Journal of Statistical Research *https://doi.org/10.3329/jsr.v58i1.75417*
2024, Vol. 58, No. 1, pp. 111-130 *ISSN 0256 - 422 X*

A QUANTILE-REGRESSION APPROACH TO BIVARIATE LONGITUDINAL JOINT MODELING

DAMITRI KUNDU
Applied Statistics Division, Indian Statistical Institute, Kolkata, India

KIRANMOY DAS*
Applied Statistics Division, Indian Statistical Institute, Kolkata, India
Beijing Institute of Mathematical Sciences and Applications, Beijing, China
Email: kiranmoy.das@gmail.com

SUMMARY

Joint modeling of longitudinal outcomes and time-to-event data has become a major research interest in the last thirty years. A joint model is useful since it helps (i) to understand the evolution of the outcome(s) of interest over time, (ii) to understand the effects of the outcomes on the time of occurrence of some event(s) of interest (e.g. death/relapse), and (iii) to study the effects of the time-varying and time-invariant predictors on the longitudinal and time-to-event process. Traditional linear mixed models are routinely used for modeling the longitudinal process. However, for non-Gaussian/skewed outcomes it is more appealing to use quantile-regression models since such models do not assume any specific probability distribution for the outcomes. In this article, we present a bivariate quantile-regression approach for jointly modeling longitudinal process and time-to-event. In a Bayesian setting we consider an Asymmetric Laplace Distribution (ALD) for modeling different quantiles of the outcomes, and a semi-parametric Proportional Hazards (PH) model for time-to-event. Model parameters are estimated using Markov chain Monte Carlo (MCMC) algorithm, and we discuss the computational complexities through several simulation studies. Our numerical studies illustrate the usefulness of our model over the other traditional models.

Keywords and phrases: Asymmetric Laplace Distribution (ALD), Bivariate longitudinal data, Joint model, MCMC, Quantile regression.

1 Introduction

Joint modeling of longitudinal outcomes and time-to-event data has drawn attention of the researchers in the last thirty years mainly because of its wide applications in various disciplines including biomedical studies. For example, in the traditional clinical trials, some specific drugs are given to a group of patients over a period of time, and the outcomes of interest are measured longitudinally along with the time of occurrence of some event of interest (e.g. survival/relapse). In agricultural studies, biomass of some plants are measured longitudinally and the time to get the first flower is

* Corresponding author

© Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka 1000, Bangladesh.

observed for each plant. Such an experiment can reveal the effect of biomass on the reproductive time of the plants. Statistical models have been developed for jointly modeling the longitudinal outcomes and the time-to-event instead of modeling them separately. A joint model is more informative than separate models since it helps to understand (i) how the outcomes of interest evolve over time, (ii) the effects of the outcomes on the time of occurrence of the event, and (iii) the effects of the covariates (time-varying and time-invariant) on the longitudinal and time-to-event process.

The literature on joint modeling is indeed quite rich. Henderson (2000) proposed a likelihood based approach for jointly modeling univariate longitudinal outcome and an event-time. In a Bayesian framework Wang and Taylor (2001) proposed a joint model for modeling CD4 count and the time to progress into AIDS for HIV patients. In a similar setting, Guo and Carlin (2004) developed joint models for comparing the efficacy of two drugs. Fiews and Verbeke (2004), Chi and Ibrahim (2006), Rizopoulos and Ghosh (2011) developed joint models for multiple longitudinal outcomes and an event-time. Das et al. (2012, 2016) developed a Bayesian model for jointly modeling biomass and the reproductive time for soybean plants with a goal of identifying the genetic markers controlling the biomass and the reproductive time. Rizopoulos et al. (2017) proposed an efficient joint model for dynamically predicting the survival probability of each subject over time. More recently, Kundu et al. (2024a) developed a Bayesian joint model for three important biomarkers (white blood cell count, neutrophil count and platelet count) and time to relapse for acute lymphocytic leukemia (ALL). However, a common feature of all these works is that a linear mixed model is used for the longitudinal process, assuming that the outcomes are Gaussian, and a proportional hazards (PH) or an accelerated failure time (AFT) model is used for the time-to-event process.

In many real applications, we come across non-Gaussian outcomes; and the interest might be on modeling (or predicting) some specific quantiles of the outcomes. For example, a medical researcher might be interested in predicting the time-to-event for patients with a higher (or other extreme) value of certain outcome(s) of interest. Traditional linear mixed models which rely on the normality for the outcome of interest fail to handle such datasets. In such cases, quantile regression models are more appropriate; and therefore, we present a (bivariate) quantile regression model for jointly modeling the longitudinal and the time-to-event process. Koenker and Bassett (1978) developed quantile regression model where the effects of the covariates can be assessed at different quantile levels. Koenker (2004) also developed quantile regression model for longitudinal outcomes. Based on an Asymmetric Laplace Distribution (ALD) Yu and Moyeed (2001) developed a Bayesian quantile regression model, and Geraci and Bottai (2007) extended such models for longitudinal outcomes. Kozumi and Kobayashi (2011) showed that an Asymmetric Laplace Distribution can be expressed as a scale mixture of normals with an exponential scale distribution. Biswas and Das (2021) developed a computationally efficient Gibbs sampler for modeling the quantiles of multivariate longitudinal outcomes. Yang et al. (2019) proposed a Bayesian quantile joint regression model for predicting the development of Huntington's disease. Similar joint models are proposed in Zhang and Huang (2020) for analyzing data from a Multicenter AIDS Cohort Study.

In this paper, we develop a Bayesian joint model for bivariate longitudinal outcomes and time-to-event at different quantile levels. For the longitudinal process we consider a linear mixed model but assume an Asymmetric Laplace Distribution for each outcome variable. The longitudinal dependence among the outcomes at different time points and the dependence between the outcomes are simultaneously modeled by considering a Bivariate Brownian Motion (BBM) for the subject-specific random effects. Proportional Hazards models are used for the time-to-event process, and the model parameters are estimated by Gibbs sampler. We address various computational issues and assess the effectiveness of the proposed approach by extensive simulation studies.

The rest of this paper is organized as follows. In Section 2, we discuss the motivation for our joint quantile modeling. In Section 3, we describe the proposed model in detail and derive the joint posterior distribution. The findings from two simulation studies are discussed and summarized in Section 4. The proposed model is also compared with some other traditional models popularly used in joint modeling literature. Finally, some concluding remarks are given in Section 5.

2 A Simulated Dataset and Motivation

We consider two simulated datasets in this paper to illustrate the effectiveness of our proposed quantile regression model for jointly modeling bivariate longitudinal and time-to-event data. We note that a real dataset could definitely be more appealing, but unfortunately due to the unavailability of an appropriate dataset at hand we describe the usefulness of our proposed model through simulated datasets.

Let us consider the scenario where two biomarkers are being recorded over time along with some time-invariant covariates with a goal of modeling the relapse-time of a disease, and also to estimate the effectiveness of the drugs used for preventing a relapse. We simulate bivariate longitudinal outcomes for 281 subjects where the number of measurements differs from one subject to the other but the starting time is the same for all subjects. We consider two time-varying covariates (which we refer to as medicines/drugs) and four time-invariant covariates. The event of interest is referred to a “relapse” which is typically the case for cancer patients, and the subjects are censored after some subject-specific time points. So, each subject in the end can be classified either as a case of relapse or a non-relapse.

Two outcomes are denoted by Y_a and Y_b, respectively, and in Figure 1 we show the bivariate quantile-quantile (QQ) plot for two outcomes. The plot indicates that the sample quantiles differ from the theoretical quantiles, indicating that the joint distribution of outcomes differs substantially from a bivariate normal distribution. Because of that we cannot use a traditional linear mixed model for modeling the longitudinal process.

Next, we show a contour plot in Figure 2. Contours are shown for some specific density levels (e.g. 0.3, 0.2, 0.1, 0.05 etc.), and for each density level we use a different color. We use solid curves for the subjects with a relapse and broken curves for those with no-relapse. The axes show five different quantile levels (i.e. 0.10, 0.25, 0.50, 0.75 and 0.90) for two outcomes. We see that along the green and the golden arrows solid curves dominate their broken counter-parts indicating that, for the higher quantile levels of Y_a, the probability of relapse increases. On the other hand, along the black and the brown arrows the broken curves dominate their solid counter-parts indicating that for the higher quantiles of Y_b, there is a higher chance of no-relapse. Hence, in one way this figure shows that the occurrence of the event (relapse/no relapse) depends on the longitudinal outcomes, and hence a joint modeling is meaningful. At the same time it also illustrates that the probability of the occurrence of the event differs from one quantile level to the other, and thus a quantile-specific joint model is more appropriate here. Thus, for jointly modeling the longitudinal outcomes and time-to-event, we recommend a QQ plot and a contour plot for understanding the underlying complexity in the dataset.

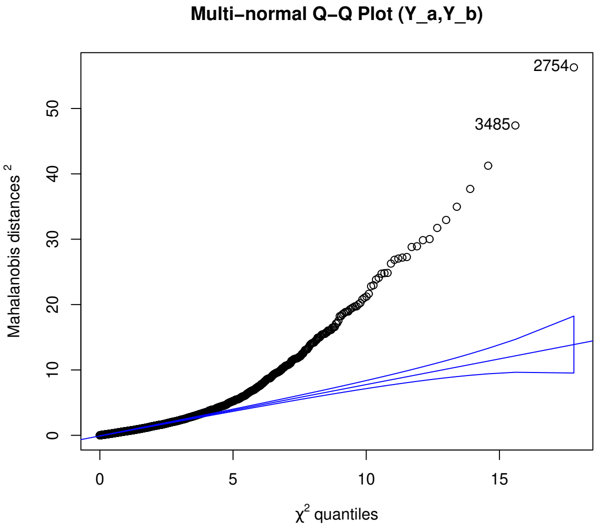

Figure 1: Quantile-Quantile plot for two outcomes in the simulated dataset.

3 Proposed Joint Model

As discussed in the earlier section, we propose a Bayesian quantile-specific joint model for bivariate longitudinal data and time-to-event. We note that the proposed methodology can be extended to a multivariate setting in a straightforward way.

We consider two longitudinal outcomes, hereafter referred to as biomarkers. We define a quantile level $\boldsymbol{\tau} = (\tau_a, \tau_b)$, where τ_k denotes the quantile level of the k-th biomarker, $k = a, b$. This notation illustrates that our proposed approach can handle different quantile levels for different biomarker, and hence quite flexible. Our proposed model has two parts, (i) a quantile-specific longitudinal submodel, and (ii) a time-to-event submodel with a semi-parametric proportional hazards model adjusted for each quantile level. Let Y_{ijk} denote the k-th biomarker measured from the i-th subject at the j-th time point ($j = 1, 2, \dots, t_i$), and the quantile-specific hazard for the i-th subject at time t is denoted by $\lambda^{(\boldsymbol{\tau})}(t)$. In addition, we define $s_i = min(T_i, C_i)$, where T_i denotes the relapse-time and C_i denotes the censoring time of the i-th subject, and thus s_i is considered as the survival time.

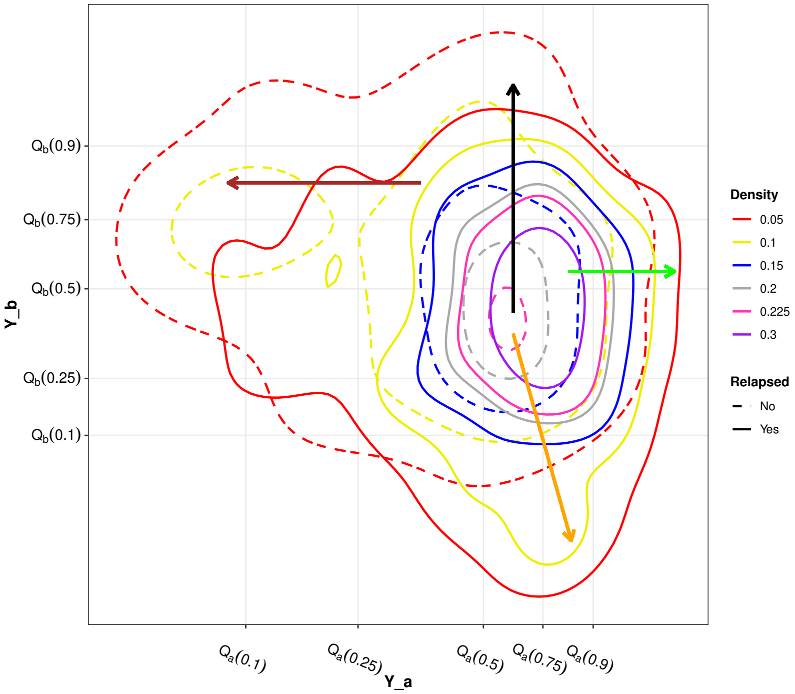

Figure 2: Contour plot of Y_a and Y_b for simulated dataset, with solid contours representing density levels for the candidates with a relapse, and broken curves for the candidates with no-relapse. Here, $\mathrm{Q}_a(u)$ and $\mathrm{Q}_b(u)$ represent the u-th quantile of Y_a and Y_b, respectively.

We define the indicator variable δ_i=1 when $T_i < C_i$; and 0, otherwise.

3.1 Longitudinal submodel

For simultaneously modeling the quantiles of the two biomarkers, we use the traditional linear mixed model following Wang and Taylor (2001), Rizupoulos and Ghosh (2011), Kundu et al. (2024b). Such models consider fixed effects of the covariates along with subject-specific random effects. In our setting, the biomarker quantiles (denoted by $Q^{(\tau)}(Y_{ijk})$) are modeled as follows:

$$Q^{(\tau)}(Y_{ijk}) = g_k^{(\tau)}(t_{ij}) + \boldsymbol{\beta}_k^{(\tau)T}\mathbf{x}_{ij} + \boldsymbol{\eta}_k^{(\tau)T}\mathbf{z}_i + \omega_{ik}^{(\tau)}(t_{ij}), \tag{3.1}$$

where g is the biomarker-specific general effect of time, and it can be modeled either by polynomial functions, or by splines, wavelets etc. For simplicity and computational ease, we consider a polynomial function of unknown order r, and model it as follows: $g_k^{(\tau)}(t) = \sum_{l=0}^{r} \zeta_{lk}^{(\tau)} t^l$. The unknown order r is obtained based on the information criteria, e.g. AIC, BIC, DIC etc. The regression coefficients, $\boldsymbol{\beta}_k^{(\tau)}$ and $\boldsymbol{\eta}_k^{(\tau)}$, are the quantile-specific fixed effects of two time-varying covariates ($\mathbf{x}_{ij}$) and the time-invariant covariates ($\mathbf{z}_i$), respectively. The subject-specific random effects $\omega_{ik}^{(\boldsymbol{\tau})}(t_{ij})$ capture the longitudinal dependence among the biomarkers at different times points and the dependence between two biomarkers over time.

We express equation (3.1) as follows:

$$Y_{ijk} = g_k^{(\boldsymbol{\tau})}(t_{ij}) + \boldsymbol{\beta}_k^{(\boldsymbol{\tau})T}\mathbf{x}_{ij} + \boldsymbol{\eta}_k^{(\boldsymbol{\tau})T}\mathbf{z}_i + \omega_{ik}^{(\boldsymbol{\tau})}(t_{ij}) + \epsilon_{ijk}, \tag{3.2}$$

where ϵ_{ijk} are random errors. Regression coefficients of this quantile regression model, for each fixed k, are estimated by minimizing the following quantile loss function:

$$L = \Big[(1-\tau_k) \sum_{Y_{ijk} < Q^{(\tau)}(Y_{ijk})} (Y_{ijk} - Q^{(\tau)}(Y_{ijk})) + \tau_k \sum_{Y_{ijk} > Q^{(\tau)}(Y_{ijk})} (Y_{ijk} - Q^{(\tau)}(Y_{ijk}))\Big]. \tag{3.3}$$

Koenker and Bassett (1978) proposed a convex optimization approach for minimizing the above loss function. Yu and Moyeed (2001), Geraci and Bottai (2007), Kulkarni et al. (2019) noticed that by assuming an Asymmetric Laplace Distribution (ALD) with location parameter=0, scale parameter=σ_k, and skewness parameter=τ_k, ($k = a, b$) for the random errors ϵ_{ijk} and then by maximizing the log-likelihood function, one can get exactly the same estimates.

Kozumi and Kobayashi (2011) showed that ALD can be written as a scale mixture of normals with exponential scales. Based on that we can write ϵ_{ijk} as follows:

$$\epsilon_{ijk} = \theta_{1k}e_{ijk} + \theta_{2k}\sqrt{\sigma_k e_{ijk}}v_{ijk},$$

where $\theta_{1k} = (1-2\tau_k)/\{\tau_k(1-\tau_k)\}$, and $\theta_{2k} = \sqrt{2/\{\tau_k(1-\tau_k)\}}$, $e_{ijk} \overset{ind}{\sim} \text{Exp}(1/\sigma_k)$, and $v_{ijk} \overset{iid}{\sim} N(0,1)$. Biswas and Das (2021) used the similar representation for modeling quantiles of multivariate longitudinal outcomes. Thus, conditional on $\omega_{ik}^{(\boldsymbol{\tau})}(t_{ij})$ and e_{ijk}, the outcomes Y_{ijk} follow a normal distribution. In particular, we get the following hierarchical structure:

$$Y_{ijk}|e_{ijk}, \omega_{ik}^{(\boldsymbol{\tau})}(t_{ij}) \sim N\left(g_k^{(\boldsymbol{\tau})}(t_{ij}) + \boldsymbol{\beta}_k^{(\boldsymbol{\tau})T}\mathbf{x}_{ij} + \boldsymbol{\eta}_k^{(\boldsymbol{\tau})T}\mathbf{z}_i + \omega_{ik}^{(\boldsymbol{\tau})}(t_{ij}) + \theta_{1k}e_{ijk}, \theta_{2k}^2\sigma_k e_{ijk}\right),$$
$$e_{ijk}|\sigma_k \sim \exp(\frac{1}{\sigma_k}).$$

The random effects $\omega_{ik}^{(\boldsymbol{\tau})}(t_{ij})$ play an important role in modeling longitudinal outcomes since they handle both the inter-outcome and intra-outcome dependence over time. Traditionally, subject-specific random intercepts and random slopes (of time) are used to capture the inter-biomarker dependence and the biomarker-specific longitudinal dependence (Das 2016, Kulkarni et al. 2019, Kundu et al. 2024a). Specifically, one may consider the following structure: $\omega_{ik}^{(\boldsymbol{\tau})}(t_{ij}) = c_{ik} + d_{ik}t_{ij}$, where c_{ik} are the subject-specific random intercepts and d_{ik} are the subject-specific random slopes. Let $R_i = [c_{ik}, d_{ik}]^T$, and assume that R_i independently follow multivariate normal distribution with mean vector=0, and covariance matrix=Σ_R. Different inter-outcomes and intra-outcomes dependence are captured by matrix Σ_R. Although this is the most commonly used approach for modeling random effects, such specifications allow the deviations of the subject-specific outcomes from their respective means to follow a straight line path, and that is quite restrictive (Wang and Taylor 2001).

We follow an alternative and more flexible approach for modeling $\omega_{ik}^{(\boldsymbol{\tau})}(t_{ij})$ as proposed in Kundu et al. (2024b) where no additional restriction on the deviations is imposed. Wang and Taylor (2001) used Integrated Ornstein-Uhlenbeck (IOU) process for the random effects, and Kundu et al. (2024b) considered a Bivariate Brownian Motion (BBM) for the same. We write $\omega_{ik}^{(\boldsymbol{\tau})}(t_{ij})$ by step functions,

$$\omega_{ik}^{(\boldsymbol{\tau})}(t) = \sum_{j=1}^{M} w_{ijk}^{(\tau_k)} \mathbf{1}_{(t_{i,j-1}^w \leq t < t_{ij}^w)};$$

where $t_{i0}^w = t_{i1}$, $t_{i1}^w, \ldots t_{i,M-1}^w$ are $(M-1)$-point Gauss-Kronrod points in the interval $(t_{i1},\, s_i)$, and $t_{i,M}^w > s_i$. Additionally, $[w_{i11}^{(\boldsymbol{\tau})}, w_{i12}^{(\boldsymbol{\tau})}]^T = w_{i1}^{(\boldsymbol{\tau})} \sim N_2(\mathbf{0}, t_{i,0}^w \Sigma_{\boldsymbol{\tau}})$, and $w_{ij}^{(\tau)} = w_{i,j-1}^{(\tau)} + \sqrt{t_{i,j-1}^w - t_{i,j-2}^w}\boldsymbol{U}_{ij}^{(\tau)}$, with $\boldsymbol{U}_{ij}^{(\tau)} \overset{iid}{\sim} N_2(\mathbf{0}, \Sigma_{\boldsymbol{\tau}})$, $j = 2, \ldots, M$. In practice, the value of M is taken from 10 to 20, although in principle, it can be any value sufficiently large. Here, we take M=16 for our computation. This choice considers different covariance specification between the longitudinal outcomes at different quantile levels which is typically the case for multivariate quantile regression model, as noted by Biswas and Das (2021), and Alfo et al. (2021).

3.2 Event-time submodel

One of the major goals of joint modeling is to assess the effects of the biomarkers and the covariates on the time-to-event. Therefore, while modeling time-to-event using the popularly used proportional hazards (PH) model or accelerated failure time (AFT) model, the biomarkers are used as covariates. However, the existing works on joint modeling have shown that the bias in the estimates of the regression coefficients can be reduced by considering the expected biomarker values rather than the observed biomarker values (Henderson 2000, Wang and Taylor 2001, Das 2016, and the references therein).

In our setting the association between the biomarkers and the time-to-event possibly differs from one quantile level to the other (as indicated in Figure 2). Yang et al. (2019), Zhang and Huang (2020) proposed quantile-specific joint models, and we build our model based on their works. We consider a PH model for quantile-specific hazards assuming that the hazard rate at any time point is associated with the estimated biomarker quantiles along with the time-invariant covariates.

We propose two different formulations of the PH model. First, assume that the time-varying covariates (i.e. medicine doses, for example) only affect the biomarker values, and can have only indirect effects on the time-to-event through the biomarkers (Kundu et al. 2024a). This assumption is realistic for many biomedical applications where the event-time is not directly controlled by the medicine doses. However, the medicine doses control the biomarker values which affect the time-to-event. For such situations, our model will be as follows:

$$\lambda_i^{(\boldsymbol{\tau})}(t) = \lambda_0^{(\boldsymbol{\tau})}(t)\text{exp}\left[\Psi^{(\boldsymbol{\tau})T}\boldsymbol{Q}_i^{(\boldsymbol{\tau})}(t) + \boldsymbol{\gamma}^{(\boldsymbol{\tau})T}\mathbf{z}_i\right], \tag{3.4}$$

where $\boldsymbol{Q}_i^{(\boldsymbol{\tau})}(t) = [Q_{i1}^{(\boldsymbol{\tau})}(t), Q_{i2}^{(\boldsymbol{\tau})}(t)]^T$, and $Q_{ik}^{(\boldsymbol{\tau})}(t) = g_k^{(\boldsymbol{\tau})}(t) + \boldsymbol{\beta}_k^{(\boldsymbol{\tau})T}\mathbf{x}_{it} + \boldsymbol{\eta}_k^{(\boldsymbol{\tau})T}\mathbf{z}_i + \omega_{ik}^{(\boldsymbol{\tau})}(t)$; for $k = a, b$. Here λ_0 denotes the baseline hazard, and Ψ and γ, respectively, denote the effects of the biomarker quantiles and the time-invariant covariates on the hazard rate. The baseline hazard function $\lambda_0^{(\boldsymbol{\tau})}(t)$ can be modeled in many different ways. For example, Rizopoulos (2016) used a cubic B-Spline and used the following expression $log(\lambda_0^{(\boldsymbol{\tau})}(t)) = \sum_{q=1}^{Q} \gamma_{0,q}^{(\boldsymbol{\tau})} B_q(t, \nu)$, where $B_q(t, \nu)$ is the q-th basis function of B-splines with knots $\nu_1, \nu_2, \ldots, \nu_Q$ (typically taken as equal percentiles of the event-times). For our illustration, however, we consider a constant baseline hazard for simplicity.

Second, if we assume that the time-varying covariates can directly affect the time-to-event (or hazard rate) as well, then these covariates can be taken as two additional covariates (time-dependent) in the PH model. The hazard rate is now modeled as follows:

$$\lambda_i^{(\boldsymbol{\tau})}(t) = \lambda_0^{(\boldsymbol{\tau})}(t)\text{exp}\left[\Psi^{(\boldsymbol{\tau})T}\boldsymbol{Q}_i^{(\boldsymbol{\tau})}(t) + \boldsymbol{\gamma}^{(\boldsymbol{\tau})T}\mathbf{z}_i + \boldsymbol{\alpha}^{(\boldsymbol{\tau})T}\mathbf{x}_i(t)\right], \tag{3.5}$$

where $\mathbf{x}_i(t) = [x_{i1}(t), x_{i2}(t)]^T$.

For real applications, sometimes, it might not be possible to figure out which of the above two choices will be better. Therefore, we recommend considering both the models given in equations (3.4) and (3.5), and then selecting the one which gives the better fit to the data. Goodness of fit can be assessed by the standard measures, for example, AIC, BIC, LPML etc. On the contrary, if the main objective of the study is prediction then one can train both the models on the data and use a super learner (Naimi and Balzer 2018) for better prediction.

3.3 Joint likelihood and Bayesian inference

We use a Bayesian approach and estimate the regression coefficients based on the joint posterior distribution. We first derive the joint likelihood function as follows.

From longitudinal submodel, considering the mixture representation of ALD, we get the following conditional distributions:
$Y_{ijk}|e_{ijk}, \omega_{ik}^{(\boldsymbol{\tau})}(t_{ij}) \sim N\left(g_k^{(\boldsymbol{\tau})}(t_{ij}) + \boldsymbol{\beta}_k^{(\boldsymbol{\tau})T}\mathbf{x}_{ij} + \boldsymbol{\eta}_k^{(\boldsymbol{\tau})T}\mathbf{z}_i + \omega_{ik}^{(\boldsymbol{\tau})}(t_{ij}) + \theta_{1k}e_{ijk}, \theta_{2k}^2\sigma_k e_{ijk}\right)$,
$e_{ijk}|\sigma_k \sim \exp(\frac{1}{\sigma_k})$.

Let $\boldsymbol{\omega}_i^{(\boldsymbol{\tau})} = \{\omega_{ik}^{(\boldsymbol{\tau})}\}$, $\mathbf{Y} = \{Y_{ijk}\}$, $\mathbf{s} = \{s_i\}$, and $\boldsymbol{\Theta}$ denotes the set of all model parameters (from the longitudinal and the event-time submodels). The joint likelihood is expressed as follows:

$$L(\boldsymbol{\Theta}|\mathbf{Y}, \mathbf{s}, \boldsymbol{\omega}_i^{(\boldsymbol{\tau})}) = \prod_{i=1}^{N}\left[\prod_{j=1}^{n_i}\prod_{k=a,b}\Big(\{f_1(Y_{ijk}|e_{ijk}, \omega_{ik}^{(\boldsymbol{\tau})}(t_{ij})\} \times \{f_2(e_{ijk}|\sigma_k)\}\Big)\times l(\boldsymbol{\omega}_i^{(\tau)})\times l(s_i|\boldsymbol{\Theta})\right], \tag{3.6}$$

where f_1 and f_2, respectively, denote the (conditional) density of $Y_{ijk}|e_{ijk}, \omega_{ik}^{(\boldsymbol{\tau})}(t_{ij})$; and the conditional density of $e_{ijk}|\sigma_k$. Here,

$$\begin{aligned} l(\boldsymbol{\omega}_i^{(\tau)}) = & \frac{1}{\sqrt{2\pi|t_{i1}\boldsymbol{\Sigma}_\tau|}} \times \exp\Big(-\frac{1}{2}\mathbf{w}_{i1}^T(t_{i1}\boldsymbol{\Sigma}_\tau)^{-1}\mathbf{w}_{i1}\Big) \\ & \times \prod_{j=2}^{16}\frac{1}{\sqrt{2\pi|\boldsymbol{\Omega}_{ij}|}} \times \exp\Big(-\frac{1}{2}(\mathbf{w}_{ij} - \mathbf{w}_{i,j-1})^T\boldsymbol{\Omega}_{ij}^{-1}(\mathbf{w}_{ij} - \mathbf{w}_{i,j-1})\Big) \end{aligned}$$

is the likelihood contribution from the random effects, and

$$l(s_i|\mathbf{\Theta}) = \big(\lambda_i^{(\tau)}(s_i)\big)^{I_i} \times \exp\Big(-\int_0^{s_i} \lambda_i^{(\tau)}(t)dt\Big)$$

is the likelihood contribution (for the i-th individual) from the event-time submodel. Here, Ω_{ij} is the 2×2 variance-covariance matrix for dependent Weiner process, i.e. $\mathbf{w}_{ij}$, where $\Omega_{ij} = (t_{ij}^w - t_{i,j-1}^w)\mathbf{\Sigma}_\tau; \quad j = 1, \ldots, M-1$. For our computation, we take M=16 since we did not see any change in the estimates for the values higher than 16.

In the Bayesian setting some prior distributions are assumed for $\mathbf{\Theta}$, and then the joint posterior distribution is derived as follows: $\pi(\mathbf{\Theta}|\mathbf{Y}, \mathbf{s}) \propto L(\mathbf{\Theta}|\mathbf{Y}, \mathbf{s}, \boldsymbol{\omega}_i^{(\tau)}) \times \pi(\mathbf{\Theta})$, where $\pi(\mathbf{\Theta})$ is the joint prior distribution for the set of all model parameters. Assuming independent prior distributions for different model parameters we derive the full conditional distribution for each regression coefficient from the joint posterior distribution, and sample from the full conditional distributions for implementing Markov Chain Monte Carlo (MCMC) algorithm. We use the more recently developed R JAGS which can automatically run the MCMC algorithm without manually sampling from each full conditional distribution. R JAGS also provides the diagnostic checks for the convergence of the chains. We note that joint models are typically implemented using JMBayes software developed by Rizupoulos et al. (2017). However, JMBayes is less flexible and therefore, we use R JAGS and also recommend using it for Bayesian joint modeling.

In many real applications, the longitudinal biomarker values might be missing at some time points. Missingness might happen when the some subject under the study become unavailable at some time points. Joint models can impute missing outcomes effectively (Kundu et al., 2024a), and in our setting we can also impute the missing biomarkers within each MCMC iteration. We simply treat the missing values as unknown parameters, and then keep updating those in each MCMC iteration assuming that the missingness is "ignorable". In particular, we assume "missing at random" (MAR) setting where the missing observations depend only on the observed data points.

In our notation, $\mathbf{\Theta}$ denotes the set of all model parameters. Let $\mathbf{\Theta}^{(m)}$ be the updated $\mathbf{\Theta}$ in the m-th iteration. Based on $\mathbf{\Theta}^{(m)}$, we sample the random effects ω_{ik} for a fixed time point, say t. Next, conditional on ω_{ik}, we sample the missing biomarker values for one specific biomarker conditional on the other biomarker and the model parameters. This will preserve the underlying dependence between the two biomarkers, and we repeat this for each biomarker and for each time point. This gives us a complete dataset with no missing biomarker. We use this complete dataset for updating $\mathbf{\Theta}$ from $\mathbf{\Theta}^{(m)}$ to $\mathbf{\Theta}^{(m+1)}$. By considering M iterations, we obtain M complete datasets based on which we estimate the model parameters. As shown in Kundu et al. (2024a), such a method improves the accuracy of the estimates, and therefore, we recommend this method for each quantile level τ in our setting. However, for our numerical illustrations we do not consider missing biomarker values.

4 Numerical Studies

We perform two simulation studies to illustrate the usefulness of the proposed quantile regression approach to longitudinal joint modeling. In the first simulation study, we illustrate the interpretation of the results from the analysis. In the second study we show that for a skewed distribution (of the biomarkers) quantile-regression joint modeling performs better than the traditionally used linear mixed joint model.

4.1 Simulation study 1

We consider two longitudinal biomarkers, Y_a and Y_b, with two time-varying covariates and six time-invariant covariates. The time-varying covariates, which we refer to as medicines (for illustration only), are denoted by X_1 and X_2. For each subject and for each time point we sample X_1 and X_2 values from a Beta (2,2) and Beta (4,2) distribution, respectively. Time-invariant covariates are denoted by $Z_1, Z_2, \ldots, Z_6$, and those are generated from different probability distributions as summarized and shown in Table 1. We consider 281 subjects for which the biomarkers are measured at different time points. We consider an irregular setting (which is more practical) where the number of measurements are different for different subjects. Total number of measurements from the i-th subject is n_i, and we take $n_i = (I_{V_i=1}\text{Poisson}(3) + I_{V_i=0}\text{Poisson}(10)) + 5$, with $V_i \sim \text{Binomial}(1, 0.3)$. The first time of measurement is same for all subjects, i.e. $t_{i1} = 1$, and the difference in two consecutive time points i.e. $t_{ij} - t_{i,j-1}$ are sampled as: $t_{ij} - t_{i,j-1} \sim \text{Binomial}(1, 0.5) + 4$. The censoring times C_i are sampled as $C_i \sim \text{Binomial}(5, 0.5) \times 4 + 280$, with at least one of the C_i is 300. The longitudinal biomarkers are sampled using equation (3.2) with g as a linear function of time, and for sampling time-to-event we use equation (3.4). We consider five different quantile combinations i.e. $\boldsymbol{\tau} = (0.25, 0.25), (0.25, 0.75), (0.50, 0.50), (0.75, 0.25), (0.75, 0.75)$.

Table 1: Simulation setting for time-invariant covariates in Simulation 1.

Covariate	Distribution
Z_1	Binomial (1,0.7)
Z_2	Binomial (2,0.7) × 0.5
Z_3	Binomial (3,0.3) × 0.5
Z_4	Binomial (4,0.2) × 0.5
Z_5	$N(0, 1)$
Z_6	$N(0, 1)$

4.1.1 Computational Details

We use Bayesian computation for estimating the model parameters, as discussed earlier, and therefore, we consider prior distributions for the model parameters. For each component of $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, $\boldsymbol{\zeta}$ in equation (3.1), and for $\boldsymbol{\gamma}$, we take $N(0, 1000)$ prior distribution. For the covariance matrix Σ_τ, we take Inverse Wishart $(I_3, 2)$ prior, and for σ_k we take Inverse Gammma (0.05,0.05) prior. We select the flat prior distributions typically used in the existing Bayesian literature where the hyper-parameters have minimal effect on the final estimates. We perform a sensitivity analysis with different choices of the hyper-parameters and noticed that the final estimates did not change.

We use MCMC iterations for the parameter estimation. We take 6,000 iterations from two different chains (with different starting values). We discard the first 1,000 iterations as burn-in, and thin the chains by saving every 10-th iteration. This results in 1,000 remaining iterations used for the estimation. Model parameters are estimated by their respective sample means computed from MCMC iterations. The convergence of the chains are assessed by trace plots which we directly get in JAGS. In Figure 3 we show the trace plots for the association parameters (γ) for different quantile levels. These plots show a good convergence of the respective chains. In addition, scale reduction factors (Brooks and Gelman 1998) were also computed by JAGS, and the values are all smaller than 1.1 indicating a good convergence of the chains. Similar results are obtained for the other model parameters (results not shown).

4.1.2 Results and Interpretations

In Figure 4, we show the quantile-specific effects of the six time-invariant covariates for two biomarkers and time-to-relapse. For the biomarker Y_a, we see that across all the five quantile-levels variables 1, 5 and 6 have positive effects. Variable 4 has negative effects for $\boldsymbol{\tau} = (0.25, 0.25)$, $(0.50, 0.50)$, $(0.75, 0.25)$, and no significant effect for the other two levels. Across all five quantile levels, variables 2 and 3 have no effect on Y_a. For the second biomarker Y_b, variable 5 has positive effects across all quantile levels, and variable 6 also has positive effects except for the quantile levels (0.25, 0.75) and (0.75,0.75). Variable 3 has negative effect for $\boldsymbol{\tau} = (0.75, 0.75)$. On the other hand, for the relapse-time, variable 5 has positive effects except for $\boldsymbol{\tau} = (0.25, 0.25)$, variable 1 has positive effects at $\boldsymbol{\tau} = (0.25, 0.75), (0.75, 0.75)$, variable 2 has positive effect for $\boldsymbol{\tau} = (0.25, 0.75)$, variable 3 has negative effects for $\boldsymbol{\tau} = (0.25, 0.25), (0.50, 0.50), (0.75, 0.25)$. We see that none of the covariates has significant effect on two outcomes for all quantile levels, and also there is no covariate with no effect on any outcome at any quantile-level. This illustrates the usefulness of a quantile-regression modeling in the way that it identifies the covariates effect at different quantiles of the data distribution.

Figure 5 shows the quantile-specific effects (and the estimated 95% credible intervals) of two medicines on two biomarkers. We see that for Y_a, across all quantile levels medicine 1 has consistent negative effects, but the medicine 2 has slightly positive effects with the credible intervals containing zero. For the second biomarker Y_b, medicine 1 has mostly insignificant effect except for the level $\boldsymbol{\tau} = (0.50, 0.50)$, and medicine 2 has significantly positive effects for $\boldsymbol{\tau} = (0.25, 0.25), (0.75, 0.25), (0.50, 0.50)$. For the other two levels the estimated effect is close to zero. This figure also indicates that the effects of the medicines change from one quantile level to the other, and therefore, quantile-based modeling is really meaningful. In real data application, such a plot indicates how the dose of the medicines could be adjusted for a higher (or lower) values of certain biomarker.

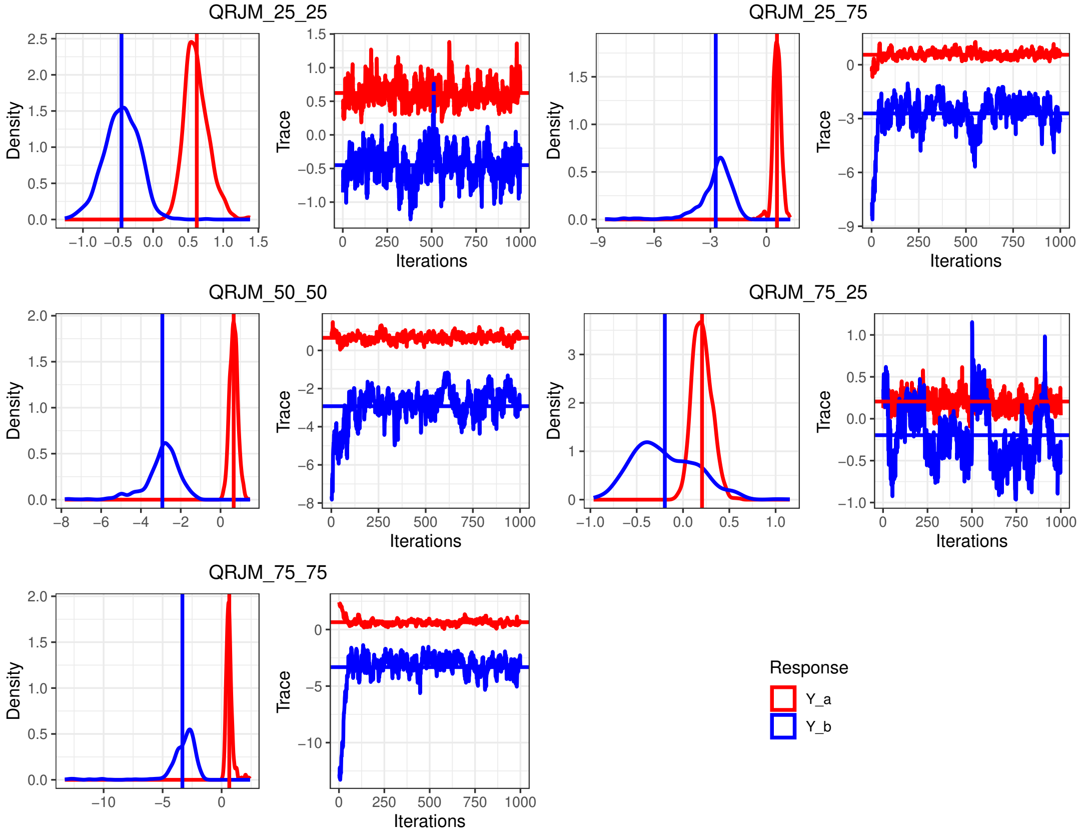

Figure 3: Trace plots for the association parameters in Simulation Study 1.

In Figure 6 the estimated association parameters and the corresponding 95% credible intervals are shown for different quantile-levels. We see that the estimated association parameters for Y_a are slightly positive and the credible intervals do not contain zero. However, for Y_b, the estimates are mostly negative, and the credible intervals do not contain zero except for the level (0.75,0.25). While the estimates for Y_a are mostly similar, those for Y_b differ across quantiles. This figure indicates that a higher values of Y_a increases relapse probability but higher values of Y_b increases the probability of no-relapse. We note that Figure 2 indicated the similar thing which is re-established in Figure 6. For cancer patients, occurrence of a relapse might be affected by some specific biomarkers (for example, platelet count, red blood cell count etc.) and such a plot would be extremely helpful in deciphering the underlying cancer dynamics.

In Figure 7 we show the estimated median non-relapse probabilities for five quantile-levels. Here the covariates are fixed at their respective median values, and the random effects (based on the bivariate Brownian motion) are averaged over all subjects. We see that the survival (non-relapse

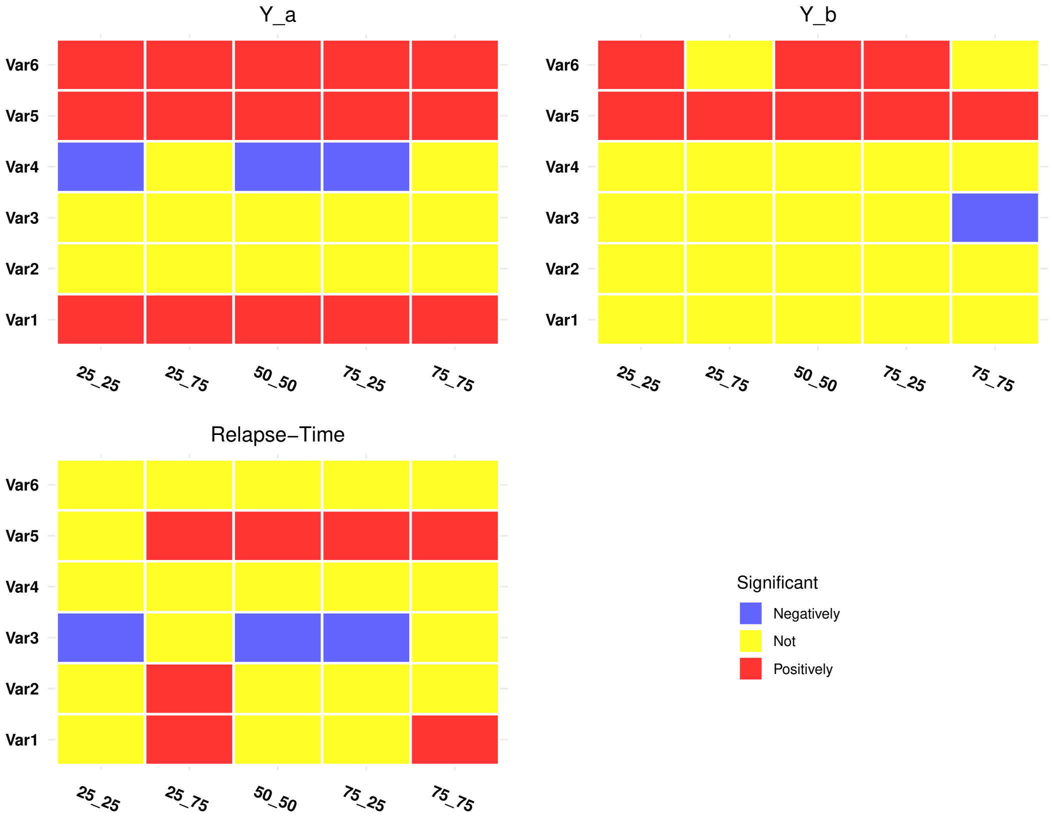

Figure 4: Quantile-specific effects of the fixed covariates in Simulation Study 1.

probability) curves for the levels (0.75,0.75) and (0.25,0.75) are uniformly higher than the curves for the other levels, and the survival probabilities are higher than 0.90. On the other hand, the curve for the level (0.75,0.25) is uniformly lower than all the other curves, and the survival probability goes down to 0.55 in the end. This figure illustrates that irrespective of the quantile level of Y_a, if Y_b is at the higher level then the survival probability is towards the higher side. On the other hand, when Y_b is at lower levels, the survival probability is lower even if Y_a is at a higher level. This indicates that the biomarker Y_b is more important and more informative for a relapse. Detection of such a biomarker is very important in a real biomedical application.

Finally, in Figure 8 we show the estimated quantiles for Y_a and Y_b at levels 0.25, 0.50 and 0.75. It is seen that the estimated quantile values for a higher quantile level is indeed higher, and hence there is no quantile cross in our case. This illustrates that our model results in meaningful estimates of the quantiles and therefore can be used in real application.

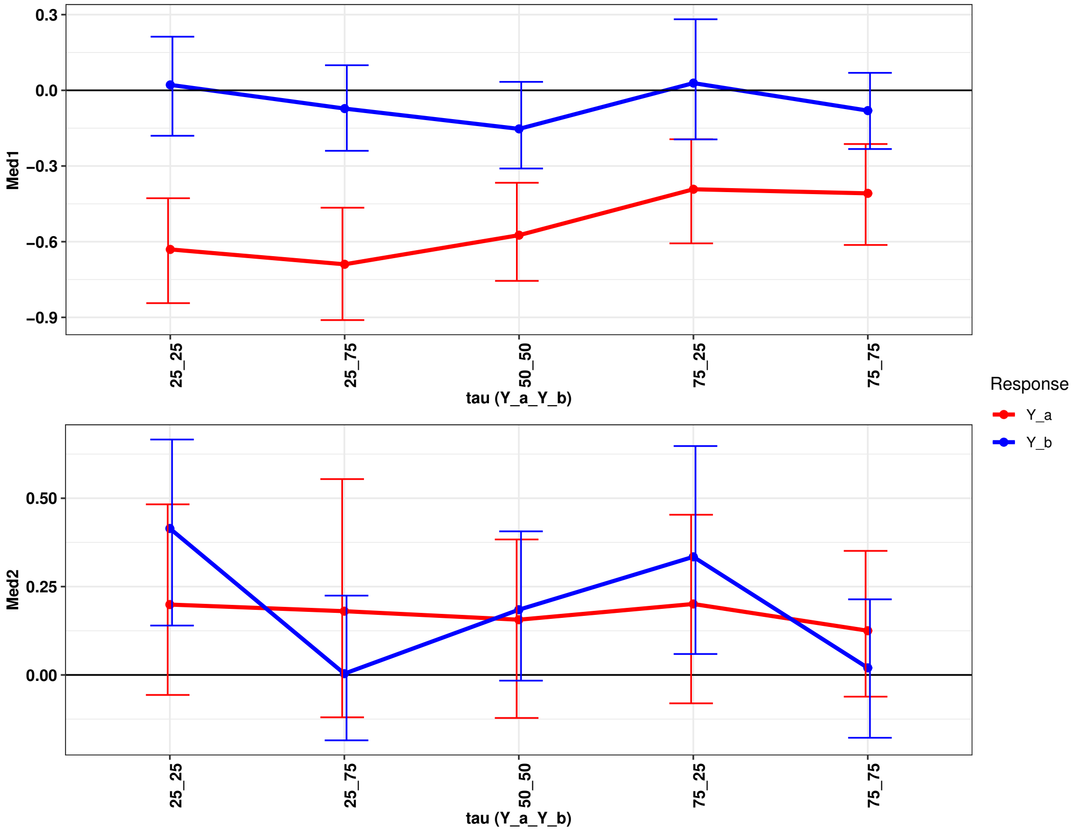

Figure 5: Quantile-specific effects of the two medicines and the respective 95% credible intervals in Simulation Study 1.

4.2 Simulation study 2

We perform a second simulation study for assessing the effectiveness of the quantile regression joint model (QRJM) over the linear mixed joint model (LMJM) when the biomarkers are generated from a skewed distribution. We simulate two longitudinal biomarkers (Y_{ijk}) from the model given in equation (3.2), without g (the general effect of time). We consider two time-varying covariates, i.e. $\mathbf{x}_{ij} = [x_{ij1}, x_{ij2}]$, and four fixed covariates, i.e. $\mathbf{z}_i = [z_{i1}, z_{i2}, z_{i3}, z_{i4}]$. All the covariates are generated from a standard normal distribution, and the subject-specific random effects are generated from a bivariate normal distribution with mean vector=0, and the covariance matrix=Σ. The variance components in the matrix Σ are 1.5, 2.5; and the correlation between the two biomarkers is 0.65. The random errors are generated from a standard normal distribution and from an asymmetric Laplace distribution with twp different skewness parameters, i.e. $\tau_k = 0.25$ and $\tau_k = 0.50$.

For generating time-to-event we use the model given in equation (3.4), with a constant baseline

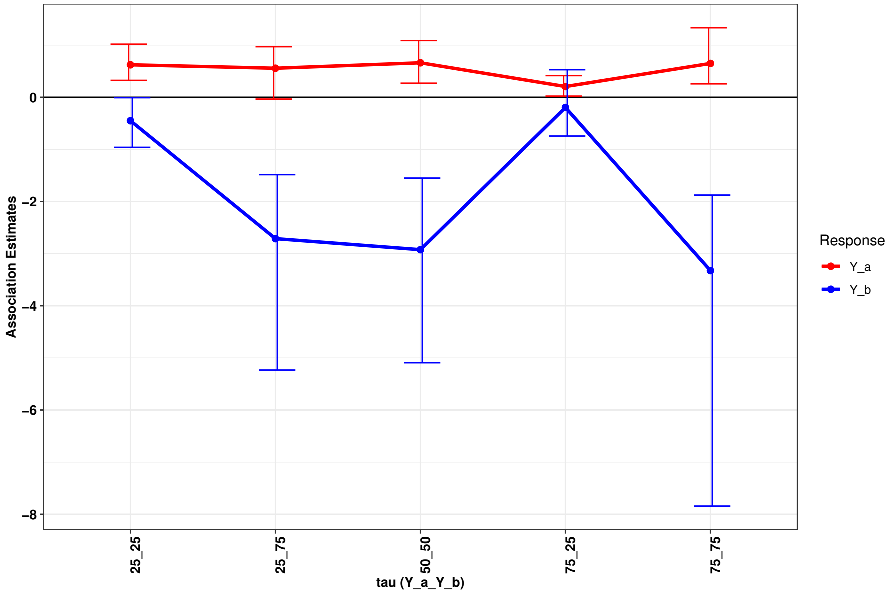

Figure 6: Quantile-specific association parameters and the respective 95% credible intervals for Simulation Study 1.

hazard. We consider fifteen longitudinal measurements for each subject, and then subjects are followed for the next fifteen time points. At T=30, subjects are all censored. We consider the following three cases.

Case I: Random errors are generated from ALD with τ_k=0.25, $k = 1, 2$ (right-skewed distribution).
Case II: Random errors are generated from ALD with τ_k=0.50, $k = 1, 2$ (symmetric distribution with heavy tails).
Case III: Random errors are generated from a standard normal distribution (symmetric distribution).

For each of the above three cases we generate 200 datasets, and for each dataset we consider 100 subjects. We fit both QRJM and LMJM, and use MCMC for parameter estimation. We note that for LMJM, we use the model in equation (3.2) without $\boldsymbol{\tau}$, and the errors ϵ_{ijk} are independent $N(0, \sigma^2)$ random variables, and for equation (3.4) we replace $\boldsymbol{Q}_i^{(\boldsymbol{\tau})}(t)$ by $E(Y_{ijk})$ obtained from equation (3.2), and there is no $\boldsymbol{\tau}$ here as well. In other words, LMJM considers the effects of the mean outcomes on time-to-event, and also aims to assess the effects of the covariates on the mean biomarkers.

For evaluating the discriminative power of a model we compute the area under the receiver operating characteristic curve (AUC) for different models. The AUC measures how effectively a joint model discriminates the subjects for which a relapse occurs from the subjects with no relapse (Rizopoulos 2016, Kundu et al. 2024b). Let $\pi_i(t + \Delta t|t)$ be the probability that for the i-th subject there is no relapse upto time $t + \Delta t$ given that it is event-free (no relapse) until time t. For any pair

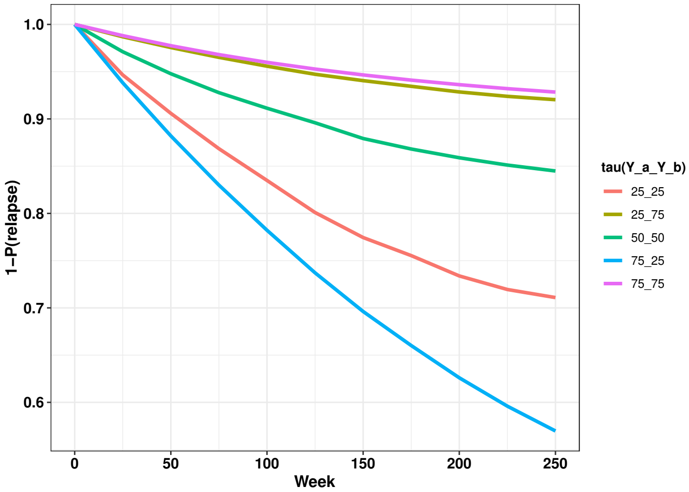

Figure 7: Median estimated non-relapse probability curves for different quantile combinations.

of subjects $[i, j]$ who are event-free until time t, the discriminative power of a model is assessed by: $AUC = P\left[\pi_i(t+\Delta t|t) < \pi_j(t+\Delta t|t) | (T_i \in (t, t+\Delta t]) \cap (T_j > t+\Delta t)\right]$, where T_i and T_j, respectively, denote the actual event-time for the i-th and the j-th subject. This means that for a fixed time-interval $(t, t+\Delta t]$ if a relapse occurs for the i-th subject but the j-th subject is event-free upto time $t+\Delta t$, then the model must assign a higher non-relapse probability to the j-th subject.

Table 2: AUC values for different models under different settings in the Simulation Study (values are rounded upto two decimal places).

				Predicted AUC(t, Δt)		
Data Distribution	t	Δt	True AUC(t, Δt)	QRJM(τ=(25,25)	QRJM(τ=(50,50)	LMJM
ALD(τ_k=0.25)	15	5	0.85	0.83	0.82	0.75
		8	0.88	0.87	0.85	0.77
		10	0.92	0.91	0.89	0.81
ALD(τ_k=0.50)	15	5	0.84	0.83	0.85	0.80
		8	0.89	0.88	0.89	0.84
		10	0.90	0.89	0.88	0.85
$N(0,1)$	15	5	0.83	0.80	0.83	0.84
		8	0.87	0.84	0.86	0.87
		10	0.91	0.87	0.89	0.89

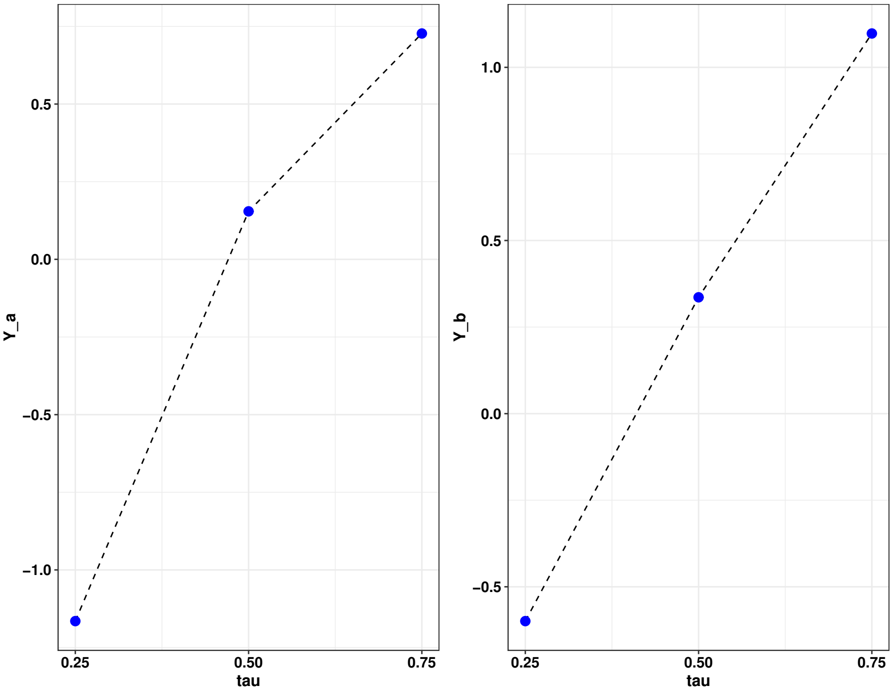

Figure 8: Estimated quantiles for the two biomarkers in Simulation Study 1.

We consider t=15, and for three different values of Δt (i.e. Δt=5,8,10) we compute the true AUC (based on the simulated dataset) and the predicted AUC for three different models. Results (average AUC values from 200 datasets) are summarized in Table 2. We note that for data generated from ALD with τ_k=0.25, QRJM provides a better prediction (i.e. higher AUC values) than LMJM. For data generated from ALD with τ_k=0.50, we observe the similar results. For data generated from a standard normal distribution, the predictive power of LMJM and a QRJM with $\boldsymbol{\tau}$=(0.50,0.50) give similar results. Thus, in general, we notice that QRJM provides a better prediction than LMJM.

We note that for assessing the performance of a prediction model Brier scores are popularly use. Brier score provides the overall model performance. In the joint modeling of longitudinal outcomes and time-to-event it is of interest to evaluate models based on the discriminative power. Therefore, AUC has been used popularly in the joint modeling literature. However, we note that for obtaining a better predictive power Rizopoulos and Taylor (2024) proposed a super learner approach where several models are combined altogether and the resulting super learner provides a better prediction than any of the individual model. These authors used Integrated Brier Scores (IBS) for assessing the models under investigation. We exclude these methods in this work for the ease of presentation and also because we consider simulated datasets only, but for a real application Brier Scores can be extremely helpful.

5 Discussion

Joint modeling is a very useful tool in biomedical research, and it can provide scientifically meaningful statistical inferences in an effective way. However, the traditional linear mixed model based approach fails to provide efficient inference when the outcomes of interest come from a non-Gaussian and/or a skewed distribution. In this article, we illustrate a Bayesian quantile-regression based joint model which can handle such datasets. Our numerical studies have shown the practical usefulness and merits of this approach over the linear mixed joint models used commonly in the existing literature. We are quite confident that in many real applications the proposed model can be used for powerful statistical inferences. Also, our model can automatically handle missing biomarker values without much computational difficulty.

There are, however, certain limitations of the proposed model and these limitations are quite common for Bayesian quantile regression models in general. First, we assume an Asymmetric Laplace Distribution with some specific scale parameter for the given dataset for modeling one particular quantile level. For modeling a different quantile level, we assume a different scale parameter for the same dataset, and this results in a lack of compatibility. In addition, although we do not come across a quantile cross in our numerical illustrations, there is no theoretical justification which can state that one would not come across such a problem in our modeling approach. Also quantile regression models are typically problematic in higher dimensions. All these typical issues of a quantile regression model are there in our proposed QRJM, unfortunately. However, as illustrated in this article, QRJM is able to handle non-Gaussian data in a more effective way than LMJM.

References

Alfò, M., Marino, M.F., Ranalli, M.G., Salvati, N., & Tzavidis, N. (2021), “M-quantile regression for multivariate longitudinal data with an application to the Millennium Cohort Study,” *Journal of the Royal Statistical Society Series C: Applied Statistics*, 70, 122-146.

Biswas, J., & Das, K. (2021), “A Bayesian quantile regression approach to multivariate semi-continuous longitudinal data,” *Computational Statistics*, 36, 241-260.

Chi, Y.Y., & J.G. Ibrahim. (2006), “ Joint models for multivariate longitudinal and multivariate survival data,” *Biometrics*, 62, 432-445.

Das, K., Li, R., Huang, Z., Gai, J., & Wu, R. (2012), “A Bayesian framework for functional mapping through joint modeling of longitudinal and time-to-event data,” *International Journal of Plant Genomics*, doi:10.1155/2012/680634.

Das, K. (2016), “ A semiparametric Bayesian approach for joint modeling of longitudinal trait and event time,” *Journal of Applied Statistics*, 43, 2850-2865.

Daniels, M.J., & Hogan, J.W. (2008), *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*, CRC press.

Fieuws, S., & Verbeke, G. (2004), “Joint modelling of multivariate longitudinal profiles: Pitfalls of the random-effect approach,” *Statistics in Medicine*, 23, 3093-3104.

Geraci, M., & Bottai, M. (2007), “ Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, 8, 140-154.

Guo, X., & Carlin, B.P. (2004), “Separate and Joint Modeling of Longitudinal and Event Time Data Using Standard Computer Packages,” *The American Statistician*, 58, 16-24.

Henderson, R., Diggle, P., & Dobson, A. (2000), “Joint modelling of longitudinal measurements and event time data,” *Biostatistics*, 1, 465-480.

Koenker, R., & Bassett, Jr G. (1978), “Regression quantiles,” *Econometrica: Journal of the Econometric Society*, 46, 33-50.

Koenker, R. (2004), “Quantile regression for longitudinal data,” *Journal of Multivariate Analysis*, 91, 74-89.

Kozumi, H., & Kobayashi, G. (2011), “Gibbs sampling methods for Bayesian quantile regression,” *Journal of Statistical Computation and Simulation*, 81, 1565-1578.

Kulkarni, H., Biswas, J., & Das, K. (2019), “A joint quantile regression model for multiple longitudinal outcomes. *Advances in Statistical Analysis*, 103, 453–473.

Kundu, D., Sarkar, P., Gogoi, M., & Das, K. (2024a), “A Bayesian joint model for multivariate longitudinal and time-to-event data with application to ALL maintenance studies,” *Journal of Biopharmaceutical Statistics*, 34, 37-54.

Kundu, D., Shekhar, K., Gogoi, M., & Das, K. (2024b), “A Bayesian quantile joint modeling of multivariate longitudinal and time-to-event data,” *Lifetime Data Analysis* (in press).

Naimi, A.I., & Balzer, L.B. (2018), “Stacked generalization: an introduction to super learning,” *European Journal of Epidemiology*, 33, 459-464.

Rizopoulos, D., & Ghosh, P. (2011), “ A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine*, 30, 1366-1380.

Rizopoulos, D. (2016), “The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. *Journal of Statistical Software*, 72, 1-45.

Rizopoulos, D., Molenberghs, G., & Lesaffre, E. (2017), “Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking,” *Biometrical Journal*, 59, 1261-1276.

Rizopoulos, D., & Taylor, J.M.G. (2024), “Optimizing dynamic predictions from joint models using super learning,” *Statistics in Medicine*, 43, 1315-1328.

Wang, Y., & Taylor, J.M.G. (2001), “Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome," *Journal of the American Statistical Association*, 96, 895-905.

Yang, M., Luo, S., & DeSantis, S. (2019), "Bayesian quantile regression joint models: inference and dynamic predictions," *Statistical Methods in Medical Research*, 28, 2524-2537.

Yu, K., & Moyeed, R.A. (2001), "Bayesian quantile regression," *Statistics* & *Probability Letters*, 54, 437-447.

Zhang, H., & Huang, Y. (2020), " Quantile regression-based Bayesian joint modeling analysis of longitudinal–survival data, with application to an AIDS cohort study," *Lifetime Data Analysis*, 26, 339-368.

Received: February 7, 2024

Accepted: May 12, 2024