

Joint Modeling of Binary Response and Survival
for Clustered data in Clinical Trials

Bingshu E. Chen¹

Canadian Cancer Trials Group

and Department of Public Health Sciences, Queen's University,

Kingston, Ontario, Canada

bechen@ctg.queensu.ca

Jia Wang²

Population Health Research Institute,

Hamilton, Ontario, Canada

¹Corresponding author.

²Both authors are co-first authors and contribute equally to the work.

Abstract

In clinical trials, it is often desirable to evaluate the effect of a prognostic factor such as a marker response on a survival outcome. However, the marker response and survival outcome are usually associated with some potentially unobservable factors. In this case, the conventional statistical methods that model these two outcomes separately may not be appropriate. In this paper, we propose a joint model for marker response and survival outcomes for clustered data, providing efficient statistical inference by considering these two outcomes simultaneously. We focus on a special type of marker response: a binary outcome, which is investigated together with survival data using a cluster-specific multivariate random effect variable. A multivariate penalized likelihood method is developed to make statistical inference for the joint model. However, the standard errors obtained from the penalized likelihood method are usually underestimated. This issue is addressed using a Jackknife resampling method to obtain a consistent estimate of standard errors. We conduct extensive simulation studies to assess the finite sample performance of the proposed joint model and inference methods in different scenarios. The simulation studies show that the proposed joint model has excellent finite sample properties compared to the separate models when there exists an underlying association between the marker response and survival data. Finally, we apply the proposed method to a symptom control study conducted by Canadian Cancer Trials Group to explore the prognostic effect of covariates on pain control and overall survival.

Key words: Generalized Linear Mixed Model; Laplace transformation; Jackknife; Marker response; Penalized likelihood; Random effects model; Survival Analysis

1 Introduction

To ensure adequate sample size and generalization of a study, most clinical trials are now conducted in a multi-centre manner. The patients within each centre potentially share similar characteristics. For example, they could have similar socioeconomic status and exposure to the same environmental factors. In immunotherapy trials, a new drug is often tested in multiple types of cancer patients in a single study. Accordingly, it is crucial to introduce cluster-level (e.g., either centre-specific or cancer-specific) random effects to account for the dependence of patients within each centre.¹ In some situations, it is of interest to study the treatment efficacy on a survival outcome. Early response to treatment is often considered as a surrogate measurement for potential efficacy. In many situations, marker responses are measured repeatedly, which generates longitudinal data, for example, CD4 counts or viral load over time by Neuhäus.² Over the past few decades extensive research has been conducted on joint modeling of longitudinal measurements and survival data. For example, Wu provided a comprehensive review of this topic.³ However, the association between binary marker response at a specific time point and survival outcome, although typical and important in cancer clinical trials, has not yet been addressed thoroughly. Inoue developed a parametric mixture model for binary tumor response and survival outcome under the Bayesian framework.⁴ Lai proposed a semi-parametric model that connected the binary response and survival outcome based on a frequentist approach in a sequential designing study of phase II - phase III cancer clinical trials.⁵ Recently, Hwang and others studied a joint model of longitudinal binary and survival data using the Monte Carlo Expectation Maximization (EM) algorithm.⁶

The joint model approach proposed in this paper for binary and survival data was motivated by a symptom control study from Canadian Cancer Trials Group (CCTG).⁷ This CCTG study was a multi-centre randomized controlled non-inferiority trial, in which 850 patients were randomized to receive either single or multiple fraction radiotherapy to control pain caused by metastatic bone lesions. The trial was conducted on multiple types of cancers with bone metastasis (e.g. bladder, breast, colon, esophageal, kidney, lung, marrow, prostate, and rectum cancers) and different painful bone lesion sites (e.g. limbs, hips, lumbar spine, thoracic spine, thoracolumbar spine, superficial bones, etc). This leads to many possible combinations of cancer types and painful bone lesions (more than 60 in this CCTG symptom control study). The effect of cancer with bone metastasis (e.g. colon cancer with painful lumbar bone lesion) was modeled via cluster-specific random effects. The study was designed to test the hypothesis that a radiotherapy of 8 Gy in single fraction is non-inferior to (or as good as) a radiotherapy of 20 Gy in multiple fraction for cancer patients with painful bone

lesions for pain control. Response to pain control treatment was evaluated two months after randomization. With the proposed model, we aim to investigate the potential association between treatment and binary response (pain control), and the association between treatment and overall survival, accounting for the effect of different cancer types and bone metastasis sites. Furthermore, we aim to elucidate if response to pain control treatment can be used as a potential surrogate for survival.

In this paper, we develop a Multivariate Penalized Likelihood (MPL) method to make statistical inference in the proposed joint model of the binary and survival outcomes. The MPL method is built on the penalized partial likelihood depicted by Ripatti for multivariate frailty models.⁸ The multivariate random effects in the proposed joint model are assumed to follow a bivariate normal distribution, similar to those used by Breslow.⁹ In addition, normally distributed random effects allow one to handle both positive and negative associations between marker response and survival outcome in the joint model, which is not possible when assuming a Gamma distributed frailty variable.⁸ We use the Laplace approximation to tackle the challenge of intensive computation for numerical integration.^{9,10} To make inference on the proposed MPL method, two approaches of variance estimation for the parameters are proposed: one based on the asymptotic distribution of the maximum likelihood estimates and the other based on the Jackknife method for clustered data.

We conduct numerical simulations to evaluate the finite sample properties of the proposed MPL method and show that it provides a computationally efficient alternative to the existing inference approaches. The proposed joint model is compared with models that treat the binary outcome and the survival outcome separately. The simulation studies suggest that the asymptotic method underestimates the variances of the parameter estimates, while the Jackknife method provides consistent variance estimation. Finally, we apply the proposed MPL method to a symptom control clinical trial. We identify a significant effect of pain control on overall survival outcome by taking the correlation among these two outcomes and the cancer types and bone metastasis sites specified effects into account.

2 Joint model for binary and survival data

We propose a joint model for clustered binary response and survival time through a bivariate random effect. Let m be the total number of clusters. For each cluster, let n_i be the number of subjects in center i , where $i = 1, 2, \dots, m$. Let $y_{ij} = 0$ or 1 be the observed marker response outcome Y_{ij} for subject j in cluster i and $\Pr\{Y_{ij} = 1\} = \pi_{ij}$, for $i = 1, \dots, m$ and $j = 1, 2, \dots, n_i$. Here $n = n_1 + n_2 + \dots + n_m$ is the total number of subjects in the study.

Let T and C be the failure and censoring time, respectively. We define $x_{ij} = \min\{t_{ij}, c_{ij}\}$ to be the observed failure or censoring time, and let $\delta_{ij} = 1$ if subject j in center i failed and $\delta_{ij} = 0$ otherwise. Let $Z_{n \times p}$ be the covariate matrix for the binary response outcome and $W_{n \times q}$ be the covariate matrix for the survival outcome, respectively. Here $W_{n \times q}$ may or may not be exactly the same as $Z_{n \times p}$. For example, to treat the short term marker response outcome $Y = (Y_{11}, \dots, Y_{1n_1}, \dots, Y_{m1}, \dots, Y_{mn_m})'$ as a potential predictor variable for the long term survival outcome, we may assume that Y is one of the columns in matrix W , and may interact with other covariates in W .

The Logistic regression model for the binary response outcome and the Cox model for the survival outcome are linked through a cluster specific zero mean random effect $\mathbf{u}_i = (u_{1i}, u_{2i})' \sim N(\mathbf{0}, \Sigma)$. The variance matrix of the random effect \mathbf{u}_i is given by

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix},$$

where σ_{11} and σ_{22} are the variances of random effects u_{1i} and u_{2i} , respectively, and σ_{12} is the covariance between u_{1i} and u_{2i} , which can be either a positive or a negative value. Let $\boldsymbol{\sigma} = (\sigma_{11}, \sigma_{22}, \sigma_{12})$ denote the parameters describing Σ . When $\sigma_{12} = 0$, there is no association between two outcomes and the joint model reduces to two separate models.

Conditional on the covariate vectors $\mathbf{z}_{ij}, \mathbf{w}_{ij}$ and random effect \mathbf{u}_i , we assume that the response Y_{ij} follows a Bernoulli distribution with $P\{Y_{ij} = 1 | \mathbf{z}_{ij}, u_{1i}\} = \pi_{ij}$ and the survival time T_{ij} has a hazard function $\lambda_{ij}(t)$ with a Cox proportional hazards model,¹¹

$$\begin{cases} \text{logit}(\pi_{ij} | \mathbf{z}_{ij}, u_{1i}) &= \mathbf{z}'_{ij} \boldsymbol{\beta} + u_{1i}, \\ \lambda_{ij}(t | \mathbf{w}_{ij}, u_{2i}) &= \lambda_0(t) \exp(\mathbf{w}'_{ij} \boldsymbol{\gamma} + u_{2i}), \end{cases} \quad (1)$$

where $\lambda_0(s)$ is the baseline hazard function. Let $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$, given u_{2i} we have the cumulative hazard function for subject j in cluster i

$$\Lambda_{ij}(t) = \Lambda_0(t) \exp(\mathbf{w}'_{ij} \boldsymbol{\gamma} + u_{2i}).$$

Given random effects (u_{1i}, u_{2i}) , the conditional likelihood functions $p_1(\boldsymbol{\beta}, u_{1i}) = (\pi_{ij})^{Y_{ij}} (1 - \pi_{ij})^{1 - Y_{ij}}$ and $p_2(\boldsymbol{\gamma}, u_{2i}) = \{\lambda_{ij}(x_{ij})\}^{\delta_{ij}} \exp\{-\Lambda_{ij}(x_{ij})\}$ are obtained from the logistic regression model and the Cox proportional hazards model, respectively. The marginal joint likelihood function for regression coefficients $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and variance parameter $\boldsymbol{\sigma}$ is given by

$$\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma}) = \sum_{i=1}^m \log \left\{ \int p_1(\boldsymbol{\beta}, u_{1i}) p_2(\boldsymbol{\gamma}, u_{2i}) \phi(u_{1i}, u_{2i}, \boldsymbol{\sigma}) du_{1i} du_{2i} \right\}, \quad (2)$$

where $\phi(\cdot)$ is the probability density function for a bivariate normal distribution with mean 0 and variance-covariance matrix Σ . We define the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma})$. Given the expressions for $p_1(\boldsymbol{\beta}, u_{1i})$ and $p_2(\boldsymbol{\gamma}, u_{2i})$, maximization of the likelihood function (2) involves high dimensional integration. Here, we apply the first-order Laplace transformation⁹ to approximate the joint likelihood function (2). Following the approaches of Breslow⁹ and Ripatt,⁸ for a given random effect variable $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m)$, we let

$$\begin{aligned}\ell_1(\boldsymbol{\beta}, \mathbf{u}) &= \sum_{i=1}^m \sum_{j=1}^{n_i} \left[y_{ij}(\mathbf{z}'_{ij}\boldsymbol{\beta} + u_{1i}) - \log\{1 + \exp(\mathbf{z}'_{ij}\boldsymbol{\beta} + u_{1i})\} \right] \\ \ell_2(\boldsymbol{\gamma}, \mathbf{u}) &= \sum_{i=1}^m \sum_{j=1}^{n_i} \left[\delta_{ij}(\log\{\lambda_0(x_{ij})\} + \mathbf{w}'_{ij}\boldsymbol{\gamma} + u_{2i}) - \Lambda_0(x_{ij}) \exp(\mathbf{w}'_{ij}\boldsymbol{\gamma} + u_{2i}) \right] \\ \ell_3(\boldsymbol{\sigma}, \mathbf{u}) &= \sum_{i=1}^m -\frac{1}{2} \mathbf{u}'_i \Sigma^{-1} \mathbf{u}_i.\end{aligned}$$

To simplify the expression of the conditional likelihood function for survival data, we consider the full likelihood approach similar to Andersen for $\ell_2(\boldsymbol{\gamma}, \mathbf{u})$.¹² The baseline cumulative hazard function $\Lambda_0(t)$ is treated as a nuisance parameter. We then follow the same idea as Murphy and van der Vaart¹³ by considering the Cox partial likelihood estimator $\boldsymbol{\gamma}$ as a maximum profile likelihood estimator. Given $\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}$ and $\mathbf{u} = \hat{\mathbf{u}}$, the baseline cumulative hazard function $\Lambda(t)$ can be estimated using the Nelson-Aalen method,^{14,15}

$$\hat{\Lambda}_0(t) = \sum_{x_{ij} \leq t} \frac{d_{ij}}{\sum_{i^*j^* \in R(x_{ij})} \exp\{\mathbf{w}'_{i^*j^*}\hat{\boldsymbol{\gamma}} + \hat{u}_{2i^*}\}}, \quad (3)$$

where $R(x)$ is the at risk set of individuals who are alive at time x . Combining $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ together leads to the non-parametric maximum likelihood estimators (NPMLEs). Our expression of the likelihood function is similar to the likelihood function in equation (3) of Murphy and van der Vaart.¹³ This is different from the approach in Ripatti⁸ and Ye,¹⁶ which uses the partial likelihood function of the Cox model.¹⁷

For any given value of parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, we define

$$K(\mathbf{u}) = -\{\ell_1(\boldsymbol{\beta}, \mathbf{u}) + \ell_2(\boldsymbol{\gamma}, \mathbf{u}) + \ell_3(\boldsymbol{\sigma}, \mathbf{u})\} \quad (4)$$

as a function of random effects \mathbf{u} . Let $K'(\mathbf{u})$ and $K''(\mathbf{u})$ be the first and second order partial derivatives for $K(\mathbf{u})$, respectively. Detailed derivations of $K'(\mathbf{u})$ and $K''(\mathbf{u})$ are given in the Appendix. Let $\tilde{\mathbf{u}} = \tilde{\mathbf{u}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ such that $K'(\tilde{\mathbf{u}}) = \mathbf{0}$, where $K'(\mathbf{u})$ is the partial derivative of $K(\mathbf{u})$ with respect to \mathbf{u} .

The full log likelihood function in equation (2) can be approximated by the following penalized log likelihood function, which will be the base of the main statistics method of this paper,

$$\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma}) \approx -\frac{m}{2} \log |\Sigma| - \frac{1}{2} \log(|K''(\tilde{\mathbf{u}})|) - K(\tilde{\mathbf{u}}), \quad (5)$$

where $|K''(\tilde{\mathbf{u}})| = \prod_{i=1}^m |A_i + \Sigma^{-1}|$. We show in the Appendix that $A_i = \text{diag}(a_{1i}, a_{2i})$ is a 2×2 matrix of parameter $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ with $a_{1i} = \sum_{j=1}^{n_i} \pi_{ij}(1 - \pi_{ij})$ and $a_{2i} = \sum_{j=1}^{n_i} \Lambda_{ij}(t_{ij})$. Here $\boldsymbol{\lambda}$ is the parameter for the baseline cumulative hazard function $\Lambda_0(t)$, which can be estimated by the Nelson-Aalen method (3). By simple algebra, we have

$$|K''(\tilde{\mathbf{u}})| = \prod_{i=1}^m \left\{ a_{1i}a_{2i} + |\Sigma|^{-1}(a_{1i}\sigma_{11} + a_{2i}\sigma_{22} + 1) \right\}.$$

Here $\log |\Sigma|$, the first term of equation (5), does not contain the regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Furthermore, according to Ripatt⁸ and Ye,¹⁶ $\log(|K''(\tilde{\mathbf{u}})|)$ in (5) has negligible effects on the estimation of regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Therefore, we ignore these two terms in the approximate log likelihood function (5) to simplify the computation. It is important to keep in mind that the removal of the first two terms of (5) may lead to loss of information in the statistic inference for the model parameters. We will address this issue in Sections 3 and 4 of this paper.

For a given variance parameter $\boldsymbol{\sigma}$, we can maximize $\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{u}) = \ell_1(\boldsymbol{\beta}, \mathbf{u}) + \ell_2(\boldsymbol{\gamma}, \mathbf{u}) + \ell_3(\mathbf{u})$ by alternately maximizing $\ell_1(\boldsymbol{\beta}, \mathbf{u}) + \ell_2(\boldsymbol{\gamma}, \mathbf{u})$ with respect to $\boldsymbol{\beta}, \boldsymbol{\gamma}$ for the current value of \mathbf{u} and $\ell_1(\boldsymbol{\beta}, \mathbf{u}) + \ell_2(\boldsymbol{\gamma}, \mathbf{u}) + \ell_3(\mathbf{u})$ with respect to \mathbf{u} for current value of $\boldsymbol{\beta}, \boldsymbol{\gamma}$. The converged estimates that maximize the penalized log likelihood (5) are denoted as $\hat{\boldsymbol{\beta}}(\boldsymbol{\sigma}), \hat{\boldsymbol{\gamma}}(\boldsymbol{\sigma})$ and $\hat{\mathbf{u}}(\boldsymbol{\sigma})$. Given $\hat{\boldsymbol{\gamma}}(\boldsymbol{\sigma})$ and $\hat{\mathbf{u}}(\boldsymbol{\sigma})$, The Nelson-Aalen estimation for the baseline cumulative hazard function $\Lambda_0(t)$ is a step function with jumps at observed failure time t_{ij} as in equation (3).

When estimating variance components $\boldsymbol{\sigma}$, both terms ℓ_1 and ℓ_2 in (4) are ignored since they do not depend on $\boldsymbol{\sigma}$ and can be treated as constants. Conditional on $\hat{\boldsymbol{\beta}}(\boldsymbol{\sigma}), \hat{\boldsymbol{\gamma}}(\boldsymbol{\sigma})$ and $\hat{\mathbf{u}}(\boldsymbol{\sigma})$ obtained from previous steps, the approximate profile log likelihood function for $\boldsymbol{\sigma}$ can be expressed as:

$$\ell_p(\boldsymbol{\sigma}) = -\frac{m}{2} \log |\Sigma| - \frac{1}{2} \log(|K''(\mathbf{u})|) - \frac{1}{2} \sum_{i=1}^m \mathbf{u}'_i \Sigma^{-1} \mathbf{u}_i. \quad (6)$$

The estimate of variance component $\boldsymbol{\sigma}$ that maximize the profile log likelihood function (6) has a closed-form expression⁸ as:

$$\hat{\Sigma} = \frac{1}{m} \sum_{i=1}^m \left\{ \mathbf{u}_i \mathbf{u}'_i + K''_i(\mathbf{u})^{-1} \right\}. \quad (7)$$

The overall estimation procedure is summarized as follows:

1. Generate initial values of the random effects \mathbf{u}_i using variance component parameters $\sigma_{11} = \sigma_{22} = 1$ and $\sigma_{12} = 0$.
2. Repeat steps 2a) and 2b) below until $\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{u}$ converge.
 - 2a) By treating \mathbf{u}_i as an offset term for all subjects in the i^{th} cluster, the regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are obtained by maximizing the log likelihood function $\ell_1(\boldsymbol{\beta}, \mathbf{u})$ and $\ell_2(\boldsymbol{\gamma}, \mathbf{u})$, respectively.
 - 2b) For fixed values of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, the Newton-Raphson method is used to maximize $-K(u)$ in (4) by a function of \mathbf{u} .
3. Calculate the estimate of variance parameters $\boldsymbol{\sigma}$ based on (7), which maximizes the profile likelihood function $\ell_p(\boldsymbol{\sigma})$.

The maximum likelihood estimates for all parameters can then be obtained by iterating from step 2 to step 3 until convergence. The convergence criteria is defined by

$$\left\| \boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)} \right\|_{L_1} < \epsilon,$$

where $\boldsymbol{\theta}^{(k)}$ and $\boldsymbol{\theta}^{(k+1)}$ are the parameters from the k^{th} and $(k+1)^{st}$ iteration, and the difference is calculated with L_1 norm. Here ϵ is a small positive number, for example, we set $\epsilon = 10^{-3}$ and 10^{-5} in numerical simulation and application, respectively. The choice of likelihood function $\ell_2(\boldsymbol{\gamma}, \mathbf{u})$ leads to Cox's partial likelihood estimator, where the asymptotic distribution of $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ is well established.¹² Therefore, the number of nuisance parameters $\boldsymbol{\lambda}$ will not introduce convergence issues and the algorithm converges within 5 to 10 iterations in the numerical simulation and real data application. The maximum likelihood estimate of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma})$ is denoted by $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\sigma}})$.

3 Variance Estimation

Standard errors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be estimated using information matrices obtained from the log likelihood functions $\ell_1(\boldsymbol{\beta})$ and $\ell_2(\boldsymbol{\gamma})$. Estimation of standard errors for the variance components $\boldsymbol{\sigma}$ is performed by inverting the Fisher information matrix of the profile likelihood function (6), which takes the form

$$\ell_p(\boldsymbol{\sigma}) = -\frac{1}{2} \sum_{i=1}^m \left\{ \log(|\Sigma| a_{1i} a_{2i} + a_{1i} \sigma_{11} + a_{2i} \sigma_{22} + 1) + \mathbf{u}'_i \Sigma^{-1} \mathbf{u}_i \right\}.$$

We provide the expression for the second order derivative $I(\boldsymbol{\sigma}) = \frac{\partial^2 \ell_p(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}'}$ in the Appendix. The asymptotic variance of $\hat{\boldsymbol{\sigma}}$ is estimated by the inverted information matrix $\text{var}(\hat{\boldsymbol{\sigma}}) = I(\boldsymbol{\sigma})^{-1}$. In the numerical simulations we show that the standard error estimation obtained from the above method tends to underestimate the true standard error. This is because part of the information is missing when using the multivariate penalized likelihood (5) to approximate the likelihood function $\ell(\boldsymbol{\theta})$ in (2).

Ripatti⁸ suggested that the bootstrap resampling method¹⁸ can be used to estimate the standard errors of $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\sigma}}$. We tried the simple random sampling with replacement method for the observed individuals, however, this method takes a large number of iterations to converge (up to 300 or more iterations in some scenarios, compared to 5 to 10 iterations in the original data). Numerical simulation also shows that the simple bootstrap standard errors overestimate the empirical standard errors, especially for the variance component parameters. One possible reason for this may be due to the fact that simple random sampling with replace generates a large amount of duplicated samples and destroys the clustered structure of the data.

Alternatively, we consider the method of *delete- n_i Jackknife algorithm* with possible unequal cluster sizes n_i proposed by Busing et al.¹⁹ Here, n_i represents the number of patients in the i^{th} cluster, $i = 1, 2, \dots, m$. To calculate standard errors of parameter estimates, let $w_i = \frac{h_i}{n}$ be a weight indicating the percentage of subjects in cluster i . We define the pseudo value for parameter $\boldsymbol{\theta}$ when n_i subjects from cluster i is deleted by

$$\tilde{\boldsymbol{\theta}}_i = \frac{1}{w_i} \hat{\boldsymbol{\theta}} + \left(1 - \frac{1}{w_i}\right) \hat{\boldsymbol{\theta}}_{(-i)},$$

where $\hat{\boldsymbol{\theta}}_{(-i)}$ is the multivariate penalized likelihood estimate for the data with n_i subjects from cluster i removed. The *delete- n_i Jackknife* estimate for $\boldsymbol{\theta}$ is given by

$$\bar{\boldsymbol{\theta}} = \sum_{i=1}^m w_i \tilde{\boldsymbol{\theta}}_i = m \hat{\boldsymbol{\theta}} - \sum_{i=1}^m (1 - w_i) \hat{\boldsymbol{\theta}}_{(-i)},$$

and the corresponding variance matrix is estimated by

$$\hat{V} = \frac{1}{m} \sum_{i=1}^m \frac{w_i}{1 - w_i} (\tilde{\boldsymbol{\theta}}_i - \bar{\boldsymbol{\theta}})' (\tilde{\boldsymbol{\theta}}_i - \bar{\boldsymbol{\theta}}).$$

We make the algorithm available in a R package “*bhm*”, where a very simple syntax can be used to fit a joint model by calling the function “*mpl*” in the package (<https://cran.r-project.org/package=bhm>). For example, the following R code

```
fit=mpl(Surv(time,event)~arm+resp+age, resp~arm+age, ~site, jackknife=TRUE)
```

will fit a joint model for a data set with survival outcomes ($time, event$), binary response $resp$, covariates: arm , and age and cluster variable $site$. A Jackknife variance estimate will be provided with the option $jackknife = TRUE$ in the function.

While other complicated bootstrap methods that maintain the clustered structure of the data (i.e. residuals bootstrapping) may work for this kind of data, and remain important for consideration in future work, the proposed Jackknife method also provides a great solution to handle this problem. We show in Sections 4 and 5 that the proposed Jackknife method works well in both numerical simulations and real data applications.

4 Simulation Study

In this section, we conduct simulation studies to evaluate the finite sample properties of the proposed multivariate penalized likelihood method. Data sets will be simulated based on the model (1). For simplicity, we simulated data sets with only one treatment variable (denoted as z_{ij}) for the binary response variable in the generalized linear mixed model. The treatment covariate z_{ij} is generated from a Bernoulli distribution with a success probability of 0.5. We also study how the proposed method performs with an asymmetric binary covariate (i.e., the success probability = 0.25). For the survival outcome (x_{ij}, δ_{ij}), besides treatment variable z_{ij} and binary response variable y_{ij} , we also include the interaction term of these two variables in the Cox frailty model to explore the interactive effect of treatment and response on patients' survival distribution.

By letting $\mathbf{w}_{ij} = (z_{ij}, y_{ij}, z_{ij}y_{ij})$, we have

$$\begin{aligned} \text{logit}(\pi_{ij}|z_{ij}, u_{1i}) &= \beta_0 + z_{ij}\beta_1 + u_{1i}, \\ \log \left\{ \frac{\lambda_{ij}(t|\mathbf{w}_{ij}, u_{2i})}{\lambda_0(t)} \right\} &= z_{ij}\gamma_1 + y_{ij}\gamma_2 + z_{ij}y_{ij}\gamma_3 + u_{2i}. \end{aligned} \quad (8)$$

The cluster-specific random effects u_{1i} and u_{2i} are generated from a bivariate normal distribution $N_2(0, \Sigma)$, where Σ is a 2×2 variance-covariance matrix. Conditional on u_{1i} , the binary response variable Y_{ij} is generated from the first equation of (8). We assume that failure time T_{ij} follows an exponential distribution and is simulated from the second equation of (8), where the baseline hazard $\lambda_0(t)$ is set to a constant of 0.15. Censoring time C_{ij} is assumed to be non-informative and is generated from a uniform distribution on $(0, 20)$. This makes the censoring rate for failure time around 20%.

The simulated data with different parameter specifications are generated to evaluate the robustness of the proposed multivariate penalized likelihood method for joint modeling of binary and survival outcomes. Regression coefficients β_0 , γ_1 , γ_2 and γ_3 are fixed with values of $\beta_0 = -1.0$ and $\gamma_1 = \gamma_2 = \gamma_3 = \log(2.0)$, respectively. The treatment effect for binary response β_1 takes the value of $\log(0.5)$ or $\log(2.0)$. For the variance components σ_{11}, σ_{22} , we consider two different magnitudes for random effect variances: $\sigma_{11} = \sigma_{22} = 0.5$ and $\sigma_{11} = \sigma_{22} = 1.0$. We choose the values of σ_{12} such that the correlation coefficients between u_{1i} and u_{2i} are given by $\rho = -0.9, 0.9$ and 0.0 , respectively.

Two models are applied to the simulated data for comparison: (1) the proposed joint model incorporated with multivariate penalized likelihood method; (2) the separated models consisting of a generalized linear mixed model and a Cox frailty model, implemented with *R* software²⁰ functions *glmer* from package *lme4*, and *coxme* from package *coxme*, respectively. The total sample size n is set to 600, and number of centres m is set to 30. A total of $R = 500$ replications are performed for each parameter combination.

Table 1 shows that the empirical biases of estimates from the proposed multivariate penalized likelihood method and the separated models are both fairly small. This suggests that the proposed joint model provides a consistent estimate for both the regression coefficients and the variance parameters. When $\sigma_{12} \neq 0$, numerical results show that, compared to the separate models, the joint model produces smaller empirical bias. When $\sigma_{12} = 0$, the separate models don't have to fit the extra parameter σ_{12} in the model, therefore, they produce smaller empirical bias in some settings.

The asymptotic standard error (*ASE*) of a parameter estimate is based on the Fisher information matrix, presented as the average asymptotic standard error of 500 simulations. The empirical standard error (*ESE*) of a parameter estimate is calculated by taking the standard deviation of the parameter estimates in 500 simulations. In general, *ESE* can be considered as an approximation of the true standard error.²¹ Results from several sets of parameter specifications are presented in Table 2. For example, results for simulation with $\sigma_{11} = \sigma_{22} = 0.5, \sigma_{12} = -0.45$ and 0.45 (i.e., correlation coefficient $|\rho|=0.9$) are presented in the upper and middle panel of Table 2. We find that the value of *ESE* for joint model is a bit smaller than the separate models, which suggests that the proposed joint model is more efficient. On the other hand, the main concern with the proposed joint model is that in some cases the *ASE* underestimates the true standard error, especially for the variance parameter σ .

There is no obvious difference in the empirical coverage probabilities of the 95% confi-

dence intervals (CIs) between the joint and the separated models. Coverage probabilities for regression coefficients from both models are all close to the nominal level of 95%. Since the distribution of the variance component estimate is skewed, we applied a logarithm transformation to calculate the coverage probabilities of the 95% CIs for σ_{11} and σ_{22} . In the joint model, however, the coverage probability is relatively low (around 70% to 80%) for the variance component σ . This may be due to the fact that the asymptotic standard error from MPL algorithm tends to underestimate the true standard error. We will show later that this issue can be addressed by the Jackknife resampling method. The separate model does not have coverage probability for the variance component σ because it does not provide an estimate of the asymptotic variance for σ .

The efficiency of estimating a parameter is defined by the ratio of its mean square errors (*MSE*). Here we use MSE_J/MSE_S , the ratio of MSE_J (the mean square error of the joint model method), and MSE_S (the mean square error of the separate models method) to compare these two methods. *MSE* is calculated by the formula $MSE = Bias^2 + Variance$, where variance is the square of *ESE*. We did not use the *ASE* because the asymptotic method may underestimate the true standard error. When $\rho = -0.45$ and $\rho = 0.45$, we observed that $MSE_J/MSE_S < 1$. This means that the joint model yields smaller *MSE* than the separate models method. In some scenarios, *MSE* is reduced by up to 20% when the joint model is applied. We also explored many different settings of different parameter values in the co-author's thesis work and similar results were observed.²²

When the covariance σ_{12} takes the value 0, there is no association between the binary marker response and survival data through joint distribution of random effects. Therefore, we expect that the joint model should have a similar performance to the separate models with random effects. The lower panel in Table 2 shows that the joint model and separate models give similar point estimates and ESE for most parameters. There is no obvious evidence showing one model outperforming the other when the covariance σ_{12} is set to 0, which is consistent with our expectations. In general, the advantage of the proposed joint model over separate models is more pronounced as the association between two random effects becomes stronger.

We also conduct numerical simulations to evaluate the performance of the Jackknife standard error in Section 3. Due to the large scale of computation required for the Jackknife resampling, we use sample size $n = 200$ and number of clusters $m = 20$ in this part of the simulation study. The treatment covariate z_{ij} was generated from a Bernoulli distribution with a success probability of 0.25 to study the robustness of the proposed method with an

unbalanced binary covariate.

In Table 3, we compare three types of standard errors: (1) the average over 500 replications of the Jackknife standard error (JSE); (2) the average over 500 replications of the ASE and (3) the ESE . As we can see from the table, the JSE values for all parameters are closer to the gold standard ESE , with the relative differences about -4% to 4% in most settings. On the other hand, the relative differences between ASE and ESE are very large, with many of them in the range of -10% to -30% , which means the asymptotic method under estimates the true standard error. In particular, the estimation of standard errors for variance components, which is especially problematic for the asymptotic method (under estimated by 23.6% on average), is greatly improved using the Jackknife method (the average JSE under estimate the ESE by about 3.1%).

Furthermore, the coverage probabilities of the 95% CIs for regression coefficients β and γ based on the Jackknife method also perform better than the asymptotic method. For the variance parameter σ , the coverage probabilities change significantly and all of them are now close to the nominal level of 95% . For example, the coverage probabilities for parameter σ_{11} and σ_{22} were improved from about 70% - 80% to about 90% - 97% . Thus, we conclude that Jackknife resampling is able to provide decent estimates of the true standard error, and serve as a useful tool in our study to correct the underestimation of standard error in the MPL algorithm. While we expect the estimate for standard error will be improved using the Jackknife method for the separate models, existing R packages do not provide an option to do so.

To explore the association between sample size and performance of the joint model, we simulate data with a series of sample sizes, ranging from $n = 100$ to 600 . The number of clusters m is set to 20 as a constant. As shown in Figure 1, both the bias and MSE for parameter γ_1 decrease consistently as the sample size increases, and similar performances were observed for β_1 , γ_2 and γ_3 (data do not shown). This indicates that increasing sample size improves the accuracy of parameter estimation as well as the finite sample performance of the proposed joint model. If we hold the number of clusters constant, when the total sample size increases, the number of observations in each cluster also increases, which improves the accuracy of Laplace approximation. This agrees with the observations by Ripatti⁸ and Abrahantes.²³

5 Application to a cancer symptom control trial

In this section, the proposed joint model based on the MPL inference method is applied to clinical trial data from a CCTG symptom control study⁷ to explore the potential association between response to pain control radiation and overall survival, and to detect important predictors for patients’ survival.

A total of 850 patients were randomized to received either single fraction or multiple fraction radiation therapy (RT) with 425 patients in each arm. The original trial shows that a single fraction RT is non-inferior to multiple fraction RT in term of the pain control response at 2 months.⁷ In this exploratory analysis, we consider the association between pain response at 2 months (Yes vs. No, denoted by *resp*) and overall survival, and their relationship with the following two categorical and continuous variables: (1) treatment effect (denoted by *arm*: $arm = 0$ for multiple fraction and $arm = 1$ for single fraction), and (2) age of patient (denoted by *age*, as a continuous variable in years).

To remove the guarantee-time bias when assessing the association between pain response and survival, we implement a landmark analysis here.²⁴ The pain response evaluation time is selected as the “natural landmark time”, which is two months after randomization. Thus, among $n = 850$ patients, 141 with survival time less than 2 months are removed from the data set. Among 714 patients included in this study, 358 patients were in the multiple fraction arm and 356 patients were in the single fraction arm. The median age was 65 years (range 18.4 to 93.0). The overall pain control response rate at two months was 36.4%. The survival time used in data analysis is calculated by subtracting the time of response evaluation (two months) from the original survival time. A total of 350 patients out of 714 died at the end of the trial, with a median survival duration 8.8 months (95% CI 8.2 to 10.1 months) from the two months landmark time.

The clinical trial data contain cancer patients with different type of cancers and each with a potentially different painful bone lesion site. In the analysis, it is a challenge to take both factors into account (e.g. breast cancer with painful limbs site, lung cancer with painful superficial bones, etc, with a total of more than 60 different combinations) using a fixed effects model. Therefore, the proposed joint model is fit to this symptom control trial data to account for the clustered effect of cancer type and painful bone lesion. The multivariate penalized likelihood method is applied to calculate the point estimates for regression coefficients and random effects variance components. We further apply the Jackknife method to estimate the standard errors for the model parameters. The parameter estimates (*EST*) based on the joint model, *ASE*, *JSE*, Odds Ratios (OR), Hazards Ratios (HR), the corresponding 95%

CI and P-values calculated through the asymptotic method and the Jackknife method are summarized in Table 4.

Table 4 shows that for the regression parameters, the *ASEs* calculated directly by the asymptotic method are slightly smaller than those from the Jackknife method. For the variance components σ , Jackknife standard errors are much larger. For example, standard error for $\hat{\sigma}_{12} = -0.167$ is 0.043 with the asymptotic method and 0.062 with the Jackknife method, respectively. This makes the 95% CIs using the Jackknife method wider than those based on the asymptotic method. Since the Jackknife method generally produces more accurate estimates of standard errors, we recommend using this method instead of the asymptotic method to interpret the analysis results.

As shown in Table 4, covariate *arm* is not significantly associated with neither the pain control response outcome (OR = 1.22, 95% CI: 0.77 to 1.93) nor the overall survival outcome (HR = 1.06, 95% CI: 0.77 to 1.48), which is expected because of the non-inferiority trial design. The mortality risk is significantly lower for patients who responded to the pain control treatment, with a HR of 0.53 (95% CI: 0.37 to 0.76), after accounting for the treatment and patient’s age. Age is significantly associated with the pain response outcome (OR = 1.02, 95% CI 1.00 to 1.04), but not the overall survival outcome (HR = 1.01, 95% CI: 0.99 to 1.02). The covariance between the random effect for pain response and overall survival is estimated to be $\hat{\sigma}_{12} = -0.167$ (95% CI: -0.29 to -0.05, p-value 0.0068). This shows a negative correlation between these two random effect variables, with the corresponding correlation coefficient $\rho = -0.77$.

It should be noted that even if the patients who responded to the pain control treatment had longer survival time, we cannot conclude that there is a causal relationship between the pain control response and survival. As pointed out by Anderson,²⁵ response to the treatment could simply be a marker of patients with favorable prognostic characteristics, which could be the true reason for longer survival.

6 Discussion

In this paper, we propose a joint model for binary marker response and survival data, and develop a joint inference method based on a multivariate penalized likelihood (MPL) method (“*mpl*” function in the “*bhm*” package: <https://cran.r-project.org/package=bhm>). The performance of the proposed joint model and the MPL method is evaluated in extensive simulation studies. We also applied the proposed method to an actual data set from the CCTG’s symptom control study. The symptom control data that support the findings of this study are

available upon request from the corresponding author. The data are not publicly available due to privacy restrictions.

The MPL inference technique used in this paper circumvents the intractable integration in maximum likelihood analysis by utilizing Laplace approximation. It then maximizes the approximate likelihood function with respect to fixed-effect parameters, random effects (treated as parameters) and variance components in a sequential manner until all estimates converge.

In simulation studies, we observed that the asymptotic standard errors were underestimated for the MPL method. Thus, we implemented the Jackknife method, and the simulation results indicate that it can greatly improve the estimation of standard errors, especially for the frailty variance components. Therefore, the Jackknife method can be used alone with the MPL algorithm when reasonably accurate standard error estimates are required.

Performance of the proposed joint model is compared with the separate models for binary and survival outcomes in simulation studies. When the covariance of random effects in the simulated data is set to zero to mimic the situation of no association between two sub-models, the overall performance of joint models and separate models are very similar. However, when there is a positive or negative association between the binary and survival outcomes, the joint model shows some favorable features over separate models. The joint model reduces the bias in the estimator of fixed-effect parameters and variance components, decreases the MSE and increases the efficiency of parameter estimation. It also produces a slightly smaller ESE , resulting in more precise estimation and higher power. Finally, it provides a decent estimation for the frailty covariance, which is not available in the analysis of separate models.

It has been shown previously that the inference method based on penalized likelihood was much faster in computation than EM algorithm and the performance of these two methods was comparable (Ye et al. 2008).¹⁶ It is of interest to compare the proposed MPL method with more exact likelihood approaches such as those based on the Monte Carlo EM algorithm by Hwang and others.⁶ Tables 1 and 2 of Hwang's paper have similar censoring rate and sample size to our Table 3, with slightly differences in parameter settings. For most of the parameters, both the penalized likelihood method and the Monte Carlo EM algorithm method provide a relative bias between -10% and 10%, with the largest absolute value of relative bias around 16% to 18%. The coverage probabilities of the 95% CIs are also very similar for both methods (most of the coverage probabilities from both methods are around 90% to 97%). Overall, the finite sample properties of the MPL method and the Monte Carlo EM algorithm method are very similar.

In summary, the proposed joint model with MPL provides a computationally efficient

and reasonably advantageous approach to jointly analyze binary marker response and survival data. It provides a better fit than the separate models, particularly when there is a strong association between the binary response and survival. In addition, the proposed joint model is capable of detecting the association between two endpoints at both fixed-effect and random-effect levels. However, the poor estimation of asymptotic standard error with the MPL algorithm could be problematic. We adapted the Jackknife resampling method to correct the problem.

Acknowledgments

The authors wish to thank both the associate editor and the reviewers for their insightful comments and suggestions. This work was supported in part by grant from the Natural Sciences and Engineering Research Council of Canada (NSERC). The computation was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca) and Compute/Calcul Canada.

Appendix

The first and second order of partial derivatives for $K(\mathbf{u})$ can be written as below. The i^{th} component of the first order partial derivative is given by,

$$\frac{\partial K(\mathbf{u})}{\partial u_{1i}} = - \sum_{j=1}^{n_i} \left\{ Y_{ij} - \frac{\exp(W'_{ij}\boldsymbol{\beta} + u_{1i})}{1 + \exp(W'_{ij}\boldsymbol{\beta} + u_{1i})} \right\} + \frac{u_{1i}\sigma_{22} - u_{2i}\sigma_{12}}{|\Sigma|}$$

and

$$\frac{\partial K(\mathbf{u})}{\partial u_{2i}} = - \sum_{j=1}^{n_i} \left\{ \delta_{ij} - \Lambda_0(X_{ij}) \exp(Z'_{ij}\boldsymbol{\gamma} + u_{2i}) \right\} + \frac{u_{2i}\sigma_{11} - u_{1i}\sigma_{12}}{|\Sigma|}$$

for $i = 1, \dots, m$. Here $|\Sigma| = (\sigma_{11}\sigma_{22} - \sigma_{12}^2)$. Define $S_i = \sum_{j=1}^{n_i} (Y_{ij} - \pi_{ij}, \delta_{ij} - \Lambda_{ij})'$, we have

$$\frac{\partial K(\mathbf{u})}{\partial \mathbf{u}_i} = -S_i + \Sigma^{-1}\mathbf{u}_i$$

To obtain the the second order derivatives of $K(\mathbf{u})$, let

$$a_{1i} = \sum_{j=1}^{n_i} \pi_{ij}(1 - \pi_{ij}) \text{ and } a_{2i} = \sum_{j=1}^{n_i} \Lambda_{ij}.$$

We have

$$\begin{aligned} \frac{\partial^2 K(\mathbf{u})}{\partial u_{1i}^2} &= a_{1i} + |\Sigma|^{-1}\sigma_{22}, \\ \frac{\partial^2 K(\mathbf{u})}{\partial u_{2i}^2} &= a_{2i} + |\Sigma|^{-1}\sigma_{11}, \\ \frac{\partial^2 K(\mathbf{u})}{\partial u_{1i}\partial u_{2i}} &= -|\Sigma|^{-1}\sigma_{12}, \end{aligned}$$

and

$$\frac{\partial^2 K(\mathbf{u})}{\partial u_{1i}\partial u_{1j}} = \frac{\partial^2 K(\mathbf{u})}{\partial u_{2i}\partial u_{2j}} = \frac{\partial^2 K(\mathbf{u})}{\partial u_{1i}\partial u_{2j}} = 0 \text{ for } i \neq j.$$

Define $A_i = \text{diag}(a_{1i}, a_{2i})$, we have

$$K''_i(\mathbf{u}) = \frac{\partial^2 K(\mathbf{u})}{\partial \mathbf{u}_i \partial \mathbf{u}'_i} = \begin{bmatrix} a_{1i} + |\Sigma|^{-1}\sigma_{22}, & -|\Sigma|^{-1}\sigma_{12} \\ -|\Sigma|^{-1}\sigma_{12}, & a_{2i} + |\Sigma|^{-1}\sigma_{11} \end{bmatrix} = A_i + \Sigma^{-1},$$

and

$$|K''(\mathbf{u})| = \prod_{i=1}^m |K''_i(\mathbf{u})| = \prod_{i=1}^m \left\{ a_{1i}a_{2i} + |\Sigma|^{-1}(a_{1i}\sigma_{11} + a_{2i}\sigma_{22} + 1) \right\}.$$

To obtain the second order derivative of $\ell_p(\boldsymbol{\sigma})$, we define the following notation:

$$\begin{aligned}
\kappa_i &= |\Sigma| |K_i''(\mathbf{u})| = |\Sigma| a_{i1} a_{i2} + a_{i1} \sigma_{11} + a_{i2} \sigma_{22} + 1, \\
\psi_{1i} &= a_{i1} \sigma_{11}^2 + a_{i2} \sigma_{12}^2 - \sigma_{11} \\
\psi_{2i} &= a_{i1} \sigma_{12}^2 + a_{i2} \sigma_{22}^2 - \sigma_{22}, \\
\varphi_i &= a_{i1} \sigma_{11} + a_{i2} \sigma_{22} - 1, \\
\eta &= 2\sigma_{11} \sigma_{22} - \sigma_{12}^2 = \sigma_{11} \sigma_{22} + |\Sigma|.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\frac{\partial^2 \ell_p(\boldsymbol{\sigma})}{\partial \sigma_{11}^2} &= \sum_{i=1}^m \frac{\psi_{2i}(2a_{i1} a_{i2} \sigma_{22} |\Sigma| - a_{1i} \eta - a_{i2} \sigma_{22}^2 + \sigma_{22})}{2|\Sigma|^2 \kappa_i^2} \\
&\quad - \sum_{i=1}^m \frac{(u_{1i} \sigma_{22} - u_{2i} \sigma_{12})^2 \sigma_{22}}{|\Sigma|^3} + m \frac{\sigma_{22}^2}{2|\Sigma|^2} \\
\frac{\partial^2 \ell_p(\boldsymbol{\sigma})}{\partial \sigma_{22}^2} &= \sum_{i=1}^m \frac{\psi_{1i}(2a_{i1} a_{i2} \sigma_{11} |\Sigma| - a_{i1} \sigma_{11}^2 - a_{2i} \eta + \sigma_{11})}{2|\Sigma|^2 \kappa_i^2} \\
&\quad - \sum_{i=1}^m \frac{(u_{1i} \sigma_{12} - u_{2i} \sigma_{11})^2 \sigma_{11}}{|\Sigma|^3} + m \frac{\sigma_{11}^2}{2|\Sigma|^2} \\
\frac{\partial^2 \ell_p(\boldsymbol{\sigma})}{\partial \sigma_{12}^2} &= \sum_{i=1}^m \frac{\varphi_i (|\Sigma| \kappa_i + 2\sigma_{12}^2 (2a_{i1} a_{2i} |\Sigma| - \varphi_i))}{2|\Sigma|^2 \kappa_i^2} \\
&\quad - \sum_{i=1}^m \frac{(\mathbf{u}'_i \Sigma^{-1} \mathbf{u}_i) \sigma_{12} (|\Sigma| - 4\sigma_{12}^2) + 4u_{1i} u_{2i}}{|\Sigma|^2} + m \frac{\sigma_{11} \sigma_{22} + \sigma_{12}^2}{|\Sigma|^2} \\
\frac{\partial^2 \ell_p(\boldsymbol{\sigma})}{\partial \sigma_{11} \partial \sigma_{22}} &= \sum_{i=1}^m \frac{\psi_{2i}(2a_{i1} a_{i2} |\Sigma| \sigma_{11} - a_{1i} \sigma_{11}^2 - a_{2i} \eta + \sigma_{11}) - (2a_{2i} \sigma_{22} - 1) |\Sigma| \kappa_i}{2|\Sigma|^2 \kappa_i^2} \\
&\quad - \sum_{i=1}^m \frac{(u_{1i} \sigma_{22} - u_{2i} \sigma_{12})^2 \sigma_{11} - u_{1i} (u_{1i} \sigma_{22} - u_{2i} \sigma_{12}) |\Sigma|}{|\Sigma|^3} + m \frac{\sigma_{12}^2}{2|\Sigma|^2} \\
\frac{\partial^2 \ell_p(\boldsymbol{\sigma})}{\partial \sigma_{11} \partial \sigma_{12}} &= \sum_{i=1}^m \frac{\psi_{2i} \sigma_{12} (2a_{i1} a_{i2} |\Sigma| + \varphi_i) - a_{1i} \sigma_{12} |\Sigma| \kappa_i}{|\Sigma|^2 \kappa_i^2} \\
&\quad - \sum_{i=1}^m \frac{2(u_{1i} \sigma_{22} - u_{2i} \sigma_{12})^2 \sigma_{12} - u_{2i} (u_{1i} \sigma_{22} - u_{2i} \sigma_{12}) |\Sigma|}{|\Sigma|^3} - m \frac{\sigma_{22} \sigma_{12}}{2|\Sigma|^2} \\
\frac{\partial^2 \ell_p(\boldsymbol{\sigma})}{\partial \sigma_{22} \partial \sigma_{12}} &= \sum_{i=1}^m \frac{\psi_{1i} \sigma_{12} (2a_{i1} a_{i2} |\Sigma| + \varphi_i) - a_{2i} \sigma_{12} |\Sigma| \kappa_i}{|\Sigma|^2 \kappa_i^2} \\
&\quad - \sum_{i=1}^m \frac{2(u_{1i} \sigma_{12} - u_{2i} \sigma_{11})^2 \sigma_{12} - u_{1i} (u_{2i} \sigma_{11} - u_{1i} \sigma_{12}) |\Sigma|}{|\Sigma|^3} - m \frac{\sigma_{11} \sigma_{12}}{2|\Sigma|^2}.
\end{aligned}$$

References

1. Glidden DV, Vittinghoff E. Modelling clustered survival data from multicentre clinical trials. *Statistics in medicine*. 2004;23(3):369–388.
2. Neuhaus A, Augustin T, Heumann C, Daumer M. A review on joint models in biometrical research. *Journal of Statistical Theory and Practice*. 2009;3(4).
3. Wu L, Liu W, Yi GY, Huang Y. Analysis of longitudinal and survival data: Joint modeling, inference methods, and issues. *Journal of Probability and Statistics*. 2012;12:1-17.
4. Inoue LYT, Thall PF, Berry DA. Seamlessly expanding a randomized phase II trial to phase III.. *Biometrics*. 2002;58:823–831.
5. Lai T, Lavori P, Shih M. Sequential design of phase II-III cancer trials.. *Statistics in Medicine*. 2012;31:1944-1960.
6. Hwang YT, Huang CH, Wang CC, Y. LT, Tseng YK. Joint modelling of longitudinal binary data and survival data. *Journal of Applied Statistics*. 2019;Published Online:1-15.
7. Chow E, Linden YM, Roos D, *et al.* A Randomized Comparison of Single and Multiple Fractions of Repeat Radiation for Painful Bone Metastases. *Lancet Oncology*. 2014;15:164-171.
8. Ripatti S, Palmgren J. Estimation of multivariate frailty models using penalized partial likelihood.. *Biometrics*. 2000;56(4):1016–1022.
9. Breslow NE, Clayton DG. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*. 1993;88:9-25.
10. Rizopoulos D, Verbeke G, Lesaffre E. Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2009;71(3):637–654.
11. Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*. 1972;34:187-220.
12. Andersen PK, Borgan O, Gill RD, Keiding N. *Statistical Models Based on Counting Processes*. Springer-Verlag, New York 1993.

13. Murphy S, Van Der Laan M, Robins J. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*. 2001;96(456):1410–1423.
14. Nelson W. Theory and applications of hazard plotting for censored failure time data. *Technometrics*. 1972;14:945–965.
15. Aalen O. Nonparametric inference for a family of counting processes. *Annals of Statistics*. 1978;6:701–726.
16. Ye W, Lin X, Taylor J. A penalized likelihood approach to joint modeling of longitudinal measurements and time-to-event data.. *Statistics and Its Interface*. 2008;1:33-45.
17. Cox DR. Partial likelihood. *Biometrika*. 1975;62:269-276.
18. Efron B. *The jackknife, the bootstrap, and other resampling plans*. Society of Industrial and Applied Mathematics CBMS-NSF Monographs 1982.
19. Busing FM, Meijer E, Van der Leeden R. Delete-m Jackknife for Unequal m.. *Statistics and Computing*. 1999;9:3-8.
20. R Core Team . *R: A language and environment for statistical computing*. R Foundation for Statistical ComputingVienna, Austria 2018.
21. Hanley JA, Negassa A, Edwardes MD, Forrester JE. Statistical analysis of correlated data using generalized estimating equations: an orientation. *American Journal of Epidemiology*. 2003;157:364-375.
22. Wang J. Joint Modeling of Binary Response and Survival Data in Clinical Trials. Master’s thesisQueen’s University 2013.
23. Abrahantes JC, Burzykowski T. A version of the EM algorithm for proportional hazard model with random effects.. *Biometrical Journal*. 2005;47(6):847–862.
24. Giobbie-Hurder A, Gelber RD, Regan MM. Challenges of Guarantee-Time Bias.. *Journal of Clinical Oncology*. 2013;31:2963–2969.
25. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response and other comparisons of time-to-event by outcome variables.. *Journal of Clinical Oncology*. 2008;26:3913-3915.

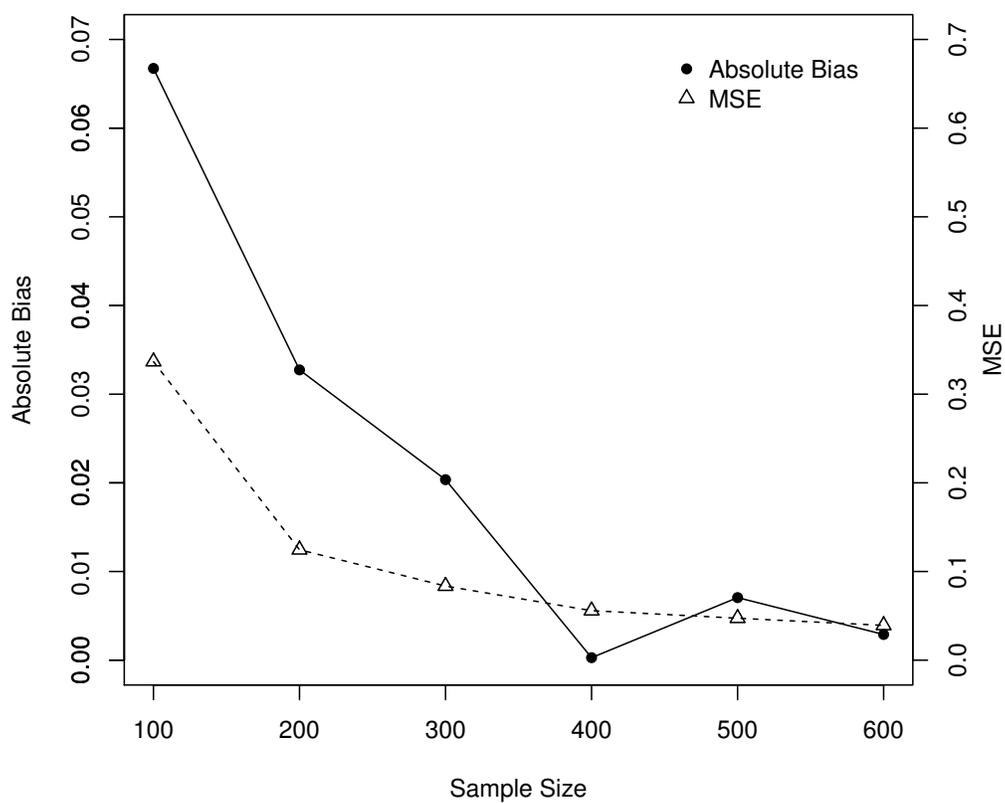


Figure 1: Simulation results for relationship between sample size and *Bias* and mean square error (*MSE*) for regression parameter γ_1 , based on average of 500 replications.

Table 1: Simulation results for comparison of empirical bias for joint model and separate models. Here $\gamma_1 = \gamma_2 = \gamma_3 = \log(2)$. (sample size $n = 600$, number of clusters $m = 30$, results are based on average of 500 replications).

			Joint Model				Separate Model			
σ_{11}	σ_{22}	σ_{12}	$\hat{\beta}_1$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\beta}_1$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$
$\beta_1 = \log(2.0)$										
1.0	1.0	-0.90	-0.027	0.007	0.001	-0.001	0.004	-0.068	0.007	0.008
1.0	1.0	0.90	-0.004	0.016	0.009	-0.015	-0.008	0.061	-0.007	-0.001
0.5	0.5	-0.45	-0.007	0.021	0.008	-0.010	0.017	0.065	-0.018	0.001
0.5	0.5	0.45	-0.008	0.009	0.007	-0.007	0.017	-0.037	0.012	-0.010
1.0	1.0	0.00	0.017	0.004	-0.004	-0.004	-0.003	0.003	0.002	0.010
0.5	0.5	0.00	-0.020	0.013	0.007	-0.007	0.013	-0.012	0.002	0.016
$\beta_1 = \log(0.5)$										
1.0	1.0	-0.90	0.024	0.003	-0.001	-0.008	-0.011	-0.059	0.001	0.018
1.0	1.0	0.90	0.012	0.007	-0.005	0.012	0.014	0.042	0.004	0.019
0.5	0.5	-0.45	-0.002	-0.002	0.003	-0.005	-0.004	-0.065	0.001	0.020
0.5	0.5	0.45	0.014	0.009	0.001	0.008	0.017	0.051	0.007	0.024
1.0	1.0	0.00	0.038	0.007	-0.001	0.009	-0.029	0.008	-0.003	0.014
0.5	0.5	0.00	0.029	0.012	0.006	-0.010	0.023	0.006	-0.008	0.024
Variance components										
σ_{11}	σ_{22}	σ_{12}	$\hat{\sigma}_{11}$	$\hat{\sigma}_{22}$	$\hat{\sigma}_{12}$			$\hat{\sigma}_{11}$	$\hat{\sigma}_{22}$	
$\beta_1 = \log(2.0)$										
1.0	1.0	-0.90	-0.090	-0.037	0.054			-0.026	-0.001	
1.0	1.0	0.90	-0.057	-0.017	-0.032			-0.020	-0.002	
0.5	0.5	-0.45	-0.016	-0.008	-0.015			-0.028	-0.020	
0.5	0.5	-0.45	-0.010	-0.007	0.009			-0.026	0.002	
1.0	1.0	0.00	-0.145	-0.049	-0.002			-0.018	0.009	
0.5	0.5	0.00	-0.049	-0.015	0.001			-0.022	0.005	
$\beta_1 = \log(0.5)$										
1.0	1.0	-0.90	-0.080	-0.044	0.052			-0.014	0.010	
1.0	1.0	0.90	-0.079	-0.047	-0.052			-0.025	0.013	
0.5	0.5	-0.45	-0.031	-0.017	0.026			-0.017	-0.014	
0.5	0.5	-0.45	-0.031	-0.025	-0.032			-0.014	-0.001	
1.0	1.0	0.00	-0.174	-0.053	-0.017			-0.038	0.015	
0.5	0.5	0.00	-0.059	-0.017	0.005			-0.030	0.023	

Table 2: Simulation results of asymptotic standard error (ASE), empirical standard error (ESE), coverage probability (CP) of 95% confidence intervals and model efficiency of joint model and separate model for (Sample size $n = 600$, number of clusters $m = 30$, results are based on average of 500 replications).

		Joint Model			Separate Model			Efficiency [†]
Parameter		ASE	ESE	CP	ASE	ESE	CP	
$\sigma_{12} = -0.45$								
β_1	$\log(2.0)$	0.181	0.181	95.2	0.183	0.184	94.4	0.960
γ_1	$\log(2.0)$	0.151	0.160	94.8	0.162	0.163	94.0	0.918
γ_2	$\log(2.0)$	0.120	0.126	93.6	0.125	0.133	94.2	0.885
γ_3	$\log(2.0)$	0.197	0.201	94.8	0.206	0.213	94.0	0.891
σ_{11}	0.5	0.129	0.195	78.4	-	0.215	-	0.811
σ_{22}	0.5	0.128	0.155	83.8	-	0.158	-	0.971
σ_{12}	-0.45	0.120	0.145	84.4	-	-	-	-
$\sigma_{12} = 0.45$								
β_1	$\log(2.0)$	0.181	0.180	96.0	0.183	0.182	95.2	0.967
γ_1	$\log(2.0)$	0.150	0.160	93.6	0.157	0.158	91.6	0.895
γ_2	$\log(2.0)$	0.125	0.129	94.2	0.130	0.130	95.0	0.973
γ_3	$\log(2.0)$	0.197	0.205	93.6	0.205	0.213	94.8	0.927
σ_{11}	0.5	0.128	0.195	76.8	-	0.201	-	0.928
σ_{22}	0.5	0.128	0.149	86.0	-	0.156	-	0.901
σ_{12}	0.45	0.119	0.152	84.0	-	-	-	-
$\sigma_{12} = 0.0$								
β_1	$\log(2.0)$	0.179	0.178	95.6	0.183	0.181	95.8	0.981
γ_1	$\log(2.0)$	0.150	0.160	94.4	0.159	0.154	96.8	1.090
γ_2	$\log(2.0)$	0.122	0.131	92.6	0.128	0.125	96.0	1.096
γ_3	$\log(2.0)$	0.197	0.203	94.6	0.205	0.196	96.6	1.067
σ_{11}	0.5	0.124	0.185	74.4	-	0.196	-	0.934
σ_{22}	0.5	0.126	0.142	86.0	-	0.160	-	0.796
σ_{12}	0.0	0.087	0.122	86.2	-	-	-	-

[†] Efficiency is define by the mean square error (MSE) of the joint model divided by the MSE of the separated models.

Table 3: Simulation results for empirical bias, empirical standard error (ESE), asymptotic standard error (ASE) and Jackknife standard error (JSE) with $\beta_1 = \gamma_1 = \gamma_2 = \gamma_3 = \log(2.0)$, $\sigma_{11} = \sigma_{22} = 0.5$. (Sample size $n = 200$, number of clusters $m = 20$, results are based on average of 500 replications.)

Parameter	Bias	RelBias [‡]	ESE	Asymptotic Method			Jackknife Method		
				ASE	A/E [†]	CP	JSE	J/E [†]	CP
$\sigma_{12} = -0.45$									
β_1	-0.042	-6.06%	0.376	0.359	0.955	93.4	0.359	0.955	91.2
γ_1	0.014	2.02%	0.260	0.245	0.942	93.8	0.269	1.034	94.8
γ_2	-0.010	-1.44%	0.250	0.215	0.860	93.2	0.247	0.988	94.0
γ_3	0.001	0.14%	0.430	0.358	0.832	91.0	0.435	1.012	92.8
σ_{11}	0.044	8.80%	0.300	0.191	0.637	77.8	0.272	0.907	90.0
σ_{22}	-0.011	-2.20%	0.223	0.158	0.709	78.8	0.216	0.969	93.4
σ_{12}	0.068	15.11%	0.208	0.145	0.697	72.2	0.199	0.957	97.2
$\sigma_{12} = 0.45$									
β_1	-0.059	-8.50%	0.367	0.358	0.975	94.6	0.365	0.994	93.4
γ_1	-0.002	-0.29%	0.293	0.253	0.863	92.8	0.278	0.948	93.6
γ_2	0.060	8.66%	0.247	0.215	0.870	90.2	0.237	0.959	92.8
γ_3	0.019	2.74%	0.429	0.387	0.902	92.8	0.431	1.005	94.0
σ_{11}	0.045	9.00%	0.275	0.191	0.695	80.8	0.279	1.015	91.8
σ_{22}	-0.033	-6.60%	0.216	0.152	0.704	78.4	0.206	0.954	93.0
σ_{12}	-0.074	-16.40%	0.201	0.144	0.716	71.6	0.195	0.970	96.8
$\sigma_{12} = 0.0$									
β_1	-0.029	-4.18%	0.352	0.353	1.002	96.0	0.358	1.017	94.0
γ_1	0.004	0.58%	0.287	0.249	0.868	90.8	0.274	0.955	92.6
γ_2	0.015	2.16%	0.242	0.214	0.884	90.6	0.242	1.000	92.8
γ_3	-0.005	0.72%	0.438	0.384	0.876	93.0	0.430	0.982	94.0
σ_{11}	-0.063	-12.60%	0.246	0.163	0.662	70.2	0.262	1.065	92.8
σ_{22}	-0.038	-7.60%	0.215	0.150	0.698	77.8	0.204	0.948	92.2
σ_{12}	0.001	—	0.180	0.108	0.600	80.6	0.166	0.922	95.2

‡ RelBias: Relative bias (no available for $\hat{\sigma}_{12}$ when $\sigma_{12} = 0$)

† A/E = ASE/ESE and J/E = JSE/ESE

Table 4: Application of the proposed joint model to a cancer symptom control trial

	Asymptotic Method					Jackknife Method		
Binary response outcome								
	Est [†]	OR	ASE	95%CI	p-value	JSE	95%CI	p-value
arm ¹	0.201	1.22	0.451	0.90, 1.67	0.203	0.572	0.77, 1.93	0.390
age ²	0.019	1.02	0.007	1.01, 1.03	0.004	0.009	1.00, 1.04	0.033
Overall survival outcome								
	Est	HR	ASE	95%CI	p-value	JSE	95%CI	p-value
arm	0.061	1.06	0.108	0.86, 1.31	0.568	0.168	0.77, 1.48	0.715
age	0.006	1.01	0.005	0.99, 1.01	0.198	0.007	0.99, 1.02	0.418
resp ³	-0.631	0.53	0.120	0.42, 0.67	<0.0001	0.179	0.37, 0.76	<0.0001
Variance Components								
	Est	ASE	95% CI	p-value	JSE	95% CI	p-value	
σ_{11}	0.142	0.039	0.08, 0.24	<0.0001	0.078	0.05, 0.42	0.0004	
σ_{22}	0.328	0.071	0.21, 0.50	<0.0001	0.125	0.16, 0.69	0.0033	
σ_{12}	-0.167	0.043	-0.25,-0.08	0.0001	0.062	-0.29,-0.05	0.0068	

†: Est for parameter estimate, OR for Odds Ratio and HR for Hazards Ratio

1. arm: Single fraction vs Multiple fraction
2. age: Age in a continuous scale
3. resp: Response to pain treatment