Journal of Medical Statistics and Informatics
ISSN 2053-7662 | Volume 9 | Article 2

HOAJ
Herbert Open Access Journals

Original Research Open Access

Joint Modeling Analysis of Multivariate Skewed-longitudinal and Time-to-event Data with Application to Primary Biliary Cirrhosis Study

Lan Xu[1], Yangxin Huang[1*], Henian Chen[1], Alfred Mbah[1] and Feng Cheng[2]

*Correspondence: yhuang@usf.edu

[1]Department of Epidemiology and Biostatistics, College of Public Health, University of South Florida, Tampa, FL 33612, U.S.A.
[2]Department of Pharmaceutical Science, College of Pharmacy, University of South Florida, Tampa, FL 33612, U.S.A.

Abstract

Background: Many clinical and public health researches collect data including multiple longitudinal measures and time-to-event outcomes, where characteristics of the pattern of exposure change and the association between features of longitudinal biomakers and the primary survival endpoint are of interest.

Methods: Many existing statistical models for longitudinal-survival data might not provide robust inference when more than one longitudinal exposures which were significantly correlated and longitudinal measurements exhibit skewness and/or heavy tails; ignoring these data features may lead to biased estimation. In this article, we offered a multivariate joint model with the skew-normal (SN) distribution with application to the Mayo clinic primary biliary cirrhosis (PBC) study to assess simultaneous effects.

Results: With the multivariate joint modeling associated with the skew-normal (SN) distribution, the subject-specific baseline (HR=2.390 with 95% CI: (1.429, 4.112)) and change rate (HR=2.588 with 95% CI: (1.845, 3.967)) of Bilirubin in natural log scale were positively associated with the risk of death; the higher the subject-specific change rate (HR=0.191 with 95% CI: (0.037, 0.915)) of Albumin in natural log scale was associated with a decrease in mortality rate; the subject-specific of SGOT levels in natural log scale did not affect the risk of death for PBC patients significantly. The results of the skewness parameters of natural log-transformed Bilirubin (δ_1=0.42), Albumin (δ_2=−0.03) and SGOT (δ_3=0.095) were estimated to be significant, indicating the skewness of three biomarkers existed.

Conclusions: Our results revealed the Bilirubin and Albumin levels may be involved in predicting risk of death for PBC patients, except for SGOT. The multivariate joint modeling associated with SN distribution provides better fit to the data, gives less biased parameter estimates for those longitudinal biomarkers in comparison with its counterpart where the normal distribution is assumed (data not shown here). The introduced modeling approach is generally applicable to other situations where longitudinal measurements and time-to-event outcomes are available.

Keywords: Bayesian inference, longitudinal-survival data, multivariate joint model, primary biliary cirrhosis, skew-normal distribution

Introduction

Primary biliary cirrhosis (PBC), recently known as primary biliary cholangitis, is a relatively rare disease caused by inflammatory destruction of the small bile ducts from the liver. Eventually, this pressure build- up will harm the bile ducts leading to liver cell damage and cirrhosis. The cause of PBC is unknown, but because the body's immune system attacks its own cells, it is most likely thought to be an auto immune disease. In this disease, the bile ducts are under attack and are destroyed [16]. Women are more likely than men to have PBC, it is most often in the woman above the age of 40 [26,39].

Mathematical models based on PBC clinical study have been developed to predict disease progression. Cox proportional hazards model for survival analysis was performed to identify the two significant biomarkers, Alkaline Phosphatase and

doi: 10.7243/2053-7662-9-2

Serum Bilirubin, regarding the risk of an event (death or liver trans- plantation) for patients diagnosed with PBC [26,27]. An increase in the levels of these biomarkers was positively associated with the hazards of PBC patients. Typically, serum bilirubin concentration was the best prognostic biomarker from all the other laboratory measurements. When the serum bilirubin exceeded 6.0 (mg/dl), the survival time was around 25 months [40]. Many other risk covariates such as age, sex, ascites, prothrombin were also used to prognostic models for PBC. However, there were several potential limitations of the previous approach-based survival analysis when the covariates were repeatedly measured over time. Firstly, only capture the biomarker observation until a certain time point, mean value or at particular time point was taken into account, but not all observations over time were take into account; using only one observation of the biomarker obviously discarded useful information about the biomarker and its trajectory. Secondly, the inter relationships between longitudinal and time-to-event processes were ignored. Thirdly, longitudinal growth trajectory as a time-varying covariate was not fully considered to assess the effect of longitudinal measures on the risk of the event. The longitudinal processes will affect the hazard of survival; therefore, an appropriate statistical model is needed to capture the unobservable quantities of the growth profile and overall growth trajectory for association with risk of death for PBC patients. The joint modeling had been applied to remedy the deficiencies.

The joint modeling approach can capture the rate of change in the biomarker levels, which contains the differences between patients and also the difference over time for the same patient. Joint modeling of longitudinal and survival data is an active area of scientific fields such as in biology, biomedical and clinical research, since it allows simultaneous analysis of longitudinal (repeated) measurements and time- to-event (survival) outcome [4,17,23,34,36,44,46]. For example, Allen et al. [2] inspected the relationship between five longitudinally collected cytokines (Interleukin (IL)-6, IL-8, growth-related oncogene-1 (GRO- 1), vascular endothelial growth factor (VEGF), and hepatocyte growth factor (HGF)) measured from serum plasma and survival, focused on whether the values of these multiple cytokines were associated with survival. Survival outcome is always associated with multiple longitudinal outcomes. Important features in clinical studies of this type are that there might be a relatively large number of biomarkers [1,5,7,8,14,24,28,32,33,37,42,43], these biomarkers are subject to sizable measurement error due to laboratory error and biological variation and which may be significantly correlated, like the PBC study from the Mayo clinic [31], patients with PBC were followed longitudinally and multiple longitudinal biomarkers were measured. The PBC data collected at the Mayo Clinic between 1974 to 1984 [31] have been widely analyzed using joint modeling methods [1,3,9,11,18,12,13,35]; researchers that risk of death was significantly impacted by the logarithm of serum bilirubin levels. However, multiple longitudinal biomarkers were collected in the PBC study, it was important to understand the relationship among those biomakers' growth over time and the risk of PBC death. The adoption of a multivariate joint modeling (MVJM) to reassess the impact of multiple serum levels on the risk of death will reduce potential bias imposed by ignoring the correlation be- tween the longitudinal exposures pervading in the more commonly used univariate joint modeling (UVJM) approach. Joint modeling considering multiple longitudinal biomarkers of prognosis simultaneously can provide more accurate prediction of survival, modeling their interrelationship, correlation and uncertainty. Moreover, in traditional linear mixed-effects models, random errors are often under a normality assump- tion due to the mathematical tractability and computational convenience. However, normality assumption may not be realistic. Alternatively, skew-normal (SN) distribution should be more appropriate to model the skewed data [22,38]. The aim of this paper is to demonstrate an introductory overview on MVJM approach for longitudinal exposures and time-to-death in a specific application to Mayo Clinic PBC data, which enables fitting of such model have high dimensional longitudinal exposures. Here, we examined the association of the three longitudinal biomarkers serum bilirubin (SB), serum albumin (SA) and serum Glutamic-Oxaloacetic transaminase (SGOT) with skew-normal (SN) distribution (i.e., estimated bilirubin at baseline and change rate over time) with the risk of death simultaneously; the result of UVJM with SN distribution was presented in Supplemental for comparison.

Methods

Primary biliary cirrhosis (PBC) background and datasets

This dataset is from Mayo Clinic trail in PBC of the liver conducted between 1974 to 1984 [15,31]. The data were collected to examine the progress of PBC patients. A total of 424 PBC patients met the eligibility criteria for the randomized placebo controlled trail of the drug D-penicillamine, referred to Mayo Clinic during that ten-year interval. This dataset contained multiple laboratory results, but only on the first 312 patients in the dataset participated in the randomized trail and obtain large complete data. The additional 112 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival.

The original clinical protocol for these patients specified visits at 6 months, 1 year, and annually thereafter. This is an ideal dataset to illustrate the various features of MVJM. The dataset collected clinical, demographic and biochemical risk factors for each patient. Demographic factors: age and sex of patients; biochemical factors: drug (D-penicillin and placebo group), ascites (accumulation of wa- ter in the

doi: 10.7243/2053-7662-9-2

abdomen due to liver failure, presence of ascites 0=No, 1=Yes), hepatomogaly (liver growth status, presence of hepatomogaly 0=No, 1=Yes), edema (presence of edema, 0=No, 1=Yes: edema present without diuretics or edema despite diuretic therapy) and histological stage(≥3 is yes).

SB (*mg/dl*), SA (*mg/dl*) and SGOT (*U/ml*) values were taken as biochemical properties. Sample sizes and descriptive statistics for key variables at study entry were shown in **Table 1**. Only the patients who had three or more measurements of all the biomarkers were considered in the analysis due to the feature of our longitudinal modeling (quadratic of year was included). Thus, 259 patients who had sufficient repeated biomarkers observations were included in the analysis, where 111(42.9%) died during the study. **Table 1** summarized the demographic characteristics of patients in whom died and censored. Of the 259 patients in the PBC dataset, mean age at baseline of this PBC dataset was 45.53±10.42 years and mean age of death for PBC patients was 52.50±10.30 years. Baseline value of natural logarithms of SB, SA and SGOT were 0.47±0.95(*mg/dl*), 1.21±0.16(*mg/dl*) and 4.70±0.45(*U/ml*), respectively. **Figure 1** displays the histogram of repeated SB, SA and SGOT measurements(in log scale) for 259 subjects enrolled in Mayo Clinic trail study [**15**,**31**]. It is seen that for this data set to be analyzed in this paper, these longitudinal data (even after log-transformation) are relatively skewed, and thus a normality assumption may be violated.

Longitudinal exposures and survival outcome

Patients with PBC had abnormalities in several blood tests, such as elevated levels of SB. Several laboratory tests had a baseline measurement and were followed longitudinally at 6 months and at yearly intervals thereafter. Data collected at each lab visit include: total SB, SA, SGOT, gender, presence of ascites and other covariates for the patients. Due to the skewness of the observed biomarkers, specially SB, we took the natural logarithm of them and used the logarithms of those biomarkers for the remainder of this analysis. **Figures 2A-2C** showed the randomly picked 50 sample trajectories of natural logarithms of SB, SA and SGOT; moreover, Kaplan-Meier (K-M) survival curve depicted in **Figure 2D** may be dependent on the multivariate longitudinal exposures.

Traditionally, the longitudinal measurements are modeled using the linear mixed-effects model for a continuous and normally distributed outcome [**25**]. The subject-specific

Table 1. Descriptive statistics for variables measured for PBC dataset (N=259). Proportion for categorical variables and mean (SD) for quantitative variables.

Variables	Total	Death Patients	Survived Patients
Number of Patients	259	111	148
Gender			
Female	228(88.03%)	89(80.18%)	139 (93.92%)
Male	31(11.97%)	22(19.82%)	9(6.08%)
Drug			
Yes	128(49.42%)	56(50.45%)	72(48.65%)
No	131 (50.58%)	55(49.55%)	76(51.35%)
Edema			
Yes	41(15.83%)	25(22.52%)	26(10.81%)
No	218(84.7%)	86(77.48%)	132(89.19%)
Histologic≥3			
Yes	186(71.81%)	90(81.08%)	96(64.86%)
No	73(18.19%)	21(18.92 %)	52(35.14%)
Ascites			
Yes	10(3.86%)	9(8.11%)	1(0.68%)
No	249(96.14%)	102(91.89%)	147(99.32%)
Hepatomegaly			
Yes	125(48.26%)	73(65.77%)	52(53.14%)
No	134(51.73%)	38(34.23 %)	96(64.86%)
Baseline Value			
Age(SD*)	49.53(10.42)	52.50(10.30)	47.30(9.99)
ln(SB)(SD*)	0.58(1.09)	0.93(1.01)	0.12(0.74)
ln(SA)(SD*)	1.26(0.11)	1.24(0.13)	1.28(0.09)
ln(SGOT)(SD*)	4.70(0.446)	4.83(0.45)	4.61(0.42)

*: Standard deviation.

doi: 10.7243/2053-7662-9-2

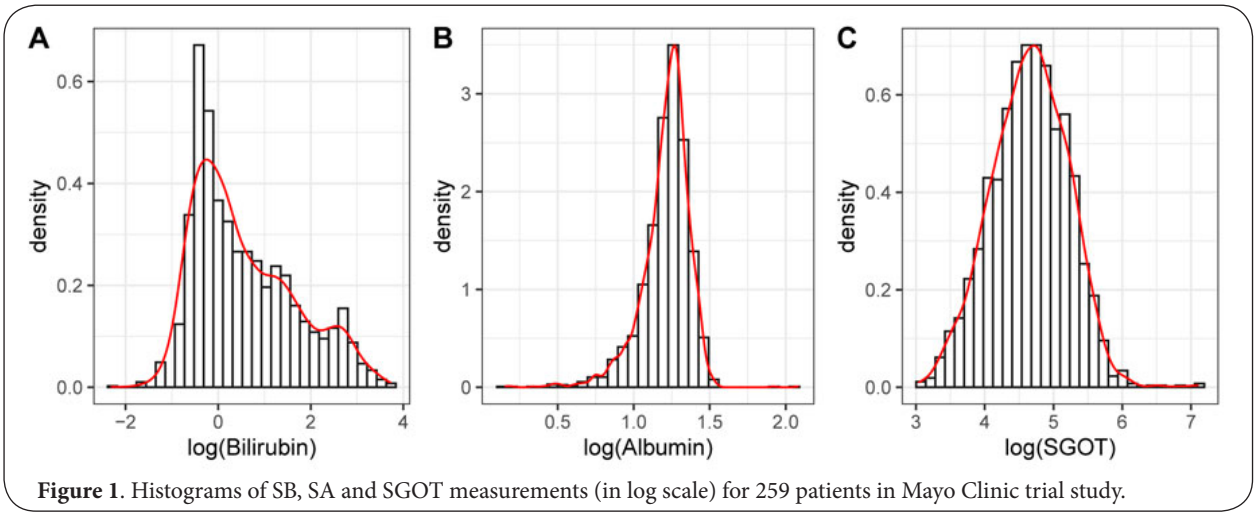

Figure 1. Histograms of SB, SA and SGOT measurements (in log scale) for 259 patients in Mayo Clinic trial study.

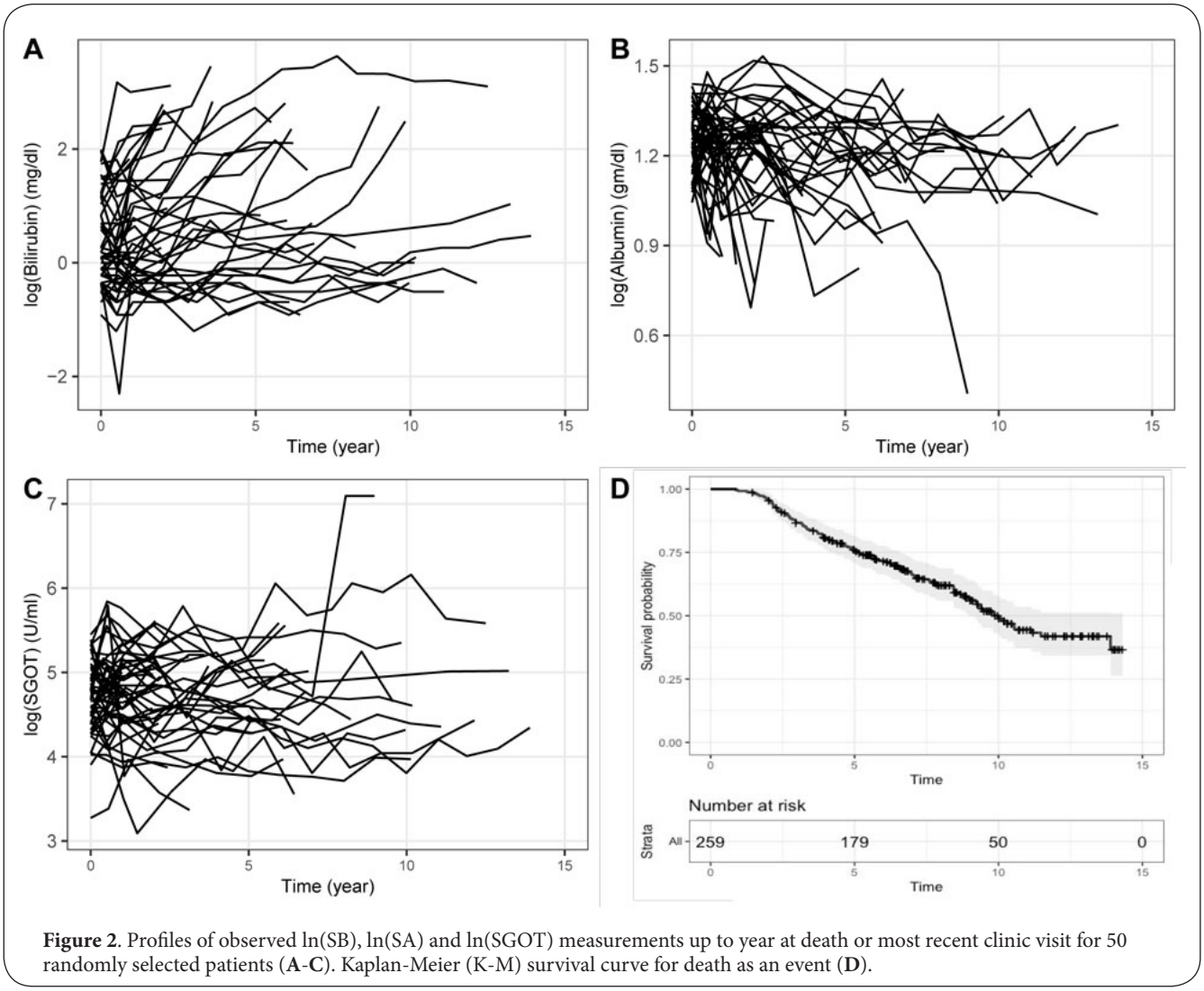

Figure 2. Profiles of observed ln(SB), ln(SA) and ln(SGOT) measurements up to year at death or most recent clinic visit for 50 randomly selected patients (**A**-**C**). Kaplan-Meier (K-M) survival curve for death as an event (**D**).

doi: 10.7243/2053-7662-9-2

random-effects in this mixed-effects model are included in the relative risk model. The various extensions under this topic have recently been investigated including, but not limited to, Dai and Pan [**10**] introduced nonparametric random mixture effect models to solve such joint modeling problems; Huang et al. [**19**,**20**,**21**,**30**] considered the various mixed-effects models with skewed distributions-based joint analysis for skewed-longitudinal and survival data. The random-effects not only account for the association between the longitudinal and survival out- comes, but also the correlation between the repeated measurements in the longitudinal process. In the PBC dataset, multiple longitudinal outcomes had been recorded. Extending the UVJM to accommodate those multiple longitudinal exposures allows us to incorporate more information and thereby improves the prognostic ability of the modeling. In this analysis, we were interested in the association between the three longitudinal biomarkers (SB, SA, SGOT) and the risk of death.

We observed the heterogeneity of each of the biomarkers for those patients. In clinical study, longitudinal exposures of the measurements of SB, SA and SGOT were usually measured with substantial errors. Meanwhile, after taking the natural logarithms of the three longitudinal exposures, they were still skewed. Various covariate mixed-effects models were discussed in the literature, but most of them assumed normal distribution for the error term which may lack the robustness against departures from normality in practice [**6**,**29**,**45**]. In order to relax the normality assumption and make a robust inference, the model was assumed to follow SN distribution. We offered an MVJM with subject-specific random intercept, random slope and quadratic of year and gender of the three serum levels for the longitudinal sub-model; and those six subject-specific random quantities were served as surrogate covariates in the survival sub-model. Therefore, an appropriate statistical model was needed to capture all the different biomarkers trajectories with the risk of death.

Multivariate joint modeling

There are two basic components of a joint modeling: the longitudinal component and the time-to-event (survival) component. In this regard, longitudinal measurements and time-to-event outcomes need to be modeled simultaneously in order to account for all information and uncertainty from both components, and understand the relationship between the underlying longitudinal data and the hazard for the event. **Figure 2** presented the underlying causal diagram for joint modeling mechanism.

The MVJM with multiple longitudinal exposures and SN distribution here was considered as an extension of a traditional UVJM. The MVJM consisted of a multivariate linear mixed-effects model for longitudinal SB, SA and SGOT exposures (the longitudinal sub-model) and a Cox proportional hazards model for the time of death as the outcome (the survival sub-model), which linked through the common subject-specific random-effects (intercepts and slopes) from the three functional forms to bring these two data types together into a single (multivariate joint) model, enabling better inference of the correlation, interplay and association between multiple longitudinal and time-to-event data. Moreover, in order to relax the normality assumption and make a robust inference, we assumed the multivariate longitudinal model with SN distribution. Based on both clinical significance and model selection criteria, we included the following covariates in the longitudinal sub-model: quadratic of year and gender (male as reference). In the survival sub-model, we assessed the simultaneous associations of the six subject-specific random baseline and change rate estimated from the longitudinal sub-model as time-varying covariates with risk of death. The survival sub-model was also adjusted for age at baseline, gender(male as reference), ascites(yes or no), hepatomegaly(yes or no), edema(yes or no) and histologic stage ≥3 (yes or no).

As described above of a general multivariate linear mixed-effects model with SN distribution, the fol-lowing was the specific MVJM for analyzing three longitudinal exposures of SB, SA and SGOT measures and time-to-

$$\begin{cases} \ln(\mathrm{SB}) = (\beta_{01} + b_{i01}) + (\beta_{11} + b_{i11})\mathrm{year}_{ij} + \beta_{21}\mathrm{year}_{ij}^2 + \beta_{31}\mathrm{gender}_i + \epsilon_{i1}, \\ \ln(\mathrm{SA}) = (\beta_{02} + b_{i02}) + (\beta_{12} + b_{i12})\mathrm{year}_{ij} + \beta_{22}\mathrm{year}_{ij}^2 + \beta_{32}\mathrm{gender}_i + \epsilon_{i2}, \\ \ln(\mathrm{SGOT}) = (\beta_{03} + b_{i03}) + (\beta_{13} + b_{i13})\mathrm{year}_{ij} + \beta_{23}\mathrm{year}_{ij}^2 + \beta_{33}\mathrm{gender}_i + \epsilon_{i3}, \\ \boldsymbol{b}_i \sim N_6(\mathbf{0}, \boldsymbol{\Sigma}_b), \quad \boldsymbol{\epsilon}_i \sim SN_{3n_i}\left(-\sqrt{2/\pi}[\boldsymbol{\Delta}_3 \otimes \mathbf{1}_{n_i}], \boldsymbol{\Sigma}_3 \otimes \boldsymbol{I}_{n_i}, \boldsymbol{\Delta}_3 \otimes \boldsymbol{I}_{n_i}\right), \\ \lambda_i(t_i|\boldsymbol{b}_i, Z_i) = \lambda_0(t_i) exp(\boldsymbol{\gamma}^T \boldsymbol{b}_i + \boldsymbol{\alpha}^T \boldsymbol{Z}_i), \end{cases} \quad (1)$$

where ln(SB), ln(SA) and ln(SGOT) are the respective natural logarithms of SB, SA and SGOT observations for the i^{th} subject at time $year_{ij}$. The vector of random errors $\boldsymbol{s}_i = (s_{i1}, s_{i2}, s_{i3})^T$ follows a multivariate SN distribution with unknown variance-covariance matrix $\boldsymbol{\Sigma}_3 = (\sigma^2_{kk'})_{3\times3}$ ($k, k' = 1, 2, 3$), unknown skew- ness parameter matrix $\boldsymbol{\Delta}_3 = \mathrm{diag}(\delta_1, \delta_2, \delta_3)$, where the vector of skewness parameters $\boldsymbol{\delta}_3 = (\delta_1, \delta_2, \delta_3)^T$, and $\mathbf{1}_{n_i} = (1, \ldots, 1)^T$. Note that $-\sqrt{2/\pi}[\boldsymbol{\Delta}_3 \otimes \mathbf{1}_n]$ is specified here to make the SN distribution with mean zero. b_{i01}, b_{i02} and b_{i03} are subject-specific random baselines (intercepts), and b_{i11}, b_{i12} and b_{i13} are subject-specific random change rates (slopes) for the three longitudinal exposures applied in this study. $\boldsymbol{b}_i = (b_{i01}, b_{i02}, b_{i03}, b_{i11}, b_{i12}, b_{i13})^T$ is a vector of random-effects which follows $N_6(\mathbf{0}, \Sigma_b)$ with Σ_b being a covariance matrix. $\beta = (\beta_{01}, \beta_{11}, \beta_{21}, \beta_{31}\beta_{02}, \beta_{12}, \beta_{22}, \beta_{32}, \beta_{03}, \beta_{13}, \beta_{23}, \beta_{33})^T$. This longitudinal sub-model assumes the mean baseline measurement, mean change rate and quadratic rate of year are different between male and female. For the survival sub-model, where $\gamma = (\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6)^T$ is the vector of the parameters corresponding to the random-effects $\boldsymbol{b}_i$ and the vector $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6)^T$ is the coefficient parameters corresponding to the risk factor vector $\boldsymbol{Z}_i$ which includes age at baseline, gender (male as reference), ascites (yes

 doi: 10.7243/2053-7662-9-2

or no), hepatomegaly (yes or no), edema (yes or no) and histologic stage ≥3 (yes or no). In the PBC study data, after taking natural logarithms of the three longitudinal exposures (SB, SA and SGOT), they still exhibits skewness and outliers as shown in **Figure 1**. Thus, we assume the multivariate linear mixed-effects models with SN distribution.

The MVJM with SN distribution was applied to assess the simultaneous effects of the three longitu- dinal biomarkers on risk of death. The fully Bayesian methodology using Markov Chain Monte Carlo (MCMC) techniques was adopted for the MVJM fitting and data analysis using the R software with as- sociated R2Winbugs [**41**]. The simultaneous statistical inference on all unknown population parameters $\theta = (\beta, \alpha, \gamma, \Sigma_3, \Sigma_b, \delta_3)^T$ can capture the underlying association between the longitudinal responses and the time-to-event data. To carry out the Bayesian inference, we need to specify prior distributions with the values of the hyper-parameters for all population parameters. Due to the absence of historical data, we applied weakly informative prior distributions for the unknown population parameters in MVJM. In particular, we specified the values for the hyper-parameters in the prior distributions as follows. (i) each element of the population coefficient vectors β, α, and γ was taken to be the normal distribution *N* (0, 100); (ii) the priors for the variance covariance matrices Σ_3 and Σ_b were taken to be inverse Wishart distribu- tions *IW* (diag(0.01, 0.01, 0.01), 4) and *IW* (diag(0.01, 0.01, 0.01, 0.01, 0.01, 0.01), 7), respectively; (iii) each of the skewness parameters δ_1, δ_2 and δ_3, which represent skewness of SB, SA and SGOT measurements, respectively, follows the normal distribution *N* (0, 100).

Multivariate Joint Modeling Results

Table 2 and **Figure 4** presented the results of joint analysis for three longitudinal SB, SA and SGOT levels in natural log-transformed and time-to-death of PBC data based on the MVJM with SN distribution. We fitted multivariate linear mixed-effects model for each biomarkers with a patient-specific baseline value and patient-specific change rate with covariates quadratic term of year and gender. It was shown from the results summarized in the upper half of **Table 2** that in the longitudinal sub-model, the estimated results indicated the skewness in SB (δ_1=0.42), SA (δ_2=−0.03) and SGOT (δ_3=0.095) after taking natural logarithms were estimated to be significant. This indicated after the natural log-transformed of the three biomarkers, the skewness with lightly right tail of the SB and SGOT, and fair lightly left tail of SA was still remain. Thus, it might suggest that accounting for an MVJM with the skewness distribution assumption provided a better fit to the data which exhibit skewness and, in turn, gave more reliable estimates of the parameters in comparison with its counterpart where the normal distribution is assumed. The estimated results of fixed-effects presented in **Table 2** indicated that the growth rate of SB, SA and SGOT with covariates quadratic term of year and gender might be approximated by $\ln(\hat{S}B)$=0.601+2.413year+0.523year^2−0.240gender, $\ln(\hat{S}A)$= 1.285-0.47 year−0.11year^2−0.013gender, and $\ln(\hat{SG}OT)$ =4.73−0.005year+0.502year^2−0.019gender, respectively.

On average, ln(SB) increased linearly, about 2.41 per year over time, combined a 0.52 acceleration; ln(SA) decreased linearly, about 0.47 per year over time, combined a 0.11 deceleration; and ln(SGOT) decreased linearly, about 0.005 per year over time, combined a 0.50 acceleration. The quadratic of year was all significant, gender was not significantly associated for the longitudinal sub-model of the MVJM analysis.

For the parameters of the survival sub-model, **Table 2** (bottom panel) presented simultaneous effects of the three serum levels as time-varying covariates on the risk of death for the PBC patients. The estimated hazard ratios of SB, SA and SGOT levels at baseline as time-varying covariates with the risk of death were 2.390 (SB, with 95% CI:(1.429,

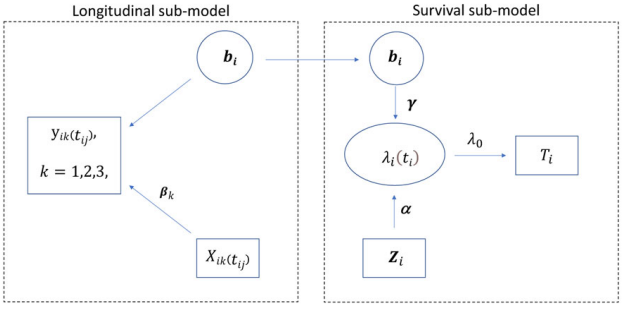

Figure 3. Causal diagram of the underlying mechanism of MVJM for the longitudinal and survival processes: Rectangles denote observed data; ellipse denotes unobserved quantities; circles denote un-observed (including subject-specific random bi) terms and unknown parameters. The arrows direct statistical dependencies. The causal chain of interest runs from longitudinal exposure y_{ik}(k = 1, 2, 3) to hazard λ_i(ti) for event, whereas y_{ik}(k = 1, 2, 3) does not "cause"event, but is statistically related to event through its dependence on the unobserved random quantities bi.

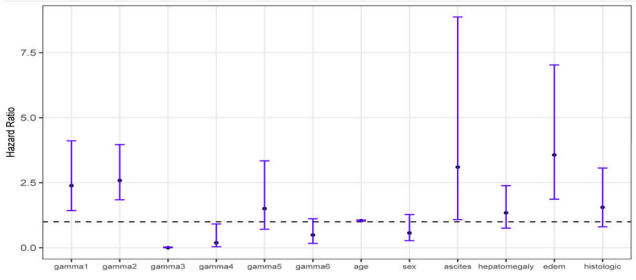

Figure 4. Hazard ratio (HR) denoted by solid dot along with 95% CI of the covariates using MVJM analysis. Here the baseline covariates are age, gender (Female vs. male as reference), Ascites status (yes or no), hepatomegaly status (yes or no), edema (yes or no) and histologic ≥3 (yes or no).

doi: 10.7243/2053-7662-9-2

Table 2. Summary of estimated posterior mean(PM), standard deviation(SD) of population (fixed- effects) parameters, and posterior mean(PM), standard deviation(SD) & hazard rate (HR) of Cox model parameters, the corresponding lower limit (LCI) and upper limit (UCI) of 95% equal-tail credible interval (CI).

Parameter	PM	SD	95%CI	HR	95%CI
Multivariate longitudinal Parameter estimates					
ln(SB)					
β_{01}	0.601	0.135	(0.346, 0.849)		
β_{11}	2.413	0.180	(2.014, 2.743)		
β_{21}	0.523	0.221	(0.081, 0.955)		
β_{31}	-0.240	0.146	(-0.512, 0.045)		
$\delta 1$	0.420	0.030	(0.354, 0.475)		
ln(SA)					
β_{02}	1.285	0.021	(1.243, 1.327)		
β_{12}	-0.470	0.045	(-0.558, -0.381)		
β_{22}	-0.110	0.051	(-0.222, -0.001)		
β_{32}	-0.013	0.023	(-0.058, 0.032)		
$\delta 2$	-0.030	0.010	(-0.051, -0.011)		
ln(SGOT)					
β_{03}	4.730	0.070	(4.594, 4.875)		
β_{13}	-0.005	0.1-8	(-0.221, 0.206)		
β_{23}	0.502	0.156	(0.204, 0.818)		
β_{33}	-0.019	0.076	(-0.162, 0.136)		
δ_3	0.095	0.032	(0.034, 0.161)		
Survival parameter estimates					
γ_1	0.871	0.268	(0.357, 1.414)	2.390	(1.429, 4.112)
γ_2	0.951	0.193	(0.613, 1.378)	2.588	(1.845, 3.967)
γ_3	-6.147	1.399	(-9.151,-3.549)	0.002	(0.001, 0.029)
γ_4	-1.655	0.812	(-3.307,-0.089)	0.191	(0.037, 0.915)
γ_5	0.409	0.394	(-0.338,1.206)	1.506	(0.713, 3.340)
γ_6	-0.712	0.467	(-1.777, 0.110)	0.490	(0.169, 1.116)
α_1	0.042	0.012	(0.019, 0.066)	1.043	(1.020, 1.069)
α_2	-0.567	0.397	(-1.297, 0.247)	0.568	(0.273, 1.280)
α_3	1.132	0.534	(0.080, 2.183)	3.102	(1.084, 8.873)
α_4	0.295	0.289	(-0.285, 0.871)	1.343	(0.752, 2.389)
α_5	1.272	0.333	(0.624, 1.950)	3.568	(1.866, 7.029)
α_6	0.442	0.341	(-0.216, 1.120)	1.556	(0.806, 1.065)

4.112)), 0.002 (SA, with 95% CI:(0.001, 0.029)) and 1.506 (SGOT, with 95% CI:(0.713, 3.340)); moreover, the estimated hazard ratios of change rate of SB, SA and SGOT as time-varying covariates with the risk of death were 2.588 (SB, with 95% CI:(1.845, 3.967)), 0.191 (SA, with 95% CI:(0.037, 0.915)) and 0.490 (SGOT, with 95% CI:(0.169, 1.116)) with adjustment for the additional covariates shown in **Figure 4**. These findings implied that there were significantly positive association of the SB levels and significantly negative association of the SA levels with the risk of death for PBC patients, but the SGOT levels did not significantly affect the risk of death for PBC patients. In other words, a 1-unit increase in baseline of natural log-transformed SB levels increased the death risk more than 2-fold (HR=2.390, 95% CI:(1.429, 4.112)), and a 1-unit increase in change rate of natural log-transformed SB levels increased the death risk approximately 2.59-fold (HR=2.588, 95% CI: (1.845, 3.967)). There was 80.9% (HR=0.191, 95% CI:(0.037, 0.915)) reduction in hazard of death relative to a 1-unit increase in change rate of natural log-transformed SA levels, and indicating that SA was a protective predictor for the risk of death for PBC patients. The effects of

 doi: 10.7243/2053-7662-9-2

baseline covariates in hazard were also presented in **Table 2** and **Figure 4**. An increase in age of one year (1.043 with 95% CI:(1.020, 1.069)) had a 4.3% increase in hazard for death. Patients with ascites (HR=3.102, 95% CI:(1.804, 8.873)) were estimated to have 2.1 times higher hazard for death as compared to patients without ascites; patients with edema (HR=3.568, 95% CI:(1.866, 7.029)) were estimated to have above 3-fold higher hazard for death than patients without edema. However, the gender (HR=0.568, 95% CI:(0.273, 1.280)), hepatomegaly (HR=1.343, 95% CI:(0.752, 2.389)) and histologic (HR=1.556, 95% CI:(0.806, 1.065)) were not found to be significantly associated with the risk of death.

Concluding Discussion

In this article, we have introduced the MVJM approach with SN distribution for simultaneously analyzing three longitudinal exposures and time-to-event data. Although joint modeling is widely recognized in the biostatistical literature and important in many application areas, allowing accurate inference of the dependency and association between these two types of data, this paper extended the traditonal joint modeling application to consider joint modeling of multivariate longitudinal exposures with skewness distribution and Cox proportional hazard model, accounting for multiple data features simultaneously. The principal benefits of this MVJM with SN distribution over traditional UVJM analysis is its efficient use of helping practitioners to analyze complicated multiple longitudinal exposures and time-to-event data under a wide range of considerations, which enables preciser inference. By using this advanced MVJM technique, we were able to obtain more accurate and reliable estimates by taking into consideration the highly corre- lated nature of the different serum levels and infer insights into the complex relationships among multiple longitudinal exposures and the risk of death for PBC patients.

We have reported the MVJM with SN distribution of a dataset from examining the progress of PBC in 259 patients in the Mayo Clinic to assess the dependency and association between the multiple longitudinal exposures over time and time-to-death outcome. The results from this MVJM revealed that both the baseline and change rate of SB trajectory over time were positively associated with the risk of death; both the baseline and change rate of SA trajectory over time were negatively associated with risk of death; but no significant association was found between the SGOT trajectory and death. From clinical perspective, these estimates were broadly in line, that higher SB was associated with higher risk of mortality, whereas higher SA was associated with lower risk of mortality. This study provides physicians a more flexible and dynamic model to discriminate patients using multiple biomarkers. Improving the knowledge about the course of PBC and its biomarkers is essential for the development and approval of new therapies.

To our knowledge, under the framework of Bayesian joint modeling for longitudinal and time-to-event data, no studies have examined multivariate longitudinal data based on SN distribution of biomarkers trajectory over time to predict survival event for the PBC study. In spite of these strengths, the following notes should be made in our study. First, including a second order polynomial in the longitudinal component, so we only select patients with at least 3 or more observations. As a result, the sample size was different from that in the analysis done by other researchers. Second, from the trajectory plots, the model specification in linear fashion may be inadequate to describe the time course of the repeated SA, SB and SGOT measurements, and the more complex models, such as the piecewise and nonlinear function-based models, should be considered. Finally, male patients were weighted with only light underwear on and thus their biomarkers value would be affected by lighter weight measurement, while undressing was not requested in female. This might introduce a small systematic bias. However, since 88.03% PBC patients were female in this study, which might indicate the majority of PBC patients were female.

Using an UVJM with SN distribution where only one longitudinal exposure serum was assessed as compared to MVJM with SN distribution, the estimated association of the time-varying serum levels with risk of death were biased which generally attenuates the risk estimates to the null, because it failed to take into account the uncertainty caused by the variations and correlations in the three longitudinally measured serum levels. We further explored the UVJM approach in the PBC data by fitting three UVJMs, separately, for each of the three serum (SB, SA and SGOT) levels and time to death (see Supplemental Material in detail). We found that, for the survival sub-model, the results showed considerably quantitative differences between the UVJM and MVJM analyses for most of the estimated parameters. This indicated the importance of recognizing the uncertainty caused by the additional variability and correlations in the observed values of SB, SA and SGOT levels.

In summary, the PBC data analysis using this MVJM with SN distribution has certain advantages compared to traditional joint modeling. First, the majority of joint modeling only focus on univariate longitudinal outcome associated with the time-to-event endpoint. However, in practice, multiple longitude outcomes are likely to be collected together and they may be highly correlated in clinical and observational studies. The MVJM analysis can reduce the bias and increase the efficiency in parameters estimation. The predictive capability will be improved by incorporating all sources of the data. These interesting findings have important clinical indications and lead to more informative inferences for the purpose of medical decision-making. Our results suggested that the biomarkers (SA, SB) had significant association with the risk of death. The MVJM

doi: 10.7243/2053-7662-9-2

analysis is applicable in any situation where one wishes to investigate association of longitudinal outcomes and survival outcome. Secondly, since it was of importance to measure SB, SA and SGOT appropriately when they exhibited skewness and heavy tails, even though after we took the natural logarithms, this analysis considered the estimation of skewness parameters δ_1, δ_2 and δ_3 were statistically significant for ln(SB), ln(SA) and ln(SGOT), indicating that the skewness exists in those biomarkers measurements. The MVJM approach with the SN distribution provided more efficient and accurate parameter estimation, as compared to existing joint modeling for the PBC study.

To conclude, the MVJM with skewness distribution analysis is an improvement over traditional joint modeling and survival model, because they consider all the longitudinal associated observations of covariates that are predictive to the survival outcome. This MVJM method can be applied to wide variaties of research setting to obtain better parameter estimation for future prediction in the medical research in comparison with traditioanl UVJM, since they are considered to account for individual variability. These predictions can provide relatively accurate characterizations of individual disease progression, which might be important for the timing of interventions, qualification for appropriate treatments, and additional other analysis. Although the application of MVJM is limited uptake by medical researchers, but it is useful and applicable to a broader field in clinical practice when multivariate longitudinal measurements and time-to-event data are available.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions

Authors' contributions	LX	YH	HC	AM	FC
Research concept and design	--	✓	--	--	--
Collection and/or assembly of data	✓	✓	--	--	--
Data analysis and interpretation	✓	✓	✓	--	--
Writing the article	✓	✓	--	--	--
Critical revision of the article	✓	✓	✓	✓	✓
Final approval of article	✓	✓	✓	✓	✓
Statistical analysis	✓	✓	--	--	--

Acknowledgment
The authors gratefully acknowledge the Editor, Section Editor and anonymous referees for their insightful comments and constructive suggestions that led to an improvement of the article. This study was partially supported by Internal Funding from College of Public Health at University of South Florida to all authors.

Publication history
Editor: Dr. Nicola Shaw. Algoma University, Canada.
Received: 22-April-2021 Final Revised: 18-June-2021
Accepted: 23-June-2021 Published: 02-Aug-2021

References
1. P.S. Albert and J.H. Shih, An approach for jointly modeling multivariate longitudinal measurements and discrete time-to-event data, The Annals of Applied Statistics 4 (2010), p. 1517.
2. C. Allen, S. Duffy, T. Teknos, M. Islam, Z. Chen, P.S. Albert, G. Wolf, and C. Van Waes, Nu-clear factor-κb–related serum factors as longitudinal biomarkers of response and survival in advanced oropharyngeal carcinoma, Clinical Cancer Research 13 (2007), pp. 3182–3190.
3. E.R. Andrinopoulou and D. Rizopoulos, Bayesian shrinkage approach for a joint model of longitudinal and survival outcomes assuming different association structures, Statistics in Medicine 35 (2016), pp. 4813–4823.
4. E.R. Brown and J.G. Ibrahim, A Bayesian semiparametric joint hierarchical model for longitudinal and survival data, Biometrics 59 (2003), pp. 221–228.
5. E.R. Brown, J.G. Ibrahim, and V. DeGruttola, A flexible b-spline model for multiple longitudinal biomarkers and survival, Biometrics 61 (2005), pp. 64–73.
6. R.J. Carroll, D. Ruppert, L.A. Stefanski, and C.M. Crainiceanu, Measurement error in nonlinear models: a modern perspective, CRC press, 2006.
7. Y. Chen and Y. Wang, Variable selection for joint models of multivariate longitudinal measurementsand event time data, Statistics in Medicine 36 (2017), pp. 3820–3829.
8. Y.Y. Chi and J.G. Ibrahim, Joint models for multivariate longitudinal and multivariate survival data,Biometrics 62 (2006), pp. 432–445.
9. M.J. Crowther, K.R. Abrams, and P.C. Lambert, Joint modeling of longitudinal and survival data, The Stata Journal 13 (2013), pp. 165–184.
10. H. Dai and J. Pan, Joint modeling of survival and longitudinal data with informative observation times, Scandinavian Journal of Statistics: Theory and Applications 45 (2018), pp. 571–589.
11. E. Dil and D. Karasoy, Joint modeling of a longitudinal measurement and parametric survival data with application to primary biliary cirrhosis study, Pakistan Journal of Statistics and Operation Research (2020), pp. 295–304.
12. J. Ding and J.L. Wang, Modeling longitudinal data with nonparametric multiplicative random effectsjointly with survival data, Biometrics 64 (2008), pp. 546–556.
13. R. Elashoff, N. Li, et al., Joint modeling of longitudinal and time-to-event data, CRC Press, 2016.
14. S. Fieuws and G. Verbeke, Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles, Biometrics 62 (2006), pp. 424–431.
15. T. Fleming and D. Harrington, Counting processes and survival analysis john wiley & sons, Inc. NewYork (1991).
16. M.E. Gershwin, C. Selmi, H.J. Worman, E.B. Gold, M. Watnik, J. Utts, K.D. Lindor, M.M. Ka-plan, J.M. Vierling, and U.P.E. Group, Risk factors and comorbidities in primary biliary cirrhosis: a controlled interview-based study of 1032 patients, Hepatology 42 (2005), pp. 1194–1202.
17. R. Henderson, P. Diggle, and A. Dobson, Joint modelling of longitudinal measurements and event time data, Biostatistics 1 (2000), pp. 465–480.
18. G.L. Hickey, P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona, joinerml: a joint model and software package for time-to-event and multivariate longitudinal outcomes, BMC Medical Research Methodology 18 (2018), p. 50.
19. Y. Huang, J. Chen, and P. Yin, Hierarchical mixture models for longitudinal immunologic data with heterogeneity, non-normality and missingness, Statistical Methods in Medical Research 26 (2017), pp. 223–247.
20. Y. Huang, J. Hu, and G. Dagne, Joint modeling time-to-event and longitudinal data: A bayesian approach, Statistical Method &

Applications 23 (2014), pp. 95–121.

21. Y. Huang and J. Chen, Bayesian quantile regression-based nonlinear mixed-effects joint models for time-to-event and longitudinal data with multiple features, Statistics in Medicine 35 (2016), pp. 5666–5685.
22. Y. Huang, R. Chen, G. Dagne, Y. Zhu, and H. Chen, Bayesian bivariate linear mixed-effects mod- els with skew-normal/ independent distributions, with application to aids clinical studies, Journal of Biopharmaceutical Statistics 25 (2015), pp. 373–396.
23. Y. Huang and G. Dagne, A bayesian approach to joint mixed-effects models with a skew-normal dis- tribution and measurement errors in covariates, Biometrics 67 (2011), pp. 260–269.
24. S. Kim and P.S. Albert, A class of joint models for multivariate longitudinal measurements and abinary event, Biometrics 72 (2016), pp. 917–925.
25. N.M. Laird and J.H. Ware, Random-effects models for longitudinal data, Biometrics (1982), pp. 963– 974.
26. W.J. Lammers, G.M. Hirschfield, C. Corpechot, F. Nevens, K.D. Lindor, H.L. Janssen, A. Floreani, C.Y. Ponsioen, M.J. Mayo, P. Invernizzi, et al., Development and validation of a scoring system to predict outcomes of patients with primary biliary cirrhosis receiving ursodeoxycholic acid therapy, Gastroenterology 149 (2015), pp. 1804–1812.
27. W.J. Lammers, H.R. Van Buuren, G.M. Hirschfield, H.L. Janssen, P. Invernizzi, A.L. Mason, C.Y. Ponsioen, A. Floreani, C. Corpechot, M.J. Mayo, et al., Levels of alkaline phosphatase and bilirubin are surrogate end points of outcomes of patients with primary biliary cirrhosis: an international follow-up study, Gastroenterology 147 (2014), pp. 1338–1349.
28. H. Lin, C.E. McCulloch, and S.T. Mayne, Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables, Statistics in Medicine 21 (2002), pp. 2369–2382.
29. W. Liu and L. Wu, Simultaneous inference for semiparametric nonlinear mixed-effects models with covariate measurement errors and missing responses, Biometrics 63 (2007), pp. 342–350.
30. X. Lu, Y. Huang, J. Chen, R. Zhou, S. Yu, and P. Yin, Bayesian joint analysis of heterogeneous- and skewed-longitudinal data and a binary outcome, with application to aids clinical studies, Statistical Methods in Medical Research 27 (2018), pp. 2946–2963.
31. P.A. Murtaugh, E.R. Dickson, G.M. Van Dam, M. Malinchoc, P.M. Grambsch, A.L. Langworthy, and C.H. Gips, Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits, Hepatology 20 (1994), pp. 126–134.
32. J. Proudfoot, W. Faig, L. Natarajan, and R. Xu, A joint marginal-conditional model for multivariatelongitudinal data, Statistics in Medicine 37 (2018), pp. 813–828.
33. E. R. Brown and J. G. Ibrahim, A bayesian semiparametric joint hierarchical model for longitudinaland survival data, Biometrics 59 (2003), pp. 221–228.
34. D. Rizopoulos, Dynamic predictions and prospective accuracy in joint models for longitudinal andtime-to-event data, Biometrics 67 (2011), pp. 819–829.
35. D. Rizopoulos, Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptivegaussian quadrature rule, Computational Statistics & Data Analysis 56 (2012), pp. 491–501.
36. D. Rizopoulos, Joint models for longitudinal and time-to-event data: With applications in R, CRCPress, 2012.
37. D. Rizopoulos and P. Ghosh, A bayesian semiparametric multivariate joint model for multiple longi-tudinal outcomes and a time-to-event, Statistics in Medicine 30 (2011), pp. 1366–1380.
38. S.K. Sahu, D.K. Dey, and M.D. Branco, A new class of multivariate skew distributions with applicationsto Bayesian regression models, Canadian Journal of Statistics 31 (2003), pp. 129–150.
39. C. Selmi, R.L. Coppel, and M.E. Gershwin, Primary biliary cirrhosis, in The Autoimmune Diseases,Elsevier, 2006, pp. 749–765.
40. J. Shapiro, H. Smith, and F. Schaffner, Serum bilirubin: a prognostic factor in primary biliary cirrho-sis., Gut 20 (1979), pp. 137–140.
41. S. Sturtz, U. Ligges, and A.E. Gelman, R2winbugs: a package for running winbugs from r (2005).
42. A.M. Tang, N.S. Tang, and H. Zhu, Influence analysis for skew-normal semiparametric joint modelsof multivariate longitudinal and multivariate survival data, Statistics in Medicine 36 (2017), pp. 1476– 1490.
43. A.M. Tang, X. Zhao, and N.S. Tang, Bayesian variable selection and estimation in semiparametric joint models of multivariate longitudinal and survival data, Biometrical Journal 59 (2017), pp. 57–78.
44. A.A. Tsiatis and M. Davidian, Joint modeling of longitudinal and time-to-event data: an overview, Statistica Sinica 14 (2004), pp. 809–834.
45. L. Wu, A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to aids studies, Journal of the American Statistical Association 97 (2002), pp. 955–964.
46. H. Zhang, Y. Huang, W. Wang, H. Chen, and B. Langland-Orban, Bayesian quantile regression- based partially linear mixed-effects joint models for longitudinal data with multiple features, Statistical Methods in Medical Research 28 (2019), pp. 569–588.

Citation:
Xu L, Huang Y, Chen H, Mbah A and Cheng F. **Joint Modeling Analysis of Multivariate Skewed-longitudinal and Time-to-event Data with Application to Primary Biliary Cirrhosis Study**. *J Med Stat Inform*. 2021; **9**:2.
http://dx.doi.org/10.7243/2053-7662-9-2