modelling approach. Biometrika 90, 157–169.
Maples, J.J., Murphy, S.A., Axinn, W.G., 2002. Two-level proportional hazards models. Biometrics 58, 754–763.
McGilchrist, C.A., 1993. REML estimation for survival models with frailty. Biometrics 49, 221–225.
McGilchrist, C.A., 1994. Estimation in generalized mixed models. J. Roy. Statist. Soc. (Ser. B) 56, 61–69.
McGilchrist, C.A., Aisbett, C.W., 1991. Regression with frailty in survival analysis. Biometrics 47, 461–466.
Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. Comput. J. 7, 308–313.
Newling, D.W., Robinson, M.R., Smith, P.H., Byar, D., Lockwood, R., Stevens, I., De Pauw, M., Sylvester, R., 1995. Tryptophan metabolites, pyridoxine (vitamin B6) and their influence on the recurrence rate of superficial bladder cancer Results of a prospective randomized phase III study performed by the EORTC GU Group. EORTC Genitourinary Tract CancerCooperative Group. European Urologist 27, 110–116.
Oakes, D., 1982. A model for association in bivariate survival data. J. Roy. Statist. Soc. (Ser. B) 44, 414–422.
Oosterlinck, W., Kurth, K.H., Schröder, F.H., Bultinck, J., Hammond, B., Sylvester, R., 1993. A prospective European Organization for Research and Treatment of Cancer Genitourinary Group randomized trial comparing transurethral resection followed by a single intravesical instillation of epirubicin or water in single stage TaT1 papillary carcinoma of the bladder. J. Urology 149, 749–752.
Ripatti, S., Palmgren, J., 2000. Estimation of multivariate frailty models using penalized partial likelihood. Biometrics 56, 1016–1022.
Ripatti, S., Larsen, K., Palmgren, J., 2002. Maximum likelihood inference for multivariate frailty models using an automated Monte Carlo EM algorithm. Lifetime Data Anal. 8, 349–360.
Royston, P., Parmar, M.K.B., Sylvester, R., 2004. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. Statist. Medicine 23, 907–926.
Sargent, D., 1998. A general framework for random effects survival analysis in the Cox proportional hazard setting. Biometrics 54, 1486–1497.
Sastry, N., 1997. A nested frailty model for survival data, with an application to the study of child survival in northeast Brazil. J. Amer. Statist. Assoc. 92, 426–435.
Schall, R., 1991. Estimation in generalised linear models with random effects. Biometrika 78, 719–727.
Sinha, D., Dey, D., 1997. Semiparametric Bayesian analysis of survival data. J. Amer. Statist. Assoc. 92, 1195–1212.
Therneau, T., 2003. On mixed effect Cox models, sparse matrices, and modelling data from large pedigree. Technical Report, July 2003.
Vaida, F., Xu, R., 2000. Proportional hazards model with random effects. Statist. Medicine 19, 3309–3324.
Vaupel, J.W., Manton, K.G., Stallard, E., 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 16, 439–454.
Witjes, J.A., van der Meijden, A.P., Sylvester, R.C., Debruyne, F.M., van Aubel, A., Witjes, W.P., 1998. Long term follow up of an EORTC randomized prospective trial comparing intravesical bacille Calmette-Guérin-RIVM and Mitomycin C in superficial bladder cancer. Urology 52, 403–410.
Xue, X., 1998. Multivariate survival data under bivariate frailty: an estimating equation approach. Biometrics 54, 1631–1637.
Xue, X., 2001. Analysis of childhood brain tumor data in New York City using frailty models. Statist. Medicine 20, 3459–3473.
Xue, X., Brookmeyer, R., 1996. Bivariate frailty model for the analysis of multivariate survival time. Lifetime Data Anal. 2, 277–289.
Xue, X., Ding, Y., 1999. Assessing heterogeneity and correlation of paired failure times with the bivariate frailty model. Statist. Medicine 18, 907–918.