BAYESIAN QUANTILE REGRESSION JOINT MODELS: INFERENCE AND DYNAMIC PREDICTIONS

by

MING YANG, MS

APPROVED:

STACIA DESANTIS, PhD

SHENG LUO, PhD

DAVID LAIRSON, PhD

XIAOMING LIU, PhD

DEAN, THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH

DEDICATION

To my family.

BAYESIAN QUANTILE REGRESSION JOINT MODELS: INFERENCE AND DYNAMIC PREDICTIONS

by

MING YANG,
BS, Xiamen University, 2008
MS, The University of Texas School of Public Health, 2012

Presented to the Faculty of The University of Texas
School of Public Health
in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS
SCHOOL OF PULIC HEALTH
Houston, Texas
December, 2016

ACKNOWLEDGEMENTS

I would like to gratefully and sincerely thank my dissertation co-advisors Dr. Stacia DeSantis and Dr. Sheng Luo. It would not have been possible to write this dissertation without their ideas, patience and guidance. They have devoted their precious time to serve as my advisors and I appreciate for all their helpful discussion and comments throughout our countless meetings and email communications. I would like to thank my minor advisor Dr. David Lairson and breadth advisor Dr. Xiaoming Liu for their support and inputs to my dissertation. Thanks also go to my external reviewer Dr. Soeun Kim for her consistent support and precious time towards my requests during this work. Last but not least, I would like to thank my wife Yalan, for her encouragement, quiet patience and unwavering love. Her tolerance of my occasional vulgar moods is a testament in itself of her unyielding devotion and love.

BAYESIAN QUANTILE REGRESSION JOINT MODELS: INFERENCE AND DYNAMIC PREDICTIONS

Ming Yang, PhD
The University of Texas
School of Public Health, 2016

Dissertation Chair, Stacia DeSantis, PhD

In the traditional joint models (JM) of a longitudinal and time-to-event data, a linear mixed model (LMM) assuming normal random error is frequently used to model the longitudinal continuous outcome. However, in many circumstances, the normality assumption cannot be satisfied and LMM is not appropriate to use. In addition, as a mean regression based methods, LMM only models the conditional mean of the longitudinal outcome, thus its application is limited when clinical interest lies in making inference or predictions on median, lower, or upper ends of the outcome variable. In contrast, quantile regression (QR) models provide a more flexible, distribution-free way to study covariate effects at different conditional quantiles of the outcome and it is robust against deviations from normality as well as outlying observations. In addition, the JM framework provides a convenient way to make subject-specific predictions of event probability. However, current predictive algorithms are all based on the traditional JM that uses LMM. In the first paper, we proposed a new version of JM that adopts a linear quantile mixed model (LQMM) for the longitudinal process and we named it quantile regression joint models (QRJM). We developed a Gibbs sampling algorithm based on the location-scale representation of the asymmetric Laplace distribution, assessed its performance through extensive simulation studies, and demonstrated how the

QRJM approach can be used for making subject-specific dynamic predictions of the risk of Huntington's disease onset. As another type of time-to-event outcome, recurrent events are commonly encountered in longitudinal biomedical studies. In contrast to survival data, multiple event times are observed in a single subject during the study follow-up. In the second paper, we extended the proposed QRJM in paper 1 to joint analysis of longitudinal and recurrent event data and developed a fully Bayesian algorithm for model inference. In the third paper, we developed a subject-specific dynamic prediction algorithm for recurrent event probability based on the QRJM proposed in paper 2. We conducted extensive simulation studies to validate the proposed algorithm in inference (paper 2) and to quantify its predictive performance (paper 3). In the data applications of paper 2 and 3, we illustrated the flexibility of the QRJM and its advantages over the traditional JM by jointly modeling the risk of coronary heart disease (CHD) recurrences and longitudinal systolic blood pressure (SBP) measurements (paper 2) and by making predictions of the risk of CHD recurrences (paper 3). QRJM was able to provide more insight into the disease progression and the association between the two disease processes in terms of various quantile-based estimations and dynamic predictions.

Contents

List of Tables

List of Figures

1 Background

1.1 Literature Review

1.1.1 Joint Models for Longitudinal And Time-to-Event Data

In many longitudinal biomedical studies, time-to-event outcome is always of clinical interest as it indicates health condition and disease progression of the patients. Meanwhile, some continuous outcome(s) is often collected repeatedly during the study follow-up or at the time of events as the disease biomarker(s). Study interest often lies in modeling both outcomes as well as the relationship between them. A familiar example is the study of the association between repeatedly measured blood pressure and the risk of cardiovascular diseases or death. Ignoring the dependence between the two processes and fit models for them separately will lead to loss of information and result in biased or inefficient inference results (Tsiatis and Davidian, 2004). In addition, traditional survival model with time-varying covariate may not be appropriate to use due to its limiting assumption of external time dependent covariates that are not related to the event mechanism. This is especially true when we are interested in making predictions of the time-to-event outcome in the future, when the longitudinal biomarker is impossible to observe. Joint models (JM) of longitudinal and time-to-event data are more suitable to use under such situation as they are highly applicable in the setting of survival analysis with a time dependent covariate measured with error, or for longitudinal data analysis with event-related dropouts(Self and Pawitan, 1992; Tsiatis et al., 1995). By accommodating such joint processes, the simultaneous covariate effects on the repeated instances of the longitudinal sequence, but also across sequences, can be modeled and examined. Moreover, association between the two processes can be made an integral part of the models.

Joint analysis of longitudinal and time-to-event outcomes has been studied by many authors since the seminal work of Self and Pawitan (1992); Tsiatis et al. (1995). Wulfsohn and Tsiatis (1997) developed the EM algorithm for parameter estimation in JM for survival analysis with a time dependent covariate measured with error. Henderson et al. (2000) introduced the shared random effects JM for longitudinal and time-to-event data. As a summary paper, Tsiatis and Davidian (2004) gives an excellent review of the JM methods. In recent years, many extensions have been added to the original version of JM, including considering multiple longitudinal outcomes (Brown et al., 2005; Rizopoulos and Ghosh, 2011), incorporating multiple failure times (Elashoff et al., 2008), among others. Rizopoulos (2011); Taylor et al. (2013) introduced the novel idea of making subject-specific dynamic predictions of future event-free probability based on the JM framework of longitudinal continuous outcome and survival. Farcomeni and Viviani (2015) first considered using longitudinal quantile regression model for the repeated continuous outcome in the JM. Compared with the extensive work on single time-to-event outcome, joint analysis of longitudinal and repeated time-to-event (or recurrent event) data has received less attention so far. To our knowledge, Henderson et al. (2000) developed a JM for longitudinal data and recurrent event outcome. Kim et al. (2012) considered a JM of longitudinal and recurrent event data with informative terminal event. Efendi et al. (2013) proposed a JM of longitudinal data and recurrent events that accommodates overdispersion.

1.1.2 Quantile Regression And Linear Quantile Mixed Model

Within the traditional JM framework, a linear mixed model (LMM) is frequently used to model the longitudinal continuous outcome. Under LMM measurements from the same subject share the same random effects to account for within-subject correlation and random errors are assumed to be normally distributed and independent with the random effects(Laird and Ware, 1982). However, if the normality assumption of the error term is violated (even after applying various outcome transformations), an LMM is not appropriate to use. Furthermore, LMM only models covariate effects on the conditional mean of the longitudinal outcome. However, in many clinical studies it is more desirable to make inference on the median, lower or higher conditional quantiles of the outcome. For example, low birth weight is known to be closely associated with infant mortality and development of chronic diseases. Koenker and Hallock (2001) studied the impact of various health factors on low birth weight using quantile regression (QR) model. They found several covariate effects on lower birth weight were significantly different compared with their effects on the mean birth weight. Thus, by modeling the conditional quantile of the outcome, QR provides a much more comprehensive way of examining the association between covariates and outcome (Koenker, 2005). Specifically, estimators of regression coefficients are functions of the quantile in QR, and we are able to study the heterogeneity of the outcome that is associated with the covariates. Moreover, as QR doesn't impose any distribution assumption on the data, it is more robust against the deviation from normality as well as outliers in the data.

Since the seminal work of Koenker and Basset (1978), QR is attracting increasing interest in the statistical community. QR has also been extended to longitudinal analysis by many authors. For example, Jung (1996) developed a quasi-likelihood method of median regression model for longitudinal data. Geraci and Bottai (2007) proposed to fit QR for longitudinal data using asymmetric Laplace distribution (ALD) and parameter estimation is made by Monte Carlo EM algorithm. Fu and Wang (2012) proposed a working correlation QR model for longitudinal data. From Bayesian perspective, (Kozumi and Kobayashi, 2011; Luo et al., 2012) developed a Gibbs sampling algorithm for linear quantile mixed model (LQMM) by assuming ALD for the error term.

Let $Y_i(t_{ij})$ be the longitudinal outcome for subject i measured at time t_{ij} where $i = 1, \cdots, N$

and $j = 1, \cdots, n_i$. Consider the following LMM:

$$Y_i(t) = \boldsymbol{X}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i + \varepsilon_i(t), \varepsilon_i(t) \sim N(0, \sigma^2), \tag{1.1}$$

where $\boldsymbol{\beta}$ is a $p-$dimensional fixed effects, $\boldsymbol{X}_i(t)$ contains the corresponding fixed-effect covariates, $\boldsymbol{u}_i$ is a $k-$dimensional random effects for subject i, and $\boldsymbol{Z}_i(t)$ contains the corresponding random-effect covariates.

A linear quantile mixed model (LQMM) assumes that the conditional quantile of the outcome is a linear function of the covariates,

$$Q_{Y_i(t)|\boldsymbol{X}_i(t),\boldsymbol{Z}_i(t)}(\tau) = \boldsymbol{X}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i, \tag{1.2}$$

where the τth quantile of a random variable Y is defined as $Q_Y(\tau) = F_Y^{-1}(\tau) = \inf\{y : F_Y(y) \geq \tau\}$ for $\tau \in [0, 1]$. The quantile regression estimates $\hat{\boldsymbol{\beta}}_\tau$ can be obtained by minimizing the loss function $\sum_{i,t}\left[\rho_\tau\left(Y_i(t) - \boldsymbol{X}_i^\top(t)\boldsymbol{\beta} - \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i\right)\right]$, where $\rho_\tau(\cdot)$ is defined as $\rho_\tau(Y) = Y(\tau - I(Y < 0))$. Parameters $\boldsymbol{\beta}_\tau$ are functions of quantile τ, as denoted by the subscript.

As discussed in Koenker and Machado (1999) and Yu and Moyeed (2001), the above minimization problem can be rephrased as a maximum-likelihood problem by assuming that the random error $\varepsilon_i(t)$ in (1.1) follows ALD, denoted by $ALD(0, \sigma, \tau)$, with location parameter equals 0, scale parameter $\sigma > 0$ and skewness parameter $\tau \in (0, 1)$. Parameter τ controls the skewness of the ALD. For example, it is skewed to left when $\tau > 0.5$, skewed to right when $\tau < 0.5$, and symmetric when $\tau = 0.5$. Adopting the ALD, an LQMM is written as $Y_i(t) = \boldsymbol{X}_i^\top(t)\boldsymbol{\beta}_\tau + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i + \varepsilon_i(t), \varepsilon_i(t) \sim ALD(0, \sigma, \tau)$. The conditional likelihood function of the longitudinal outcome is then given by $\ell(Y_i(t)|\boldsymbol{\beta}_\tau, \boldsymbol{u}_i, \sigma) = \frac{\tau(1-\tau)}{\sigma}\exp\left[-\rho_\tau\left(\frac{Y_i(t)-\boldsymbol{X}_i^\top(t)\boldsymbol{\beta}_\tau-\boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i}{\sigma}\right)\right]$.

In Bayesian quantile regression context a Gibbs sampling algorithm for model inference is developed when we utilize a location-scale mixture representation of the ALD (Kotz et al., 2012). Under such parameterization the random error is represented as $\varepsilon_i(t) = \kappa_1 e_i(t) + \kappa_2\sqrt{\sigma e_i(t)}v_i(t)$ with $v_i(t) \sim \mathcal{N}(0,1), e_i(t) \sim \exp(1/\sigma)$ and

$$\kappa_1 = \frac{1-2\tau}{\tau(1-\tau)} \qquad \kappa_2^2 = \frac{2}{\tau(1-\tau)}.$$

This re-parameterization leads to the following linear mixed model,

$$Y_i(t) = \boldsymbol{X}_i^\top(t)\boldsymbol{\beta}_\tau + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i + \kappa_1 e_i(t) + \kappa_2\sqrt{\sigma e_i(t)}v_i(t),$$

or equivalently,

$$\ell(Y_i(t)|\boldsymbol{\beta}_\tau, \boldsymbol{u}_i, e_i(t), \sigma) = \frac{1}{\sqrt{2\pi\kappa_2^2\sigma e_i(t)}} \exp\left[-\frac{(Y_i(t) - \boldsymbol{X}_i^\top(t)\boldsymbol{\beta}_\tau - \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i - \kappa_1 e_i(t))^2}{2\kappa_2^2\sigma e_i(t)}\right]. \tag{1.3}$$

As discussed in Yu and Moyeed (2001), irrespective of the actual distribution of the data, Bayesian quantile regression using ALD distribution works quite well for different error distributions and the performance is quite robust and satisfactory.

1.2 Proposed Statistical Methods

1.2.1 Quantile Regression Joint Models

In this work, we propose a new Bayesian quantile regression joint models (QRJM) framework that replace the LMM with an LQMM for the longitudinal process in the traditional linear model JM. For the time-to-event outcome, our QRJM works for both single event time data as well recurrent events.

For single time-to-event (e.g. death) outcome, let $T_i = \min(T_i^*, C_i)$ be the observed event time for subject i, where T_i^* is the true underlying event time and C_i is the censoring time. Let Δ_i be the event indicator defined as $\Delta_i = I(T_i^* < C_i)$, where $I(\cdot)$ is the indicator function. If $\Delta_i = 1$, i.e. $T_i^* < C_i$, we say an event is observed during the study period; in contrast, $\Delta_i = 0$ when there is no event observed at the end of the study or the patient is lost to follow-up (i.e., censored). Let $Y_i(t)$ be the continuous longitudinal outcome for subject i measured at time t. Note that we can only observe $Y_i(t)$ when $t \leq C_i$, so the complete longitudinal measurements for subject i can be written as $\mathcal{Y}_i(t) = \{Y_i(s) : 0 \leq s \leq t\}$. We denote the true underlying longitudinal measurement for subject i at time t with $m_i(t)$ and his/her complete history of true longitudinal trajectory as $\mathcal{M}_i(t) = \{m_i(s) : 0 \leq s \leq t\}$. The proposed QRJM for terminal event can be written as follows:

$$\begin{cases} Y_i(t) = m_i(t) + \varepsilon_i(t) = \boldsymbol{X}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i + \varepsilon_i(t), \varepsilon_i(t) \overset{i.i.d.}{\sim} ALD(0, \sigma, \tau) \\ h(T_i|\mathcal{M}_i(T_i), \boldsymbol{W}_i; \boldsymbol{\gamma}, \alpha) = h_0(T_i)\exp(\boldsymbol{W}_i^\top\boldsymbol{\gamma} + \alpha m_i(T_i)) \end{cases}$$

where the first equation is the LQMM introduced in Section 1.1.2. Here $\boldsymbol{X}_i(t)$ are the fixed effect covariates and $\boldsymbol{Z}_i(t)$ are the covariates associated with $k-$dimensional random effects $\boldsymbol{u}_i$. Individual heterogeneity is captured by $\boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i$, which is the deviation of subject i from the population. The second equation takes the format of Cox proportional hazards model (PHM) where $h_0(\cdot)$ is the baseline hazard function and $\boldsymbol{W}_i$ are the $q-$dimensional fixed effect covariates only associated with event time (not the longitudinal outcome). These two models are linked by treating the τth conditional quantile of the longitudinal outcome as a time dependent covariate in the time-to-event process, and the degree of associations is measured by parameter α.

For subject i, the likelihood function for survival data is:

$$\ell(T_i, \Delta_i|\boldsymbol{u_i}) = h(T_i|\mathcal{M}_i(T_i), \boldsymbol{W}_i)^{\Delta_i} S(T_i|\mathcal{M}_i(T_i), \boldsymbol{W}_i), \tag{1.4}$$

where $S(\cdot)$ is the survival function,

$$S(T_i|\mathcal{M}_i(T_i), \boldsymbol{W}_i) = \exp\left\{-\int_0^{T_i} h_0(s)\exp(\boldsymbol{W}_i{}^\top \boldsymbol{\gamma}_\tau + \alpha(\boldsymbol{X}_i^\top(s)\boldsymbol{\beta}_\tau + \boldsymbol{Z}_i^\top(s)\boldsymbol{u}_i))ds\right\}.$$

Similarly, for longitudinal and recurrent event data the proposed QRJM is given by:

$$\begin{cases} Y_i(t) = m_i(t) + \varepsilon_i(t) = \boldsymbol{X}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i + \varepsilon_i(t), \varepsilon_i(t) \overset{i.i.d.}{\sim} ALD(0, \sigma, \tau) \\ r_i(T_{ik}|\mathcal{M}_i(T_{ik}), \boldsymbol{W}_i; \boldsymbol{\gamma}, \alpha) = r_{i0}(T_{ik})\exp(\boldsymbol{W}_i{}^\top \boldsymbol{\gamma} + \alpha m_i(T_{ik})) \end{cases}$$

where $r_{i0}(\cdot)$ is the baseline intensity function and $T_{ik} = \min(T_{ik}^*, C_i)$, $k = 1, \cdots, m_i$, is the kth event time for subject i, assuming a total number of m_i events are observed within the censoring time C_i. The likelihood function for recurrent event data can be written as:

$$\begin{aligned} \ell(\boldsymbol{T}_i, \boldsymbol{\Delta}_i; \boldsymbol{\theta}) &= \prod_{k=1}^{m_i}\left[r_i(T_{ik}; \boldsymbol{\theta}|\mathcal{M}_i(T_{ik}), \boldsymbol{W}_i)^{\Delta_{ik}}\exp\left(-\int_{T_{ik-1}}^{T_{ik}} r_i(s; \boldsymbol{\theta}|\mathcal{M}_i(s), \boldsymbol{W}_i)ds\right)\right] \\ &= \prod_{k=1}^{m_i}\left[r_i(T_{ik}; \boldsymbol{\theta}|\mathcal{M}_i(T_{ik}), \boldsymbol{W}_i)^{\Delta_{ik}}\right]\exp\left(-\int_0^{T_{im_i}} r_i(s; \boldsymbol{\theta}|\mathcal{M}_i(s), \boldsymbol{W}_i)ds\right), \end{aligned} \tag{1.5}$$

and $\Delta_{ik} = I(T_{ik} < C_i)$ is the event indicator.

1.2.2 Bayesian Inference

Let O_i be the maximum of follow-up time for subject i, that is $O_i = T_i$ for terminal event outcome and $O_i = C_i$ for recurrent event outcome according to our previous notation. In general, the complete likelihood function of the bivariate outcomes takes the form as follows:

$$L_i(\boldsymbol{\theta}; \boldsymbol{T}_i, \boldsymbol{\Delta}_i, \mathcal{Y}_i(O_i), \boldsymbol{u}_i) = \ell(\mathcal{Y}_i(O_i); \boldsymbol{\theta}|\boldsymbol{u}_i)\ell(\boldsymbol{T}_i, \boldsymbol{\Delta}_i; \boldsymbol{\theta}|\boldsymbol{u}_i) f(\boldsymbol{u}_i|\boldsymbol{\Sigma}), \tag{1.6}$$

where vector $\boldsymbol{\theta}$ represents a set of all the parameters from each distribution function, $\ell(\mathcal{Y}_i(O_i); \boldsymbol{\theta}|\boldsymbol{u}_i) = \prod_{0 \le t \le O_i} \ell(Y_i(t); \boldsymbol{\theta}|\boldsymbol{u}_i)$, where $\ell(Y_i(t), \boldsymbol{\theta}|\boldsymbol{u}_i)$ takes the format of (1.3), and $\ell(\boldsymbol{T}_i, \boldsymbol{\Delta}_i; \boldsymbol{\theta}|\boldsymbol{u}_i)$ is given in (1.4) for terminal event and in (1.5) for recurrent event outcome, respectively. We propose a fully Bayesian inference algorithm for unknown parameters by taking advantage of the location-scale mixture representation of the ALD that is described in Section 1.1.2. According to the Bayes theorem, the posterior distributions of the model parameters are proportional to the product of likelihood function and prior:

$$f(\boldsymbol{\theta}|\boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{\mathcal{Y}}, \boldsymbol{u}) \propto \prod_{i=1}^{N} L_i(\boldsymbol{\theta}; \boldsymbol{T}_i, \boldsymbol{\Delta}_i, \mathcal{Y}_i(O_i), \boldsymbol{u}_i) f(\boldsymbol{\theta}), \tag{1.7}$$

where N is the total number subjects, $\boldsymbol{T} = (\boldsymbol{T}_1, \boldsymbol{T}_2, \cdots, \boldsymbol{T}_N)$, $\boldsymbol{\mathcal{Y}} = (\mathcal{Y}_1(O_1), \mathcal{Y}_2(O_2), \cdots, \mathcal{Y}_N(O_N))$, $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \cdots, \boldsymbol{\Delta}_N)$, $\boldsymbol{u} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_N)$, and $f(\boldsymbol{\theta})$ is the product of the prior distributions, i.e. $f(\boldsymbol{\theta}) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma})\pi(\alpha)\pi(\sigma)\pi(\boldsymbol{\Sigma})$. Here $\boldsymbol{\Sigma}$ is a $k \times k$ covariance matrix of the random effects. We adopt the following prior specifications: $\boldsymbol{\beta} \sim \mathcal{N}_p(\mathbf{0}, 10^3\mathbf{I}), \boldsymbol{\gamma} \sim \mathcal{N}_q(\mathbf{0}, 10^3\mathbf{I}), \alpha \sim \mathcal{N}(0, 10^3), \sigma \sim \mathcal{IG}(10^{-3}, 10^{-3}), \boldsymbol{\Sigma}^{-1} \sim Wishart(\mathbf{I}, k+1)$.

In QR, all parameter estimators are functions of the quantile. This is also true in the proposed QRJM. That is, parameter estimations in the time-to-event submodel, such as α and $\boldsymbol{\gamma}$, also change depending which τ is chosen. It is straightforward to code the proposed QRJM and implement the Bayesian algorithm in JAGS (Plummer, 2003) or other Bayesian sampling software.

1.2.3 Subject-Specific Dynamic Predictions of Terminal Event Risk

Upon fitting the QRJM to a study population that consists of N subjects (i.e. training data), we can then make predictions of survival probability for a new subject based on a set of his or her historical longitudinal measurements as well as other baseline covariates information. For terminal event outcome model, an implication of JM is that up to time t, until when the longitudinal measurements are available, the subject must be alive or free of event as $Y_i(t)$ serves as the internal time dependent covariate in the survival model. Thus, what we are really interested in is the conditional survival probability up to time $m = t + \Delta t$ ($\Delta t > 0$) given the survival up to time t, i.e.,

$$p_i(m|t) = Pr(T_i^* \geq m | T_i^* > t, \mathcal{Y}_i(t), \mathcal{D}_N; \boldsymbol{\theta}), \tag{1.8}$$

where $\mathcal{D}_N = \{T_i, \Delta_i, \boldsymbol{Y}_i, i = 1, \cdots, N\}$ denotes the training data of size N.

Equation (1.8) can be further elaborated as follows:

$$\begin{aligned}
&Pr(T_i^* \geq m | T_i^* > t, \mathcal{Y}_i(t), \mathcal{D}_N; \boldsymbol{\theta}) \\
&= \int Pr(T_i^* \geq m | T_i^* > t, \mathcal{Y}_i(t), \mathcal{D}_N, \boldsymbol{u}_i; \boldsymbol{\theta}) Pr(\boldsymbol{u}_i | T_i^* > t, \mathcal{Y}_i(t), \mathcal{D}_N; \boldsymbol{\theta}) d\boldsymbol{u}_i \\
&= \int Pr(T_i^* \geq m | T_i^* > t, \boldsymbol{u}_i; \boldsymbol{\theta}) Pr(\boldsymbol{u}_i | T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) d\boldsymbol{u}_i \\
&= \int \frac{S_i[m | \mathcal{M}_i(m, \boldsymbol{u}_i, \boldsymbol{\theta}); \boldsymbol{\theta}]}{S_i[t | \mathcal{M}_i(t, \boldsymbol{u}_i, \boldsymbol{\theta}); \boldsymbol{\theta}]} Pr(\boldsymbol{u}_i | T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) d\boldsymbol{u}_i,
\end{aligned} \tag{1.9}$$

where $S(\cdot)$ is the survival function conditional on the entire longitudinal history $\mathcal{M}_i(\cdot)$.

To estimate (1.9), we can take the advantage of the proposed Gibbs sampling algorithm and use the MCMC technique to calculate the posterior mean of the prediction. Specifically, we are going to estimate

$$\begin{aligned} E_{\boldsymbol{\theta}|\mathcal{D}_N}[p_i(m|t)] &= Pr(T_i^* \geq m | T_i^* > t, \mathcal{Y}_i(t), \mathcal{D}_N) \\ &= \int Pr(T_i^* \geq m | T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}_N) d\boldsymbol{\theta}, \end{aligned}$$

where the first part of the equation is given in (1.9).

A Monte Carlo (MC) approximation of $p_i(m|t)$ can be obtained using the following procedure:

1. Draw $\boldsymbol{\theta}^{(p)} \sim Pr(\boldsymbol{\theta}|\mathcal{D}_N)$ for $p = 1, \cdots, P$;

2. For each $\boldsymbol{\theta}^{(p)}$, draw $\boldsymbol{u}_i^{(q)} \sim f(\boldsymbol{u}_i | T_i^* > t, \mathcal{Y}_i(t), \boldsymbol{\theta}^{(p)})$ for $q = 1, \cdots, Q$ and compute

$$p_i^{(p)}(m|t) = \frac{1}{Q} \sum_{q=1}^{Q} S_i[m|\mathcal{M}_i(m, \boldsymbol{u}_i^{(q)}, \boldsymbol{\theta}^{(p)}); \boldsymbol{\theta}^{(p)}] S_i[t|\mathcal{M}_i(t, \boldsymbol{u}_i^{(q)}, \boldsymbol{\theta}^{(p)}); \boldsymbol{\theta}^{(p)}]^{-1};$$

3. Approximate $p_i(m|t)$ by $\hat{p}_i(m|t) = \frac{1}{P} \sum_{p=1}^{P} p_i^{(p)}(m|t)$ after collecting all P samples of $p_i(m|t)^{(p)}$.

In above algorithm P is the total number of MC iterations, $f(\boldsymbol{\theta}|\mathcal{D}_N)$ is the posterior distributions of $\boldsymbol{\theta}$ given in (1.7), and $f(\boldsymbol{u}_i | T_i^*, \mathcal{Y}_i(t), \boldsymbol{\theta}^{(k)})$ is the posterior distribution of the random effects for subject i.

1.2.4 Subject-Specific Dynamic Predictions of Recurrent Event Risk

Similarly, for recurrent event data let $\mathcal{Y}_i(t)$ be the observed complete longitudinal measurements, $\mathcal{M}_i(t)$ be the true underlying longitudinal trajectory up to time t, and $\mathcal{T}_{it-} = \{T_{ik} : 1 \leq k \leq K, T_{iK} < t\}$ be the recurrent times before time t. The predicted event-free probability at time m ($m > t$) given previous event times and longitudinal measurements up to the follow-up time t is:

$$p_i(m|t) = Pr(T_{iK+1} \geq m | T_{iK+1} > t, \mathcal{T}_{it-}, \mathcal{Y}_i(t); \boldsymbol{\theta}).$$

With further derivation:

$$\begin{aligned} p_i(m|t) &= \int Pr(T_{iK+1} \geq m | T_{iK+1} > t, \mathcal{T}_{it-}, \mathcal{Y}_i(t), \boldsymbol{u}_i; \boldsymbol{\theta}) \cdot Pr(\boldsymbol{u}_i | T_{iK+1} > t, \mathcal{T}_{it-}, \mathcal{Y}_i(t); \boldsymbol{\theta}) d\boldsymbol{u}_i \\ &= \int Pr(T_{iK+1} \geq m | T_{iK+1} > t, \mathcal{T}_{it-}, \boldsymbol{u}_i; \boldsymbol{\theta}) \cdot Pr(\boldsymbol{u}_i | T_{iK+1} > t, \mathcal{T}_{it-}, \mathcal{Y}_i(t); \boldsymbol{\theta}) d\boldsymbol{u}_i \\ &= \int \frac{Pr(T_{iK+1} \geq m | \mathcal{M}_i(m, \boldsymbol{u}_i; \boldsymbol{\theta}), \mathcal{T}_{it-}; \boldsymbol{\theta})}{Pr(T_{iK+1} > t | \mathcal{M}_i(t, \boldsymbol{u}_i; \boldsymbol{\theta}), \mathcal{T}_{it-}; \boldsymbol{\theta})} \cdot Pr(\boldsymbol{u}_i | T_{iK+1} > t, \mathcal{T}_{it-}, \mathcal{Y}_i(t); \boldsymbol{\theta}) d\boldsymbol{u}_i. \quad (1.10) \end{aligned}$$

And we approximate the prediction with it posterior mean:

$$\begin{aligned} E_{\boldsymbol{\theta}|\mathcal{D}_N}[p_i(m|t)] &= Pr(T_{iK+1} \geq m | T_{iK+1} > t, \mathcal{T}_{it-}, \mathcal{Y}_i(t)) \\ &= \int Pr(T_{iK+1} \geq m | T_{iK+1} > t, \mathcal{T}_{it-}, \mathcal{Y}_i(t); \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}_N) d\boldsymbol{\theta}, \end{aligned}$$

where the first part of the equation is given in (1.10).

A similar Monte Carlo (MC) procedure to approximate $p_i(m|t)$ is carried out as follows:

1. Draw $\boldsymbol{\theta}^{(p)} \sim Pr(\boldsymbol{\theta}|\mathcal{D}_N)$ for $p = 1, \cdots, P$;

2. For each $\boldsymbol{\theta}^{(p)}$, draw $\boldsymbol{u}_i^{(q)} \sim Pr(\boldsymbol{u}_i|\mathcal{D}_N, \boldsymbol{\theta}^{(p)})$ for $q = 1, \cdots, Q$, and approximate $p_i(m|t)^{(p)}$ by

$$\frac{1}{Q} \sum_{q=1}^{Q} \frac{Pr(T_{iK+1} \geq m | \mathcal{M}_i(m, \boldsymbol{u}_i^{(q)}; \boldsymbol{\theta}^{(p)}), \mathcal{T}_{it-}; \boldsymbol{\theta}^{(p)})}{Pr(T_{iK+1} > t | \mathcal{M}_i(t, \boldsymbol{u}_i^{(q)}; \boldsymbol{\theta}^{(p)}), \mathcal{T}_{it-}; \boldsymbol{\theta}^{(p)})};$$

3. Approximate $p_i(m|t)$ by $\frac{1}{P}\sum_{p=1}^{P} p_i(m|t)^{(p)}$.

In above estimation procedure, $Pr(\boldsymbol{u}_i|\mathcal{D}_N, \boldsymbol{\theta}^{(p)})$, i.e., $f(\boldsymbol{u}_i|T_{iK+1} > t, \mathcal{T}_{it-}, \mathcal{Y}_i(t), \boldsymbol{\theta}^{(p)})$ is the posterior distribution of the random effects for subject i. Standard error of the prediction can be computed using the sample variance.

In Step 2 of above algorithms, the posterior predictive values of the random effects are directly results from the MCMC iterations if the subject is within the training data. For out-of-sample subjects who don not belong to the original study population, we can use the inference results from the training data to run additional MCMC iterations to obtain such predictions and the rest of the algorithm follows. Since for each individual there are only a few random effects (two in our current model) to estimate, a short MCMC with 200 iterations should be sufficient for converge (Taylor et al., 2013).

In summary, it is relatively easy to make subject-specific predictions of event-free probability from the posterior samples of the fixed effects and the posterior predictive distributions of the random effects, which are direct results of our sampling algorithm. In addition, by using the MCMC technique, uncertainty of the predictive inference is fully captured in the posterior distribution and no asymptotic theory is needed to derive the standard error.

1.2.5 Predictive Accuracy

Predictive accuracy of a model can be evaluated from different perspectives, such as discrimination, calibration, and reclassification, etc. Discrimination measures a model's ability in identifying events versus non-events. Calibration quantifies the closeness of the predictions and the observed values. While reclassification assesses the improvement of a model in prediction after adding new predictor(s). In this work, we mainly focus on the discriminative ability of our model. Area under the receiver operating characteristic curve (AUC) is a commonly used statistics to evaluate the discriminative ability in prediction, while above average risk difference (AARD) measures the difference in the risk rates comparing events versus non-events at the level of population average risk, and mean risk difference (MRD) is the average difference between TPR and FPR across the risk scale (Pepe et al., 2008). In this work, we use all these three measurements as summary statistics to evaluate the predictive performance of our model.

Following Zheng et al. (2013) and Yang et al. (2016), at a given time t, a future time $t+\Delta t$ and a threshold c, the true positive rate (TPR) and false positive rate (FPR) of the predictive results can be defined as follows:

$$\mathrm{TPR}_t^{\Delta t}(c) = Pr(\mathbf{1} - \boldsymbol{p}(t+\Delta t|t) \geq c|\boldsymbol{T} \leq t+\Delta t),$$

$$\mathrm{FPR}_t^{\Delta t}(c) = Pr(\mathbf{1} - \boldsymbol{p}(t+\Delta t|t) \geq c|\boldsymbol{T} > t+\Delta t),$$

where $\boldsymbol{p}(t+\Delta t|t)$ is a vector of predicted event-free probabilities at time $t+\Delta t$ based on the longitudinal measurements up to time t:

$$p_i(t+\Delta t|t) = S_i(t+\Delta t|\mathcal{Y}_i(t), \boldsymbol{u}_i; \boldsymbol{\theta}), i = 1, \cdots, N.$$

The estimate of $\boldsymbol{p}(t+\Delta t|t)$ is denoted by $\hat{\boldsymbol{p}}(t+\Delta t|t)$ and the estimators of TPR and FPR can be written as:

$$\widehat{TPR}_t^{\Delta t}(c) = \frac{\sum_{i=1}^{N}(1-\hat{p}_i(t+\Delta t|t))I(1-\hat{p}_i(t+\Delta t|t) \geq c)}{\sum_{i=1}^{N}(1-\hat{p}_i(t+\Delta t|t))},$$

$$\widehat{FPR}_t^{\Delta t}(c) = \frac{\sum_{i=1}^{N}\hat{p}_i(t+\Delta t|t)I(1-\hat{p}_i(t+\Delta t|t) \geq c)}{\sum_{i=1}^{N}\hat{p}_i(t+\Delta t|t)}.$$

And by definition:

$$\widehat{AUC}_t^{\Delta t} = \int \widehat{TPR}_t^{\Delta t}\left\{(\widehat{FPR}_t^{\Delta t})^{-1}(u)\right\} du,$$

$$\widehat{AARD}_t^{\Delta t} = \widehat{TPR}_t^{\Delta t}(\hat{\rho}) - \widehat{FPR}_t^{\Delta t}(\hat{\rho}),$$

$$\widehat{MRD}_t^{\Delta t} = \int_c \widehat{TPR}_t^{\Delta t}(c)dc - \int_c \widehat{FPR}_t^{\Delta t}(c)dc.$$

And in AARD, $\hat{\rho} = \frac{\sum_{i=1}^{N}(1-\hat{p}_i(t+\Delta t|t))}{N}$ is the average risk in the study population at time $t + \Delta t$.

1.3 Public Health Significance

Traditional health care interventions are designed based on the average treatment effect on the population. It's not surprising that some interventions may work perfectly on some patients but fail on others because of various individual health profiles. To make the treatment more effective, ideally, personalized health care strategy should be adopted and treatments are tailored based specific disease information of each individual patient. The practice of customized treatment does already exist in treating tumors where doctors prescribe drugs based on the tumor's growth and some specific gene mutations of the patient. However, there is much more needs to be done in order to extend the idea of personalized medicine for patients suffering with other diseases. As Precision Medicine Initiative (Collins and Varmus, 2015) strives for, development of statistics tools that facilitate the analysis for individual patient could greatly help physicians make personalized treatment decisions in order to achieve better clinical outcomes.

In biomedical context, it is common that extremer values of the disease marker are related with higher risk of various medical events, including death and disease recurrences. However, it is hard to look at those tails of the biomarker distribution using traditional linear regression based methods, which models only the covariate effects on the conditional mean of the outcome. The proposed QRJM in this dissertation work provides much more flexibility in modeling covariate effects on any specific conditional quantile of the longitudinal continuous outcome meanwhile studying its relationship with the risk of medical event(s). By using the QRJM, variation of treatment effect on the outcome can be thoroughly examined, based on which different and more specific treatment strategies can be made. In real-world applications, regression quantile(s) can be chosen to reflect specific research interest, to better understand the risk factor-and-disease relationship, and to plan more effective disease interventions.

Public health is a science discipline that aims to improve the health of the entire community by preventing disease and ensuring better health care outcomes. Another important application this dissertation work is to make subject-specific predictions of future event probability. Such predictions provide important information about patient specific disease progression in terms of the likelihood of potential event(s) in the future. Thus patient-specific intervention method can be tailored to postpone or prevent the event from happening. Moreover, such predictions can be dynamically updated as we accumulate more longitudinal biomarker data, for those who are still at risk of event, and more event history for those who experience disease recurrences. This would allow us to adjust treatment plan continuously for each individual patient and the delivery of health care would be much more efficient. This idea of subject-specific dynamic predictions fits into the big concept of "personalized medicine", which aims to provide the right patient with the right drug at the right time (FDA, 2013).

1.4 Research Questions and Specific Aims

JM of longitudinal and time-to-event data using QR has been little studied so far. To our knowledge, Farcomeni and Viviani (2015) is the first work that extended LMJM to incorporate a QR model in the longitudinal process. In their paper the parameter estimations are obtained using the Monte Carlo expectation and maximization (MCEM) method. Also, no work has been done yet to extend the subject-specific dynamic predictions method to the QRJM framework. In terms of JM of longitudinal and recurrent event data, there is a scarcity of work being done either for inference or for prediction purpose. To this end, we propose a fully Bayesian algorithm for model estimation and dynamic predictions under the new QRJM framework. In the time-to-event process, we consider both terminal and recurrent event outcomes. The algorithm consists of two parts. First, we draw statistical inference of the disease progression and the association between a longitudinal biomarker and the time-to-event outcome based on a large study population. Second, we make predictions of future event-free probability for a subject with his or her longitudinal biomarker trajectory information (as well as recurrent events history in the case of recurrent event outcome). To estimate the risk of future event(s), we explore the posterior distributions of the fixed effects and the predictions of the subject-specific random effects from our algorithm. Moreover, we dynamically update the predictions as long as new longitudinal measurements and event data can be obtained.

Specifically, we aim to achieve the following three objectives in this dissertation work:

1. In paper 1, we propose a fully Bayesian algorithm for model inference under the QRJM framework, based on which a subject-specific dynamic prediction method of survival probability is developed. The proposed inference and prediction algorithms are implemented using existing software. In data application, we use the QRJM to study risk factors of Huntinton's disease (HD) onset and to make predictions of the risk of HD onset in the future. Our hypothesis was that traditional linear model JM (LMJM) would produce biased parameter estimation and predicted survival probability when the longitudinal outcome is skewed; in contrast, QRJM should be more robust against non-normality. QRJM with correctly specified quantile would outperform LMJM in both model estimation and dynamic predictions. In predicting future event-free probability, with additional longitudinal measurements, our model would be able to incorporate such information and update the predictions accordingly.

2. In paper 2, we extend our QRJM to study the recurrent event data and develop a fully Bayesian algorithm for model inference. We compare the performance of QRJM and LMJM under different error distributions (i.e. skewed v.s. normal). The proposed model is applied to a real-world data set to study various covariate effects on different quantiles of SBP and its association with the risk of recurrent CHD. Our hypothesis was that QRJM should be more appropriate to use rather than LMJM when the longitudinal outcome is possibly skewed and covariate effects would be different according to the specification of regression quantiles in the QRJM.

3. In paper 3, we develop a subject-specific dynamic predictions algorithm for future recurrent event probability based on the QRJM of longitudinal and recurrent event data. The predictions can be dynamically updated when additional longitudinal and/or recurrent event data are available. Prediction performances between QRJM and LMJM are compared in simulation study where data are generated from skewed and normal distributions. The proposed predictive algorithm is applied to make dynamic predictions of recurrent CHD risk based on historical SBP and CHD events data. Our hypothesis was the predictive performance from QRJM should outperform than that from the LMJM at some quantiles but not for all quantiles as some conditional quantiles of SBP are more informative in predicting CHD events.

2 Journal Article 1

Title of Journal Article

Bayesian Quantile Regression Joint Models: inference and Dynamic Predictions

Journal proposed for article submission: Statistics in Medicine

Bayesian Quantile Regression Joint Models: Inference And Dynamic Predictions

Abstract

In the traditional joint models (JM) of a longitudinal and time-to-event outcome, a linear mixed model (LMM) assuming normal random errors is typically used to model the longitudinal process. However, in many circumstances, the normality assumption cannot be satisfied and the LMM is not an appropriate sub-model in the JM. In addition, as the LMM models the conditional mean of the longitudinal outcome, it is not appropriate if clinical interest lies in making inference or prediction on median, lower, or upper ends of the longitudinal process. To this end, quantile regression (QR) provides a flexible, distribution-free way to study covariate effects at different quantiles of the longitudinal outcome and it is robust not only to deviation from normality, but also to outlying observations. In this article, we present and advocate the linear quantile mixed model (LQMM) for the longitudinal process in the JM framework. Our development is motivated by a large prospective study of Huntington's Disease (HD) where primary clinical interest is in utilizing longitudinal motor scores and other early covariates to predict the risk of developing HD. We develop a Bayesian method based on the location-scale representation of the asymmetric Laplace distribution, assess its performance through an extensive simulation study, and demonstrate how this LQMM-based JM approach can be used for making subject-specific dynamic predictions of survival probability.

Key words: Asymmetric Laplace distribution; Bayesian; Dynamic predictions; Huntington's disease; Joint models; Linear quantile mixed model.

2.1 Introduction

Joint models (JM) of longitudinal and time-to-event data have been well developed in the literature as they are highly applicable in the setting of longitudinal data analysis with event-related terminal events (e.g., dropout or death which is related to the longitudinal process), or for survival analysis with a time dependent covariate measured with error (Self and Pawitan, 1992; Tsiatis et al., 1995). JM have been extensively studied in recent years and an excellent review is provided by Tsiatis and Davidian (2004). Within the traditional JM framework, a linear mixed model (LMM) sub-model is used to model the longitudinal continuous outcome. To account for within-subject correlation, measurements from the same subject share the random effects, and random errors are assumed to be normally distributed (Laird and Ware, 1982). However, if normality assumption of the random errors is violated (even after applying various outcome transformations), an LMM is not appropriate. Further, an LMM models covariate effects on the conditional mean of the longitudinal outcome. However, in many clinical studies it is of interest to make inference or prediction on median, lower or higher quantiles of the longitudinal outcome. For example, in a study of the impact of various health factors on low birth weight, which is closely associated with infant mortality and development of chronic diseases, researchers fit a quantile regression (QR) model and find several significantly different covariate effects on lower birth weight compared with that on the mean birth weight (Koenker and Hallock, 2001). Thus, QR may serve as an alternative to conditional mean regression when the aforementioned limitations are present, or when the subject matter dictates.

In contrast to mean regression such as LMM, original QR models offer a flexible framework that relaxes the distributional assumption, and provides a way to study covariate effects on various quantiles of the outcome (Koenker, 2005). QR has attracted much attention since the seminal work of Koenker and Basset (1978), and extensions for longitudinal data have been explored by many authors, i.e., Koenker (2004), Geraci and Bottai (2007), Geraci and Bottai (2014), and Kozumi and Kobayashi (2011). To our knowledge, Farcomeni and Viviani (2015) first incorporated a linear quantile mixed model (LQMM) into a JM for longitudinal and survival processes, for which they utilized a Monte Carlo Expectation Maximization (MCEM) algorithm for parameter estimation.

Our work is motivated by the Neurobiological Predictors of Huntington's Disease (HD) study (PREDICT-HD; ClinicalTrials.gov number NCT00051324), a prospective observational study (n=1078) designed to detect early neurobiological predictors of HD. A total of 40 longitudinal biomarkers were measured during the study follow-up of 12 years, where the outcome of interest was time to motor diagnosis of HD (referred to as HD onset). As the primary focus of this study was to measure the association between longitudinal biomarkers and risk of HD onset, the clinical question lends itself to a joint modeling approach. The left panel of Figure 2.1 displays the scatter plot and kernel density (on the right side) of total motor score (TMS), a commonly used rating criteria of body motion abilities based on the Unified Huntington Disease Rating Scale (UHDRS). The left panel suggests that TMS increases (deteriorates) as time progresses. The right panel of Figure 2.1 displays the mean TMS values over time for subjects with follow-up time less than 48 months (solid red line), $48-72$ months (dotted green line), and more than 72 months (slashed blue line). It indicates that subjects with shorter follow-up tend to have higher TMS values (worse mobility). Because subjects with higher TMS are at greater risk of developing HD, it is of particular interest to explore the disease progression in a higher TMS subpopulation as opposed to modeling the conditional mean of TMS. Similar plots for other biomarkers are presented in Section 2.4.

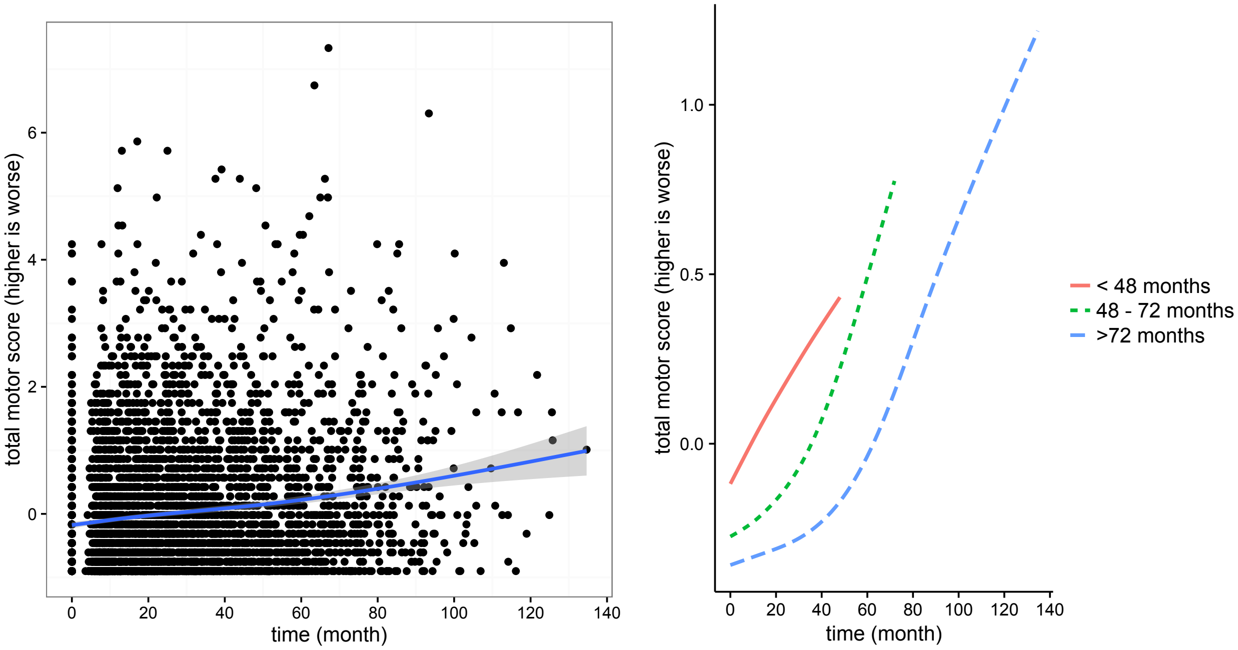


Figure 2.1: Left panel: Scatter plot (with loess curve) and kernel density plot (right side) for total motor score from the study population (time unit: month; lower total motor score is better); right panel: Mean total motor score values over time.

Moreover, we are specifically interested in estimating the risk of developing HD for those study participants who are still free of the disease. The JM framework offers a novel way of making such personalized dynamic predictions of future event-free probability (Rizopoulos, 2011; Taylor et al., 2013). A key feature of these dynamic predictions frameworks is that the predictive measures can be dynamically updated as additional longitudinal measurements become available for the target subjects, providing instantaneous risk assessment. In order to make dynamic predictions in the LQMM-based JM framework, we develop a fully Bayesian algorithm for model estimation and dynamic predictions under the new quantile regression joint models (QRJM) framework. The algorithm consists of two parts. First, we build the QRJM framework consisting of an LQMM for the longitudinal process and a proportional hazard model (PHM) for the time to HD onset process and then draw Bayesian inference, based on a large study population. Second, we make predictions of future HD-free probability for a subject using his or her longitudinal biomarker trajectory information. To estimate the risk of a future event, we explore the posterior distribution of the fixed effects and the predictions of the subject-specific random effects from our QRJM. Moreover, we dynamically update the predictions as long as an individual is free of HD and new longitudinal measurements can be obtained. Our work is different from Farcomeni and Viviani (2015) in the following: (1) we consider a fully Bayesian QRJM for statistical inference while Farcomeni and Viviani (2015) used a MCEM algorithm; (2) more importantly, by taking advantage of the posterior distributions of model parameters and subject-specif random effects, we develop a dynamic predictions procedure for future event-free probability under the proposed QRJM.

The rest of this article proceeds as follows. In Section 2.2, we give details of the QRJM and statistical methods used for inference and dynamic predictions. In Section 2.3, we present two simulation studies to validate the proposed methods. In Section 2.4, we apply the proposed methods to the motivating data set. We conclude the article with a discussion in Section 2.5.

2.2 Methods

2.2.1 Bayesian Linear Quantile Mixed Model

Let $Y_i(t_{ij})$ be the longitudinal outcome for subject i measured at time t_{ij} where $i = 1, \cdots, N$ and $j = 1, \cdots, n_i$. Consider the linear mixed effects model:

$$Y_i(t) = \boldsymbol{X}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i + \varepsilon_i(t), \varepsilon_i(t) \sim N(0, \sigma^2), \tag{2.1}$$

where $\boldsymbol{\beta}$ is a $p-$dimensional fixed effect vector, $\boldsymbol{X}_i(t)$ contains the corresponding fixed covariates, $\boldsymbol{u}_i$ is a $k-$dimensional random effect vector for subject i, and $\boldsymbol{Z}_i(t)$ contains the corresponding random covariates.

A linear quantile mixed model (LQMM) assumes that the conditional quantile of the outcome is a linear function of the covariates,

$$Q_{Y_i(t)|\boldsymbol{X}_i(t),\boldsymbol{Z}_i(t)}(\tau) = \boldsymbol{X}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i, \tag{2.2}$$

where the τth quantile of a random variable Y is defined as $Q_Y(\tau) = F_Y^{-1}(\tau) = \inf\{y : F_Y(y) \geq \tau\}$ for $\tau \in [0, 1]$. The quantile regression estimates can be obtained by minimizing the following loss function, $\hat{\boldsymbol{\beta}}_\tau = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i,t} \left[\rho_\tau\left(Y_i(t) - \boldsymbol{X}_i^\top(t)\boldsymbol{\beta} - \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i\right)\right]$, where $\rho_\tau(\cdot)$ is defined as $\rho_\tau(Y) = Y(\tau - I(Y < 0))$. In quantile regression, parameter estimators are functions of the quantile. So parameter $\boldsymbol{\beta}_\tau$ is a function of quantile τ, as denoted by the subscript.

As discussed in Koenker and Machado (1999) and Yu and Moyeed (2001), the above minimization problem can be rephrased as a maximum-likelihood problem by assuming that the random error $\varepsilon_i(t)$ in (2.1) follows the asymmetric Laplace distribution (ALD), denoted by $ALD(0, \sigma, \tau)$ with location parameter equals 0, scale parameter $\sigma > 0$ and skewness parameter $\tau \in (0, 1)$. $ALD(0, \sigma, \tau)$ is skewed to left when $\tau > 0.5$, and skewed to right when $\tau < 0.5$.

When $\tau = 0.5$, ALD reduces to the symmetric Laplace distribution. To visualize this, Web Figure 1 displays the density functions of a standard normal distribution, a Laplace distribution, and two ALDs with $\tau = 0.75$ and $\tau = 0.25$, respectively. Adopting the ALD, the LQMM in (2.2) becomes $Y_i(t) = \boldsymbol{X}_i^\top(t)\boldsymbol{\beta}_\tau + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i + \varepsilon_i(t), \varepsilon_i(t) \sim ALD(0, \sigma, \tau)$, where $i = 1, \cdots, N$ and $t = 1, \cdots, n_i$. The conditional likelihood function is $\ell(Y_i(t)|\boldsymbol{\beta}_\tau, \boldsymbol{u}_i, \sigma) = \frac{\tau(1-\tau)}{\sigma} \exp\left[-\rho_\tau\left(\frac{Y_i(t)-\boldsymbol{X}_i^\top(t)\boldsymbol{\beta}_\tau-\boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i}{\sigma}\right)\right]$.

Linear programming algorithms can be applied to obtain parameter estimates under the frequentist framework. However, to develop a Bayesian sampling algorithm for model inference, we utilize a location-scale mixture representation of the ALD (Kotz et al., 2012), which is a functional form with a mixture of common distributions. Under this parameterization, the random error is represented as $\varepsilon_i(t) = \kappa_1 e_i(t) + \kappa_2\sqrt{\sigma e_i(t)}v_i(t)$ with $v_i(t) \sim \mathcal{N}(0, 1), e_i(t) \sim \exp(1/\sigma)$, $\kappa_1 = \frac{1-2\tau}{\tau(1-\tau)}$, and $\kappa_2^2 = \frac{2}{\tau(1-\tau)}$.

This reparameterization leads to the following LQMM,

$$Y_i(t) = \boldsymbol{X}_i^\top(t)\boldsymbol{\beta}_\tau + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i + \kappa_1 e_i(t) + \kappa_2\sqrt{\sigma e_i(t)}v_i(t), \tag{2.3}$$

or equivalently, the conditional likelihood function is

$$\ell(Y_i(t)|\boldsymbol{\beta}_\tau, \boldsymbol{u}_i, e_i(t), \sigma) = \frac{1}{\sqrt{2\pi\kappa_2^2\sigma e_i(t)}} \exp\left[-\frac{(Y_i(t) - \boldsymbol{X}_i^\top(t)\boldsymbol{\beta}_\tau - \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i - \kappa_1 e_i(t))^2}{2\kappa_2^2\sigma e_i(t)}\right]. \tag{2.4}$$

As discussed in Yu and Moyeed (2001), irrespective of the actual distribution of the data, Bayesian quantile regression using ALD distribution works quite well for different error distributions and the performance is quite robust and satisfactory.

2.2.2 Joint Models Using Longitudinal Quantile Regression

We then extend the regular joint models (consisting of a linear mixed sub-model for the longitudinal process and a Cox proportional hazards model (PHM) submodel for the survival process, referred to as LMJM), by replacing the linear mixed sub-model with an LQMM as in (2.3). Let $T_i = min(T_i^*, C_i)$ be the observed event time for subject i, where T_i^* is the true underlying event time and C_i is the censoring time. Let Δ_i be the event indicator (1 if the event is observed, and 0 otherwise). Let $Y_i(t)$ be the continuous longitudinal outcome for subject i measured at time t. Note that $Y_i(t)$ is only observed when $t \leq T_i$, and the complete longitudinal measurements for subject i can be written as $\mathcal{Y}_i(t) = \{Y_i(s) : 0 \leq s \leq t\}$. We denote the true underlying longitudinal measurement for subject i at time t with $m_i(t)$ and his/her complete history of true longitudinal process as $\mathcal{M}_i(t) = \{m_i(s) : 0 \leq s \leq t\}$. The proposed quantile regression joint models (QRJM) can be written as a set of two sub-models:

$$\begin{cases} Y_i(t) = m_i(t) + \varepsilon_i(t) = \boldsymbol{X}_i^\top(t)\boldsymbol{\beta}_\tau + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i + \varepsilon_i(t), \varepsilon_i(t) \sim ALD(0, \sigma, \tau) \\ h(T_i|\mathcal{M}_i(T_i), \boldsymbol{W}_i; \boldsymbol{\gamma}_\tau, \alpha_\tau) = h_0(T_i)\exp(\boldsymbol{W}_i^\top\boldsymbol{\gamma}_\tau + \alpha_\tau(\boldsymbol{X}_i^\top(T_i)\boldsymbol{\beta}_\tau + \boldsymbol{Z}_i^\top(T_i)\boldsymbol{u}_i)), \end{cases} \tag{2.5}$$

where the first sub-model is the LQMM introduced in Section 2.2.1, in which $\boldsymbol{X}_i(t)$ are the fixed effect covariates and $\boldsymbol{Z}_i(t)$ are the covariates associated with $k-$dimensional multivariate normal random effects $\boldsymbol{u}_i$. The second sub-model takes the format of PHM where $h_0(\cdot)$ is the baseline hazard function and $\boldsymbol{W}_i$ are the $q-$dimensional fixed effect covariates only associated with event time (not the longitudinal outcome). These two sub-models are linked by incorporating $m_i(t)$ (the true underlying longitudinal measurement at time t) in the time-to-event process. The association parameter α_τ quantifies the strength of association between $m_i(t)$ and the hazard for event at the same time point, e.g., positive α_τ indicates that subjects with higher measurement tend to have an event earlier.

In the proposed QRJM (2.5), all parameters are functions of quantile τ. Thus, by choosing different quantiles, one can conduct a comprehensive analysis of the relationship between the outcome and the covariates. Depending on the research aims, we can take different strategies to utilize the flexibility of the QRJM. For example, to conduct a study over the entire conditional distribution of the longitudinal outcome, we can just fit the QRJM through a set of selected quantiles, collect and compare the resulting parameter estimations. Less varying values in the parameter estimates indicates a relatively stable covariate effects on the outcomes, and vice versa. On the other hand, the interest may lie only in assessing the effect on some pre-specified quantiles (median, lower or higher quantile) of the longitudinal outcome and its association with the event process.

2.2.3 The Survival Sub-model

For subject i, the likelihood function for survival data is:

$$\ell(T_i, \Delta_i | \boldsymbol{u_i}) = h(T_i | \mathcal{M}_i(T_i), \boldsymbol{W}_i)^{\Delta_i} S(T_i | \mathcal{M}_i(T_i), \boldsymbol{W}_i), \tag{2.6}$$

where $h(T_i|\mathcal{M}_i(T_i), \boldsymbol{W}_i)$ is given in (2.5) and $S(\cdot)$ is the survival function,

$$S(T_i|\mathcal{M}_i(T_i), \boldsymbol{W}_i) = \exp\left\{ -\int_0^{T_i} h_0(s) \exp(\boldsymbol{W}_i{}^\top \boldsymbol{\gamma}_\tau + \alpha(\boldsymbol{X}_i^\top(s)\boldsymbol{\beta}_\tau + \boldsymbol{Z}_i^\top(s)\boldsymbol{u}_i))ds \right\}.$$

For the baseline hazard $h_0(t)$, a parametric form such as exponential model can be used or it can be left unspecified. Specifically, we consider the piecewise-constant baseline hazard function, based on which a closed form survival function can be derived for each time interval. Further extension of the JM in the functional form of the two processes is also possible, as discussed in Rizopoulos et al. (2014). Although all parameters in the proposed QRJM are quantile dependent, for national ease and without ambiguity, we omit the subscript τ from all parameters in the following sections (e.g., $\boldsymbol{\theta}$ stands for $\boldsymbol{\theta}_\tau$ for all quantile-based parameters).

2.2.4 Complete Likelihood Function and Bayesian Inference

For subject i, the complete joint likelihood of the longitudinal and survival data can be written as

$$L_i(\boldsymbol{\theta}; T_i, \Delta_i, \mathcal{Y}_i(T_i), \boldsymbol{u}_i) = \ell(\mathcal{Y}_i(T_i)|\boldsymbol{u}_i)\ell(T_i, \Delta_i|\boldsymbol{u}_i)f(\boldsymbol{u}_i|\boldsymbol{\Sigma}), \tag{2.7}$$

where vector $\boldsymbol{\theta}$ represents all the parameters in (2.7), $\ell(\mathcal{Y}_i(T_i)|\boldsymbol{u}_i) = \prod_{0\le t\le T_i} \ell(Y_i(t)|\boldsymbol{u}_i)$, where $\ell(Y_i(t)|\boldsymbol{u}_i)$ is given in (2.4), and $\ell(T_i, \Delta_i|\boldsymbol{u}_i)$ is given in (2.6).

Parameter estimation can be made using Monte Carlo EM (MCEM) algorithm, in which random effects are treated as missing data (Farcomeni and Viviani, 2015). In this article, however, we take advantage of the location-scale mixture representation of the ALD described in Section 2.2.1 and propose a fully Bayesian inference approach for parameter estimation and personalized dynamic predictions. Given the complete likelihood in (2.7) and by Bayes theorem, the posterior distributions of the model parameters are given by

$$f(\boldsymbol{\theta}|\boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{\mathcal{Y}}, \boldsymbol{u}) \propto \prod_{i=1}^{N} L_i(T_i, \Delta_i, \mathcal{Y}_i(T_i), \boldsymbol{u}_i; \boldsymbol{\theta})f(\boldsymbol{\theta}), \tag{2.8}$$

where $\boldsymbol{T} = (T_1, T_2, \cdots, T_N)$, $\boldsymbol{\mathcal{Y}} = (\mathcal{Y}_1(T_1), \mathcal{Y}_2(T_2), \cdots, \mathcal{Y}_N(T_N))$, $\boldsymbol{\Delta} = (\Delta_1, \Delta_2, \cdots, \Delta_N)$, $\boldsymbol{u} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_N)$, and $f(\boldsymbol{\theta}) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma})\pi(\alpha)\pi(\sigma)\pi(\boldsymbol{\Sigma})$ is the product of the prior distributions, where $\boldsymbol{\Sigma}$ is a $k \times k$ covariance matrix of the multivariate normal random effects distribution. We adopt the following prior distributions: $\boldsymbol{\beta} \sim \mathcal{N}_p(\mathbf{0}, 10^3\mathbf{I}), \boldsymbol{\gamma} \sim \mathcal{N}_q(\mathbf{0}, 10^3\mathbf{I}), \alpha \sim \mathcal{N}(0, 10^3), \sigma \sim \mathcal{IG}(10^{-3}, 10^{-3}), \boldsymbol{\Sigma}^{-1} \sim Wishart(\mathbf{I}, k+1)$. We also consider the Cholesky decomposition prior for $\boldsymbol{\Sigma}$ in our simulation studies and find similar results as Wishart prior gives (results not shown). We have investigated other selections of vague prior distributions with various hyper-parameters and obtained very similar results.

The advantages of using fully Bayesian approach include that the uncertainty of the parameter estimates is fully captured in the posterior distributions and no asymptotic theory is needed to derive the standard error. The fully Bayesian approach provides a straightforward framework to make subject-specific prediction of survival probability using the posterior samples of the parameters and of the posterior predictive distributions for the random effects. Moreover, the proposed QRJM can be readily implemented in JAGS software (version 4.0.0) (Plummer, 2003) and the codes have been posted at the Web Supplement to facilitate easy reading and implementation of the proposed QRJM model.

2.2.5 Predictions of Survival Probability

Upon fitting the QRJM to a training dataset with N subjects, we can make prediction of survival probability for a new subject based on a set of his or her historical longitudinal measurements (denoted by $\mathcal{Y}_i(t)$) as well as other covariates information. The conditional survival probability up to time $m = t + \Delta t$ ($\Delta t > 0$), given that the subject is event-free up to censoring time t, is denoted as $p_i(m|t) = Pr(T_i^* \geq m | T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta})$, which can be further elaborated as follows:

$$
\begin{aligned}
&Pr(T_i^* \geq m | T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) \\
&= \int Pr(T_i^* \geq m | T_i^* > t, \mathcal{Y}_i(t), \boldsymbol{u}_i; \boldsymbol{\theta}) Pr(\boldsymbol{u}_i | T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) d\boldsymbol{u}_i \\
&= \int Pr(T_i^* \geq m | T_i^* > t, \boldsymbol{u}_i; \boldsymbol{\theta}) Pr(\boldsymbol{u}_i | T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) d\boldsymbol{u}_i \\
&= \int \frac{S_i[m | \mathcal{M}_i(m, \boldsymbol{u}_i, \boldsymbol{\theta}); \boldsymbol{\theta}]}{S_i[t | \mathcal{M}_i(t, \boldsymbol{u}_i, \boldsymbol{\theta}); \boldsymbol{\theta}]} Pr(\boldsymbol{u}_i | T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) d\boldsymbol{u}_i,
\end{aligned}
\tag{2.9}
$$

where $S(\cdot)$ is the survival function conditional on the entire longitudinal history $\mathcal{M}_i(\cdot)$.

To estimate (2.9), we can use the proposed Bayesian sampling algorithm in Section 2.2.4 to calculate the posterior mean of the prediction $E_{\boldsymbol{\theta}|\mathcal{D}_N}[p_i(m|t)]$ and

$$E_{\boldsymbol{\theta}|\mathcal{D}_N}[p_i(m|t)] \quad = \quad Pr(T_i^* \geq m|T_i^* > t, \mathcal{Y}_i(t), \mathcal{D}_N) = \int Pr(T_i^* \geq m|T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}_N) d\boldsymbol{\theta},$$

where $\mathcal{D}_N = \{T_i, \Delta_i, \boldsymbol{Y}_i, i = 1, \cdots, N\}$ denotes the training data of size N and the first part of the equation is given in (2.9).

A Monte Carlo (MC) estimate of $p_i(m|t)$ can be obtained using the following procedure:

1. Draw $\boldsymbol{\theta}^{(p)} \sim Pr(\boldsymbol{\theta}|\mathcal{D}_N)$ for $p = 1, \cdots, P$;
2. For each $\boldsymbol{\theta}^{(p)}$, draw $\boldsymbol{u}_i^{(q)} \sim f(\boldsymbol{u}_i|T_i^* > t, \mathcal{Y}_i(t), \boldsymbol{\theta}^{(p)})$ for $q = 1, \cdots, Q$ and compute
$$p_i^{(p)}(m|t) = \frac{1}{Q}\sum_{q=1}^{Q} S_i[m|\mathcal{M}_i(m, \boldsymbol{u}_i^{(q)}, \boldsymbol{\theta}^{(p)}); \boldsymbol{\theta}^{(p)}] S_i[t|\mathcal{M}_i(t, \boldsymbol{u}_i^{(q)}, \boldsymbol{\theta}^{(p)}); \boldsymbol{\theta}^{(p)}]^{-1};$$
3. Approximate $p_i(m|t)$ by $\hat{p}_i(m|t) = \frac{1}{P}\sum_{p=1}^{P} p_i^{(p)}(m|t)$ after collecting all P samples of $p_i(m|t)^{(p)}$.

In above algorithm, P is the total number of MC iterations, $f(\boldsymbol{\theta}|\mathcal{D}_N)$ is the posterior distributions of $\boldsymbol{\theta}$ given in (2.8), and $f(\boldsymbol{u}_i|T_i^*, \mathcal{Y}_i(t), \boldsymbol{\theta}^{(k)})$ is the posterior distribution of the random effects for subject i. And the uncertainty of the predictions is captured in the sample variance.

The posterior predictive values of the random effects $\boldsymbol{u}_i$ are direct results from the MCMC iterations if the subject is from the training dataset. For a new subject who is not in the training dataset, we can use the inference results to run additional MCMC iterations to obtain samples for the new subject's random effects $\boldsymbol{u}_i$ and the rest of the algorithm follows. Because each individual only has a few random effects (two in our current model) to estimate, a short MCMC with 200 iterations should be sufficient for convergence (Taylor et al., 2013).

2.2.6 Predictive Accuracy

Predictive accuracy of a model can be evaluated from different perspectives, such as discrimination (how well the models discriminate between subjects who had the event from those who did not), validation (how well the models predict the observed data), and reclassification (how well the model prediction improves by adding new predictors). Here we mainly focus on the discriminative ability of our model. Area under the receiver operating characteristic curve (AUC) is a commonly used statistics to evaluate the discriminative ability in prediction, while above average risk difference (AARD) measures the difference in the risk rates comparing events versus non-events at the level of population average risk, and mean risk difference (MRD) is the average difference between true positive rate (TPR) and false positive rate (FPR) across the risk scale (Pepe et al., 2008). Higher values in AUC, AARC, and MRD indicate better discriminative ability. We use all three measurements as summary statistics to evaluate the discriminative performance of our model.

Following Zheng et al. (2013) and Yang et al. (2016), at a given time t, a future time $m = t + \Delta t$, and a threshold c, the TPR and FPR of the predictive results can be defined as $\mathrm{TPR}_t(c) = Pr(\mathbf{1} - \boldsymbol{p}(m|t) \geq c|\boldsymbol{T} \leq m)$ and $\mathrm{FPR}_t(c) = Pr(\mathbf{1} - \boldsymbol{p}(m|t) \geq c|\boldsymbol{T} > m)$, where $\boldsymbol{p}(m|t) = \{p_i(m|t), i = 1, \ldots, N\}$ and $p_i(m|t)$ is illustrated in Section 2.2.5. The estimate of $\boldsymbol{p}(m|t)$ is denoted by $\hat{\boldsymbol{p}}(m|t)$, and the estimators of TPR and FPR are: $\widehat{TPR}_t(c) = \frac{\sum_{i=1}^{N}(1-\hat{p}_i(m|t))I(1-\hat{p}_i(m|t)\geq c)}{\sum_{i=1}^{N}(1-\hat{p}_i(m|t))}$ and $\widehat{FPR}_t(c) = \frac{\sum_{i=1}^{N}\hat{p}_i(m|t)I(1-\hat{p}_i(m|t)\geq c)}{\sum_{i=1}^{N}\hat{p}_i(m|t)}$. By definition, we have $\widehat{AUC}_t = \int \widehat{TPR}_t \left\{(\widehat{FPR}_t)^{-1}(u)\right\} du$, $\widehat{AARD}_t = \widehat{TPR}_t(\hat{\rho}) - \widehat{FPR}_t(\hat{\rho})$, and $\widehat{MRD}_t = \int_c \widehat{TPR}_t(c)dc - \int_c \widehat{FPR}_t(c)dc$, where $\hat{\rho} = \frac{\sum_{i=1}^{N}(1-\hat{p}_i(m|t))}{N}$ is the average risk in the study population at time m.

2.3 Simulation Studies

We conduct two simulation studies to validate the proposed QRJM. In the first simulation study, we assess the performance of the proposed Bayesian method in terms of bias and precision of the parameter estimates. In the second simulation study, we assess the predicted survival probability by comparing with the "gold standard" calculated based on the true (simulated) values of random effects and parameters.

2.3.1 Simulation Study I: Inferential Performance

In this simulation study, we consider different simulation scenarios where the random errors are generated from either a standard normal distribution or ALD distributions at different quantile τ. The simulated data are then fitted using our proposed QRJM (assuming ALD for the random errors) as well as the LMJM (assuming normality for the random errors).

We let the covariate vectors in Model (2.5) be $\boldsymbol{Z}_i(t) = (1, t)^\top, \boldsymbol{X}_i(t) = (1, x_{i1}, x_{i2} \cdot t)^\top,$ and $\boldsymbol{W}_i = (w_{i1}, w_{i2})^\top$ with covariates x_{i1}, x_{i2}, w_{i1} and w_{i2} being generated from independent standard normal distributions. We simulate the random effects, $\boldsymbol{u}_i$, from a bivariate normal distribution with mean vector $\mathbf{0}$, and standard deviations both equal to 0.3 and correlation coefficient equals to 0.16.

To simulate the survival time, we choose constant baseline hazard. We obtain event time T_i by inverting the survival function after generating n random values from the standard uniform distribution. We generate the censoring time C_i from $\text{Beta}(4, 1)$ to obtain a censoring proportion around 25%. The longitudinal data are simulated from either a standard normal distribution or a ALD with the location parameter being $\boldsymbol{\beta}^\top \boldsymbol{X}_i(t) + \boldsymbol{u}_i^\top \boldsymbol{Z}_i(t)$ and dispersion parameter being $\sigma = 1$. We keep a maximum of 6 observations for each subject, at follow-up time $t = (0, 0.25, 0.5, 0.75, 1, 3)$ respectively, after incorporating the time-to-event information.

We consider the following three scenarios in simulation study I:

1. Scenario 1: random errors follow the ALD with $\tau = 0.25$ (right-skewed);
2. Scenario 2: random errors follow the ALD with $\tau = 0.5$ (symmetric about 0 with heavy tails);
3. Scenario 3: random errors follow a standard normal distribution (symmetric about 0).

In each scenario, we simulate 200 datasets with $N = 600$ in each. Among the 600 subjects, we randomly select 500 subjects as the training dataset to build the model, and use the remaining 100 subjects as the validation dataset to make out-of-sample predictions.

We report bias, standard error (SE), mean squared error (MSE), and coverage probability (CP) for the QRJM and LMJM. Table 2.1 suggests that in Scenario 1, the true model (QRJM with $\tau = 0.25$) provides parameter estimates with very small biases and CP being close to the nominal level. In comparison, the QRJM with $\tau = 0.5$ provides reasonable estimates for most parameters, except the intercept β_0. The poor estimate of β_0 (large bias and CP far from 0.95) should not be surprising because of the incorrect specification of quantile τ. The LMJM results in very biased estimates and the CP being away from the nominal value 0.95 (especially in regression parameters $\boldsymbol{\beta}$ in the longitudinal model). In Scenario 2 (see Appendices Table 2.5) when data are symmetrical about 0 with heavier tails than the normal distribution, the LMJM still produces notably larger bias and lower CP as compared to the true model QRJM with $\tau = 0.5$. In Scenario 3 (see Appendices Table 2.6), median regression (the QRJM with $\tau = 0.5$) performs comparably to the true model LMJM, suggesting that the QRJM can provide reasonable estimates when normality assumption holds. For completion, we also consider a scenario where the random errors are simulated from ALD with $\tau = 0.75$ (left-skewed). The results (not presented) are very similar to Scenario 1, i.e., the true model

QRJM with $\tau = 0.75$ provides reasonable estimates while LMJM gives biased estimates and CP being far away from 0.95.

Table 2.1: Simulation results in Simulation study I Scenario 1 in which random errors are generated from ALD with $\tau = 0.25$.

	QRJM ($\tau = 0.25$)				QRJM ($\tau = 0.5$)				LMJM			
	Bias	SE	MSE	CP	Bias	SE	MSE	CP	Bias	SE	MSE	CP
Coefficients for longitudinal process												
β_0	-0.003	0.080	0.014	0.930	1.659	0.129	2.807	0.020	2.702	0.146	7.350	0.000
β_1	0.015	0.068	0.010	0.950	0.024	0.105	0.043	0.890	0.080	0.116	0.052	0.860
β_2	0.016	0.083	0.013	0.950	0.014	0.112	0.042	0.970	0.078	0.128	0.052	0.920
Coefficients for survival process												
γ_1	0.005	0.055	0.006	0.940	0.008	0.057	0.006	0.960	0.009	0.058	0.007	0.960
γ_2	0.006	0.055	0.006	0.930	0.010	0.056	0.007	0.910	0.010	0.058	0.007	0.940
α	-0.004	0.078	0.010	0.970	-0.051	0.119	0.070	0.930	-0.087	0.103	0.040	0.800

2.3.2 Simulation Study II: Predictive Performance

In this simulation study, we make predictions for 100 subjects in the validation dataset (out-of-sample predictions) in the three scenarios in Section2.3.1. For each subject, we use the simulated data, random effects and the true parameter values to calculate the true survival probability given by $\frac{S_i[m|\mathcal{M}_i(m,\boldsymbol{u}_i,\boldsymbol{\theta});\boldsymbol{\theta}]}{S_i[t|\mathcal{M}_i(t,\boldsymbol{u}_i,\boldsymbol{\theta});\boldsymbol{\theta}]}$ and we use it as the “gold standard”.

To assess the prediction validation (how well the models predict the survival probability), we use a Bland-Altman plot, a commonly used method to assess the agreement of the results from two measurement methods (Bland and Altman, 1986). To make the predictions “dynamic”, we choose different combinations of censoring time (i.e., t) and the prediction time interval (i.e., Δt) to mimic expected real-world time points based on our HD data set. Appendices Figure 2.5 gives an intuitive comparison of the predicted results with the “gold standard" among different models. Plots from the true model (Appendices Figure 2.4a) are horizontally spindle-shaped, suggesting that it is easier to predict a probability near 0 or 1 than the middle probability area near 0.5. Further, with an increase in Δt, there is more variation in the middle probability area, indicating that survival probability predictions for time points further into the future are less accurate than predictions for closer time points, as expected. Bland-Altman plots from the other models (QRJM with $\tau = 0.5$, Appendices Figure 2.4b and LMJM, Appendices Figure 2.4c) display systematically biased patterns in predictions.

Table 2.2 summarizes the comparison from the Bland-Altman plots for three chosen censoring time points ($t = 0.25, 0.5, 0.75$). With longer follow-up time, we tend to have more longitudinal measurements per subject and tend to have more precise predictions.On the other hand, longer follow-up time leads to fewer subjects left in the study due to event occurrence and censoring (presented as a percentage in the first column of Table 2.2), which results in higher variability in predictions. For example, in Table 2.2, at $t = 0.25$, there are 48.1% subjects remain and at $t = 0.5$ and 0.75, only 34.6% and 22.8%, respectively, of the subjects remain. As a result, we see comparable prediction results at $t = 0.25$ and 0.5 but worse predictive performance when t increases to 0.75. This is because the effect of additional longitudinal observations is "canceled out" by the variability from fewer subjects. Similarly, from the Bland-Altman plots, under the true model at the same censoring time t, an increase in Δt leads to larger bias and MSE (Appendices Figure 2.4a). Because Table 2.1 suggests that the predictions from other models (QRJM with $\tau = 0.5$ and LMJM) are systematically biased, their MSE and bias in Table 2.2 are much larger than the true model QRJM with $\tau = 0.25$. Prediction results for Scenario 3 can be found in Appendices Table 2.7 and Appendices Figure 2.5, in which QRJM with $\tau = 0.5$ performs comparably well as the true model when random errors are generated from standard normal distribution.

In summary, when data are simulated from ALD with specific skewness τ, the best predictions of survival probability are obtained using the QRJM model with the exact quantile that generated the outcome data. Instead, LMJM results in systematically biased predictions when data are skewed. When random errors are standard normally distributed, predictions from QRJM with $\tau = 0.5$ are comparable with those from the true model.

Table 2.2: Simulation study: MSE and bias of the difference between predicted survival probability and the gold standard (Scenario 1).

		QRJM ($\tau = 0.25$)		QRJM ($\tau = 0.5$)		LMJM	
t	Δt	MSE	Bias	MSE	Bias	MSE	Bias
0.25	0.25	0.006	0.009	0.137	−0.330	0.244	−0.462
	1	0.010	0.007	0.111	−0.267	0.177	−0.343
(subjects left: 48.1%)	2	0.012	0.003	0.083	−0.197	0.126	−0.249
	3	0.013	0.000	0.072	−0.168	0.107	−0.210
0.5	0.25	0.007	0.009	0.130	−0.317	0.219	−0.439
	1	0.015	0.000	0.144	−0.321	0.221	−0.408
(subjects left: 34.6%)	2	0.017	−0.015	0.121	−0.259	0.174	−0.319
	3	0.018	−0.023	0.109	−0.228	0.153	−0.278
0.75	0.25	0.009	0.005	0.125	−0.301	0.189	−0.401
	1	0.023	−0.007	0.174	−0.356	0.253	−0.447
(subjects left: 22.8%)	2	0.025	−0.033	0.159	−0.310	0.218	−0.375
	3	0.027	−0.046	0.148	−0.282	0.197	−0.336

2.4 Application

2.4.1 The Predictors of Huntington's Disease (PREDICT-HD) Study

The motivating PREDICT-HD study is an observational study that aims to identify the earliest signs of HD onset so that future HD drug trials can be targeted toward treatment that may slow the progression of the disease, or prevent it altogether (Paulsen et al., 2008). Briefly, HD is known to be caused by the mutation of the first exon of the Huntington (HTT) gene, where expansion of the cytosine-adenine-guanine (CAG) is observed for HD patients.

The study recruited participants from 33 medical centers in six countries (i.e., USA, Canada, Germany, Australia, Spain, and UK) starting from August, 2002. Qualified participants were healthy pre-HD people without any symptoms of HD; i.e., who have not had a motor diagnosis of HD based on the Unified Huntington Disease Rating Scale (UHDRS) and had more than 35 HTT CAG repeats. Detailed inclusion and exclusion criteria can be found in Paulsen et al. (2006). Baseline demographic information such as age, gender, education years, as well as clinical variables such as CAG repeat length and Beck Depression Inventory were recorded at enrollment. A total of 40 longitudinal biomarkers from five domains (i.e., motor, cognitive, psychiatric, functional, and imaging) were recorded. The data used in the current study contains 1078 individuals enrolled until July, 2014. Among those 1078 participants, 64% are female and the mean age at baseline is 39.8 years (SD=10.39, range 18.1-83.7), education is 14.5 years (SD=2.60, range 8.0-20.0), and number of CAG repeats is 42.5 (SD=2.69, range 12-62). In this study, the time variable is defined as months since enrollment. The average follow-up time is 61.2 months (SD=39.6; range 0.12-144.0) and 959 (89%) participants have data for at least two years. The survival event of interest is motor diagnosis of HD since enrollment, which is defined as having a diagnostic confidence level (DCL) score of 4 (the highest score)(Paulsen et al., 2014). During the study follow-up, 225

(21%) events (HD onsets) were observed.

We select five longitudinal biomarkers, one from each domain, that are known to be strongly predictive of HD onset (Paulsen et al., 2014) (see Appendices Section 3 for variable definitions). In the exploratory data analysis, we plot the scatter plots (with loess curves) and the kernel density plots of those selected variables (see Figure 2.1 and Appendices Figure 4). The distributions of TMS, FrBe Executive Subscale (FES), and Total Functional Capacity (TFC) all clearly deviate from normality, indicating a strong indication that the LMM may not be appropriate for these outcomes. From the loess curves, we observe the general direction of association between each longitudinal biomarker and the risk of HD onset. For example, putamen volume (intracranial-corrected) decreases over time and the brain atrophy from decreased putamen volume is associated with increased hazard of HD onset Paulsen et al. (2014). To compare the longitudinal results on the same scale, we have standardized all longitudinal biomarkers.

We consider the following joint model for our data analysis:

$$\begin{cases} y_i(t) = m_{it} + \varepsilon_{it} = \beta_0 + \beta_1 t + \beta_2 age_{0i} + u_{i1} + u_{i2}t + \varepsilon_i(t), \varepsilon_i(t) \sim ALD(0, \sigma, \tau) \\ h(T_i|\mathcal{M}_{iT_i}; \boldsymbol{\gamma}, \alpha) = \sum_{k=1}^{3} \lambda_k I_k(T_i) \exp(\gamma_1 education_i + \gamma_2 I_{male_i} + \alpha m_i(T_i)) \end{cases}$$

where $y_i(t)$ represents one of the five selected longitudinal biomarkers and age_i is the baseline age of subject i. In the survival sub-model, we specify a piecewise constant baseline hazard function with three time intervals, where λ_k is the hazard rate for time interval $[t_k, t_{k+1})$ and $I_k(t) = 1$ if $t \in [t_k, t_{k+1})$ and 0 otherwise.We have explored different numbers of time intervals and pieces and have obtained very similar results.

We randomly split the 1078 study participants into two datasets: a training dataset of 800 participants to draw statistical inference for the unknown parameters and a validation dataset of 278 participants for predictions of HD-free probability. In the model fitting step, two chains are initiated with diverse initial values and the chains are considered to converge if the potential scale reduction factors (PSRF) (Brooks and Gelman, 1998) for all parameters are below 1.1. In the prediction step we choose different censoring time (t) and prediction time window (Δt) combinations (similar to those chosen in the simulation study). Predictive accuracy for the test cohort is summarized using AUC, AARD ,and MRD defined in Section 2.2.6 for each $(t, \Delta t)$ combination as well as for each quantile in the QRJM. We also calculate the AUCs from the LMJM and compare them with those from our QRJM model.

2.4.2 Inference Results for PREDICT-HD data

Table 2.3 (for the outcome TMS) and Appendices Table 4 (for the outcomes putamen, stroop word, FES, and TFC) present the inference results from the QRJMs at $\tau = 0.25$, 0.50, and 0.75. In the longitudinal process, the coefficient for time quantifies the change rate of a specific longitudinal biomarker at a fixed quantile. For example, Table 2.3 suggests that the time effect is 0.019 (95% CI: 0.015, 0.023) for TMS at $\tau = 0.25$, indicating one month increase in time is associated with 0.019 unit increase of TMS (when $\tau = 0.25$). The positive time coefficients for TMS at all quantiles indicate that TMS increases (deteriorates) over time, which is consistent with the loess curves in Figure 2.1. The magnitude of the time coefficient is similar at different quantiles for TMS, putamen, stroop word, and FES, indicating comparable progression in each of these outcomes at those quantiles. However, some variability is observed in TFC, where the time effect is highest at lower quantiles of TFC and becomes smaller at higher quantiles. Because lower TFC is associated with higher risk of HD onset, and these data indicate that the health of those with lower TFC deteriorates faster, special targeted care may be needed for those individuals in the lower quantile of TFC, in order to reduce their risk of HD onset.

Table 2.3 also suggests that higher baseline age is associated with higher (worse) TMS and one year increase in baseline age is associated with 0.005 (95% CI: 0.001, 0.010) units increase in TMS (when $\tau = 0.5$). In the time-to-event process, across all three quantiles, education has protective effect on HD onset, i.e., those with more education years tend to have a lower risk of HD onset. Outcome TMS is strongly predictive of HD onset at all quantiles. Specifically, an unit increase in TMS increases the risk of HD onset by 4.600 (exp(1.526), 95% CI 3.747-5.726) times at $\tau = 0.25$, by 3.669 (95% CI 3.152-4.302) times at $\tau = 0.5$, and by 2.945 (95% CI 2.633-2.294) times at $\tau = 0.75$, respectively. Similarly, worse outcomes in putamen, stroop word task, FES, and TFC are also strongly associated with higher risk of HD onset across all quantiles.

Table 2.3: PREDICT-HD data analysis: Parameter estimation and 95% credible interval (in parenthesis) from QRJM at three different quantiles with TMS being the longitudinal biomarker.

	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
For longitudinal TMS process			
Intercept	-0.760 (-0.903, -0.628)	-0.525 (-0.699, -0.359)	-0.249 (-0.469, -0.035)
Time (months)	0.019 (0.015, 0.023)	0.020 (0.016, 0.024)	0.022 (0.018, 0.026)
Age	0.004 (0.001, 0.008)	0.005 (0.001, 0.010)	0.006 (0.001, 0.012)
For time to HD onset process			
Education (years)	-0.083 (-0.115, -0.052)	-0.112 (-0.142, -0.082)	-0.128 (-0.157, -0.101)
Male	0.317 (-0.037, 0.654)	0.360 (-0.020, 0.708)	0.317 (-0.010, 0.647)
association	1.526 (1.321, 1.745)	1.300 (1.148, 1.459)	1.080 (0.968, 1.192)

2.4.3 Dynamic Predictions of HD Risk

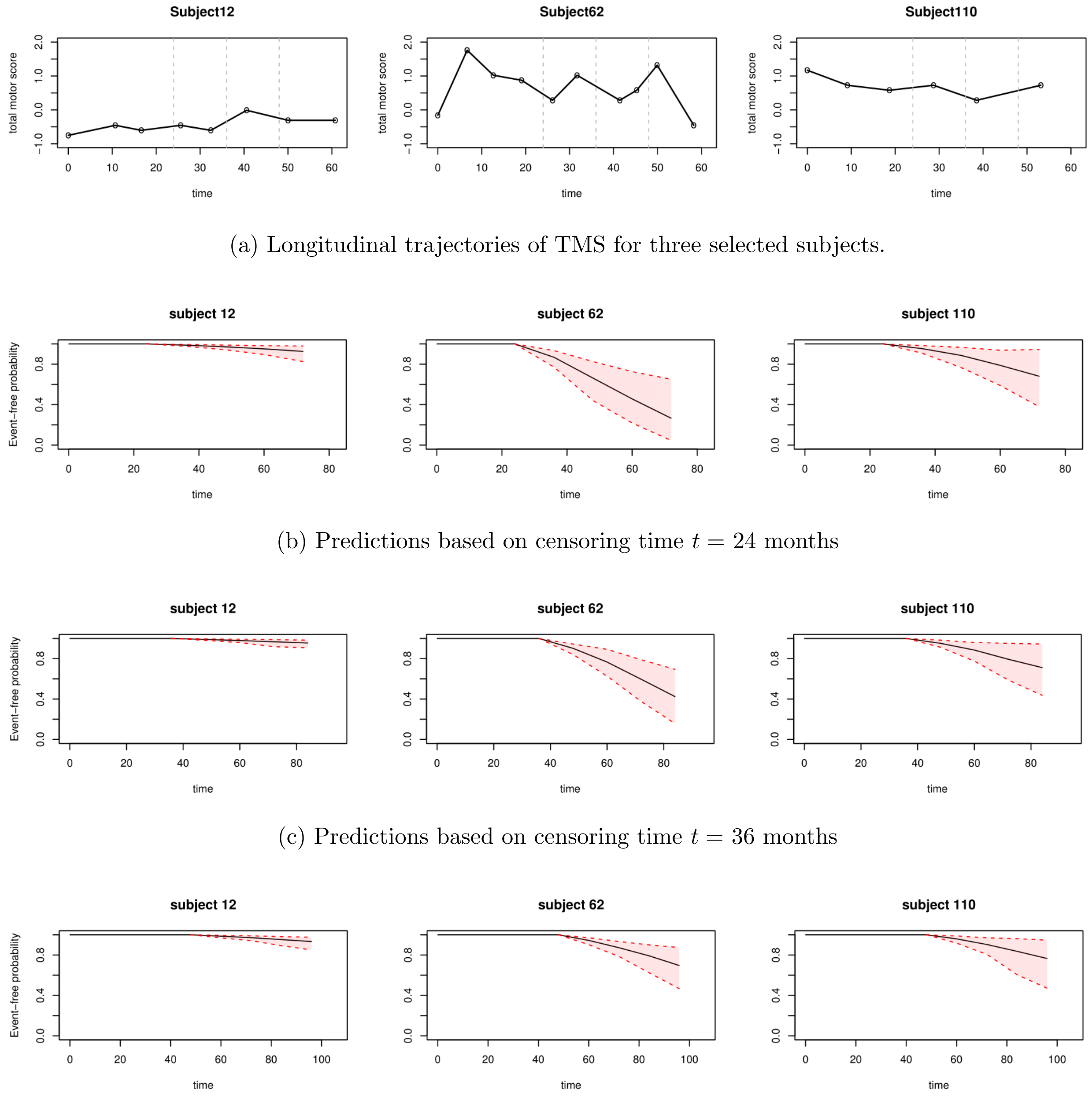
To evaluate the predictive accuracy of the model, we make out-of-sample predictions using the test dataset. Results are summarized in Table 2.4 with a column of the AUC from prediction using the LMJM. There are several interesting “patterns” observed in Table 2.4.

First, for a fixed censoring time t, all summary statistics increase with the increase in Δt (this is sensible in that HD-free probability decreases with time, thus it is easier for the model to differentiate those who will versus those who will not experience the event at later time points). Second, for the same $t + \Delta t$ value, we see an increase in all three statistics from $t = 12$ months to 24 months (e.g., $t = 12, \Delta t = 24$ v.s. $t = 24, \Delta t = 12$) followed by comparable statistics (a slight decrease in MRD) from $t = 24$ months to 48 months (e.g., $t = 24, \Delta t = 36$ v.s. $t = 48, \Delta t = 12$). The improvement in predictive performance can be due to the additional longitudinal data with increasing censoring time t. However, longer follow-up time leads to smaller sample size due to event occurrence and censoring. Similar findings are reported in the simulation. Third, compared with the LMJM, the QRJM has better prediction at some quantiles, but not all, because some quantiles of the outcome can be more informative in predicting future survival outcome than the conditional mean while some are not. There are also a few cases where the QRJM has better prediction at all quantiles ($t = 48$ at all Δt) and vice versa ($t = 24$, $\Delta t = 24$).

Table 2.4: PREDICT-HD data analysis: AUC, AARD and MRD of the predictions of HD-free probability from QRJM and AUC from LMJM with TMS as the longitudinal biomarker.

t	Δt	AUC (τ)			AARD (τ)			MRD (τ)			AUC (LMJM)
(month)		0.25	0.50	0.75	0.25	0.50	0.75	0.25	0.50	0.75	
12	12	0.647	0.683	0.738	0.213	0.261	0.356	0.010	0.020	0.059	0.679
	24	0.668	0.702	0.753	0.244	0.290	0.379	0.028	0.054	0.128	0.695
	36	0.685	0.714	0.760	0.273	0.311	0.391	0.054	0.091	0.170	0.693
24	12	0.836	0.857	0.864	0.539	0.575	0.577	0.168	0.218	0.285	0.855
	24	0.852	0.872	0.873	0.566	0.598	0.583	0.285	0.361	0.404	0.878
	36	0.866	0.877	0.872	0.581	0.599	0.575	0.368	0.420	0.430	0.836
48	12	0.875	0.878	0.868	0.583	0.598	0.589	0.326	0.320	0.303	0.669
	24	0.875	0.883	0.874	0.578	0.602	0.598	0.390	0.401	0.379	0.769
	36	0.877	0.887	0.879	0.589	0.614	0.599	0.417	0.439	0.417	0.774

We select three individuals (with IDs 12, 63, and 110) with different TMS trajectories (Figure 2.2a) to illustrate how our method provides subject-specific dynamic predictions of HD-free survival. For each individual, predictions are calculated and then updated based on increasing number of longitudinal measurements. In Figure 2.2b, we have the longitudinal measures up to 24 months ($t = 24$) for each subject, and predictions are made for Δt being 12, 24, 36, and 48 months respectively. We then increase the follow-up time to 36 and 48 months ($t = 36$ and $t = 48$) and update the predictions using the same Δt values. Updated prediction results are displayed in Figure 2.2c ($t = 36$) and Figure 2.2d ($t = 48$), respectively. The plots suggest that participants with lower and stable TMS have much higher HD-free probability (i.e., lower predicted risk of HD onset, subject 12 v.s. subjects 62 and 110). With longer follow-up times and more longitudinal measurements, predicted HD-free probabilities are updated based on the entire longitudinal history and are more accurate as indicated by the narrower point-wise 95% credible intervals as one looks at predictions from censoring time $t = 24$ months down to $t = 48$ months.



(a) Longitudinal trajectories of TMS for three selected subjects.

(b) Predictions based on censoring time $t = 24$ months

(c) Predictions based on censoring time $t = 36$ months

(d) Predictions based on censoring time $t = 48$ months

Figure 2.2: PREDICT-HD data analysis: Longitudinal trajectories and dynamic predictions of HD-free probabilities with 95% pointwise credible interval from QRJM at $\tau = 0.5$ for selected subjects.

2.5 Discussion

In our application of the LMJM (a linear mixed sub-model for the longitudinal process and a PH sub-model for the survival process) to Huntington's Disease, there are two limitations: first, the normality assumption of the random errors in the linear mixed model (LMM) was not realistic, and no obvious transformation of the longitudinal outcome to produce residual normality was applicable. This limitation is confirmed by our simulation studies where the LMJM tends to provide biased estimates to model parameters when the longitudinal data are non-normal. Consequently, predictive accuracy of the LMJM is negatively impacted. Second, the LMJM only models the conditional mean of the outcome. However, in our (and other) clinical research application(s), it may be more clinically relevant to consider the tails of the outcome distribution, e.g., the upper tail of TMS is at higher risk of developing HD.

Our proposed quantile regression joint models (QRJM) uses a linear quantile mixed model (LQMM) for the longitudinal process, improves both inference and the ability to make accurate dynamic predictions. The quantile-based estimators are more robust against skewness in the data, and as a result, our approach provides the flexibility to use median or quantile regression instead of mean regression when outliers and skewness are present in the longitudinal process. Moreover, the QRJM provides quantile-specific parameter estimates at a set of different quantiles and the researchers can choose the quantiles of interest and the corresponding inference results. The simulation studies and data application suggest that the QRJM not only inherits the good properties of an LMJM, but adds flexibility to the modeling procedure.

In this work, we develop a Bayesian algorithm to fit the proposed QRJM model and make dynamic predictions using the location-scale representation of the asymmetric Laplace distribution (ALD) for the longitudinal quantile regression. The Bayesian algorithm, which is straightforwardly implemented in JAGS software, uses a piecewise constant baseline hazard function in the survival sub-model. However, other functional forms can also be considered and the integration of the hazard function can be approximated using numerical integration such as Simpson's rule. In the real data application, we illustrate the flexibility of the QRJM and its advantages over the LMJM by jointly modeling the risk of developing HD and five commonly used early predictors of the disease. The QRJM is able to provide more insight into the disease progression and the association between the two disease processes in terms of various quantile-based estimations and dynamic predictions.

The novel application of our proposed QRJM in making personalized dynamic predictions of survival probability finds practical importance in many clinical applications. Event prediction using commonly collected biomarkers can provide clinicians with continuously updated "disease progression" information potentially allowing them to make appropriately timed intervention decisions for individual subjects. Subject-specific dynamic predictions models play an important role in the ongoing transformation of traditional medicine to patient-centric care. While in traditional medicine, treatments prescribed are guided by population-based experience, i.e., the average treatment effect in the entire population, such "one-size-fits-all" approaches are criticized for resulting in low efficacy and high adverse drug reactions in many clinical studies. Utilizing the dynamic predictions based on the proposed QRJM framework, we obtain subject-specific predictions of event risk that would actually allow physicians to target or tailor medical treatment based on a specific patient profile.

In summary, the QRJM is a good alternative to the LMJM when either the normality assumption of the errors term is concerned or the conditional quantiles are more relevant to research question. When the interest lies in the prediction of future survival probability, the "best" quantile may be chosen based on the predictive accuracy criteria. Other model selection methods or methods, e.g., Bayesian model averaging, to incorporate multiple regression results from different quantiles into a single prediction solution can be a potential direction for future work. However, in our opinion, the choice of specific quantile(s) should be more clinically oriented rather than be a statistical task.

Appendices

A Distribution Comparison

- Laplace distribution (LD) with location at 0 and scale parameter equals 1 is symmetrical about 0. It has heavier tail compared with standard normal distribution.
- Asymmetric Laplace distribution (ALD) is either positively or negatively skewed and the direction and degree of skewness are control by the skewness parameter $\tau \in [0, 1]$.

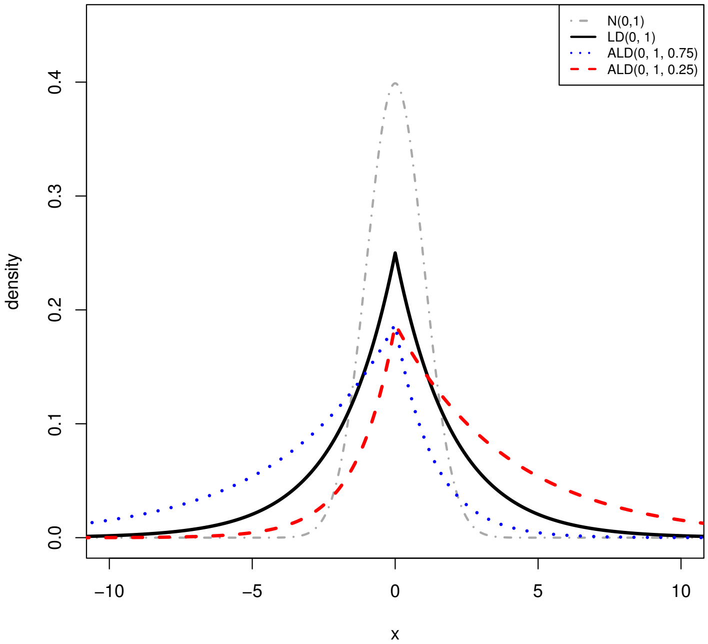


Figure 2.3: A comparison of normal, Laplace and asymmetric Laplace distributions.

B Additional Simulation Results

Table 2.5: Simulation result in Simulation study I Scenario 2 in which random errors are generated from ALD with $\tau = 0.5$.

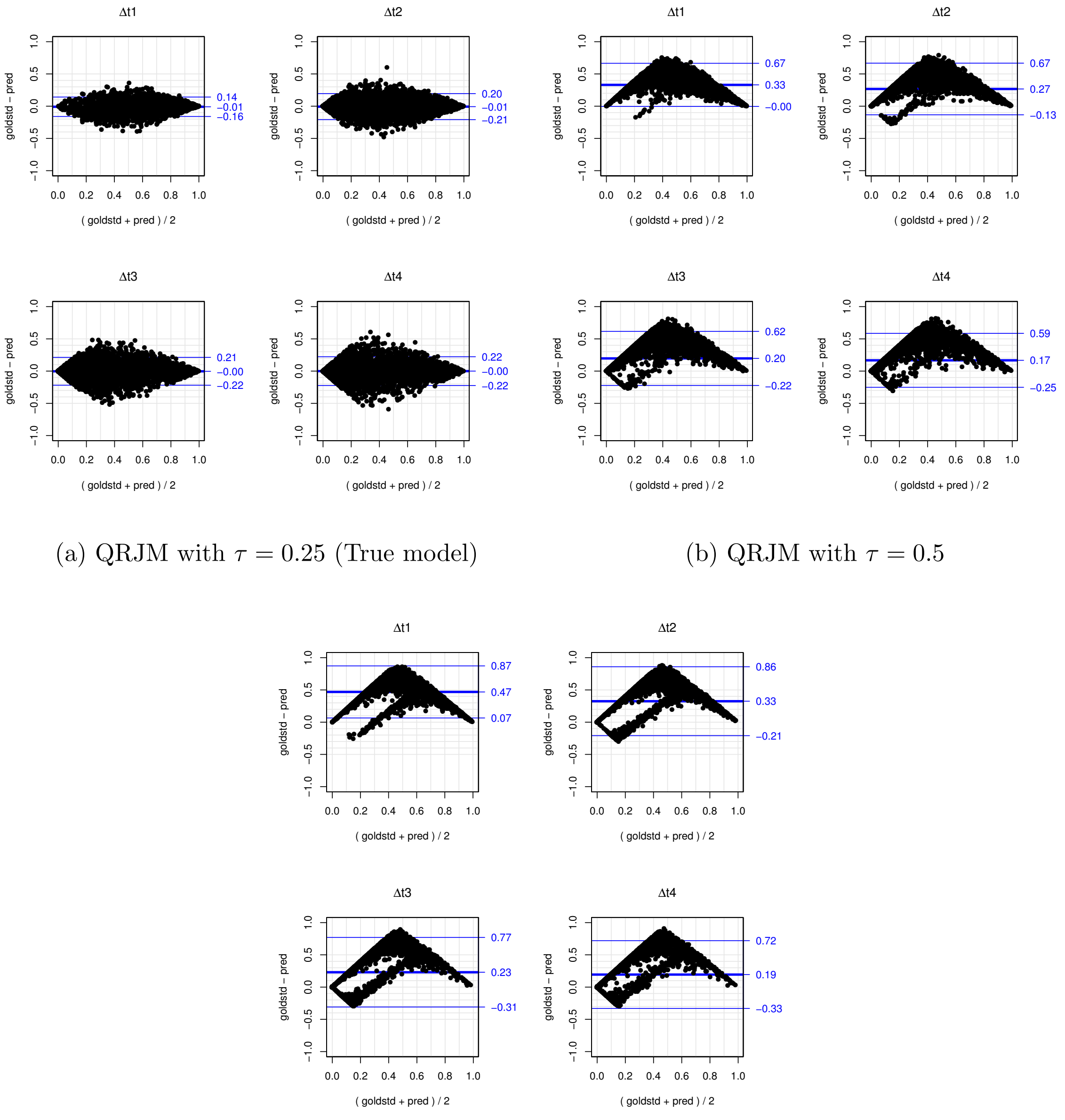
	QRJM ($\tau = 0.5$)				LMJM			
	Bias	SE	MSE	CP	Bias	SE	MSE	CP
Coefficients for longitudinal process								
β_0	-0.006	0.069	0.009	0.960	0.013	0.093	0.017	0.960
β_1	0.008	0.060	0.008	0.900	0.018	0.079	0.018	0.880
β_2	0.014	0.075	0.011	0.950	0.031	0.093	0.021	0.940
Coefficients for survival process								
γ_1	0.008	0.055	0.006	0.940	0.014	0.058	0.007	0.950
γ_2	0.007	0.055	0.007	0.950	0.013	0.057	0.007	0.950
α	-0.001	0.071	0.011	0.930	-0.028	0.101	0.086	0.920

Table 2.6: Simulation result in Simulation study I Scenario 3 in which random errors are generated from $\mathcal{N}(0, 1)$.

	QRJM ($\tau = 0.5$)				LMJM			
	Bias	SE	MSE	CP	Bias	SE	MSE	CP
Coefficients for longitudinal process								
β_0	0.015	0.037	0.003	0.950	0.000	0.035	0.002	0.980
β_1	0.004	0.034	0.002	0.960	-0.003	0.033	0.002	0.950
β_2	0.013	0.050	0.005	0.950	0.006	0.049	0.005	0.950
Coefficients for survival process								
γ_1	0.008	0.055	0.006	0.920	0.003	0.054	0.006	0.900
γ_2	0.015	0.055	0.007	0.920	0.010	0.054	0.006	0.920
α	-0.013	0.055	0.006	0.950	0.007	0.055	0.006	0.950

Table 2.7: Prediction results in Simulation study II Scenario 3 in which random errors are generated from ALD with $\tau = 0.5$.

		QRJM ($\tau = 0.5$)		LMJM	
t	Δt	MSE	Bias	MSE	Bias
0.25 (subjects left: 47.87%)	0.25	0.005	0.005	0.005	−0.001
	1	0.009	0.005	0.009	0.000
	2	0.011	0.003	0.010	0.000
	3	0.011	0.002	0.011	0.001
0.5 (subjects left: 34.78%)	0.25	0.004	0.005	0.003	−0.002
	1	0.009	0.004	0.008	−0.002
	2	0.012	0.002	0.011	−0.001
	3	0.013	0.001	0.012	−0.001
0.75 (subjects left: 27.71%)	0.25	0.002	0.004	0.002	−0.002
	1	0.009	0.005	0.007	−0.002
	2	0.012	0.002	0.011	−0.003
	3	0.014	0.001	0.012	−0.002

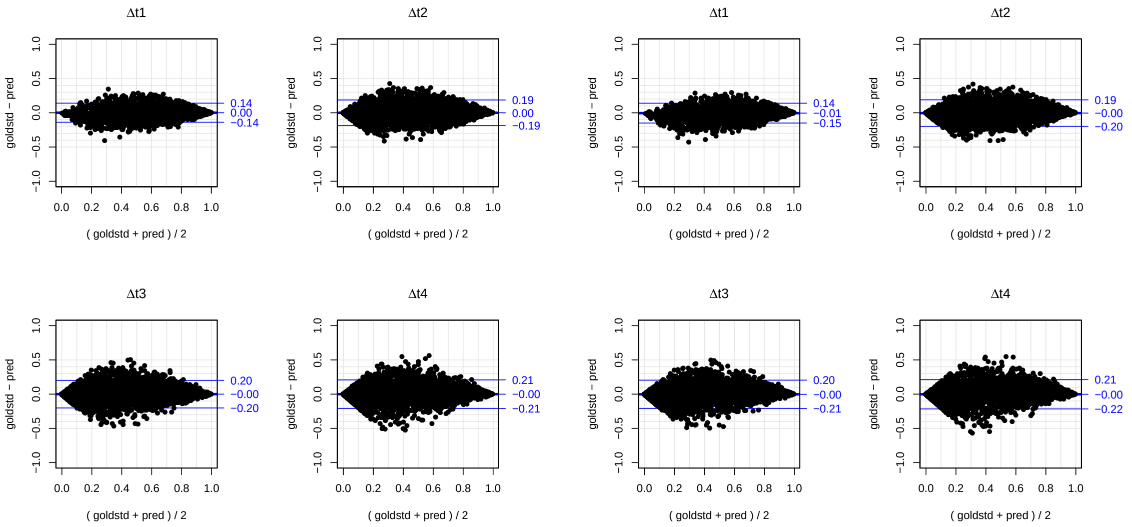


(a) QRJM with $\tau = 0.25$ (True model)

(b) QRJM with $\tau = 0.5$

(c) LMJM

Figure 2.4: Prediction results in Simulation study II Scenario 1: Bland-Altman plot (bias and 95% limits of agreement) of gold standard versus model predictions at four prediction time intervals ($\Delta t_1 < \Delta t_2 < \Delta t_3 < \Delta t_4$).



(a) LMJM (True model)

(b) QRJM with $\tau = 0.5$

Figure 2.5: Prediction results in Simulation study II Scenario 3: Bland-Altman plot (bias and 95% limits of agreement) of gold standard versus model predictions based on the first two longitudinal observations and four different prediction time intervals ($\Delta t_1 < \Delta t_2 < \Delta t_3 < \Delta t_4$).

C Variable Definitions

- Total motor score: Standardized ratings of oculomotor function, dysarthria, chorea, dystonia, gait and postural stability based on the Unified Huntington Disease Rating Scale (UHDRS).

- Putamen: the volume of putamen, which a round structure located at the base of the forebrain. The atrophy of the putamen is related with impaired movements regulation and learning ability.

- Stroop word: Stroop Color and Word Test score. The test consists of three 45-second trials, including two trials measure basic attention and process speed and a third trail tests the ability to identifying the name of a color (e.g., “blue”, “green”, or “red”) that is printed in a color not denoted by the name.

- FrBe executive subscale: Frontal System Behavioral Scale – executive subscale – companion rating scale. Part of a 46-item behavior rating scale on abstraction, problem solving, and hypothesis generation as rated by a companion focusing on dorsolateral prefrontal circuitry.

- Total functional capacity: A list of independent and common daily tasks that can be accomplished bas on the UHDRS.

D LOESS Plots of Selected Longitudinal Variables

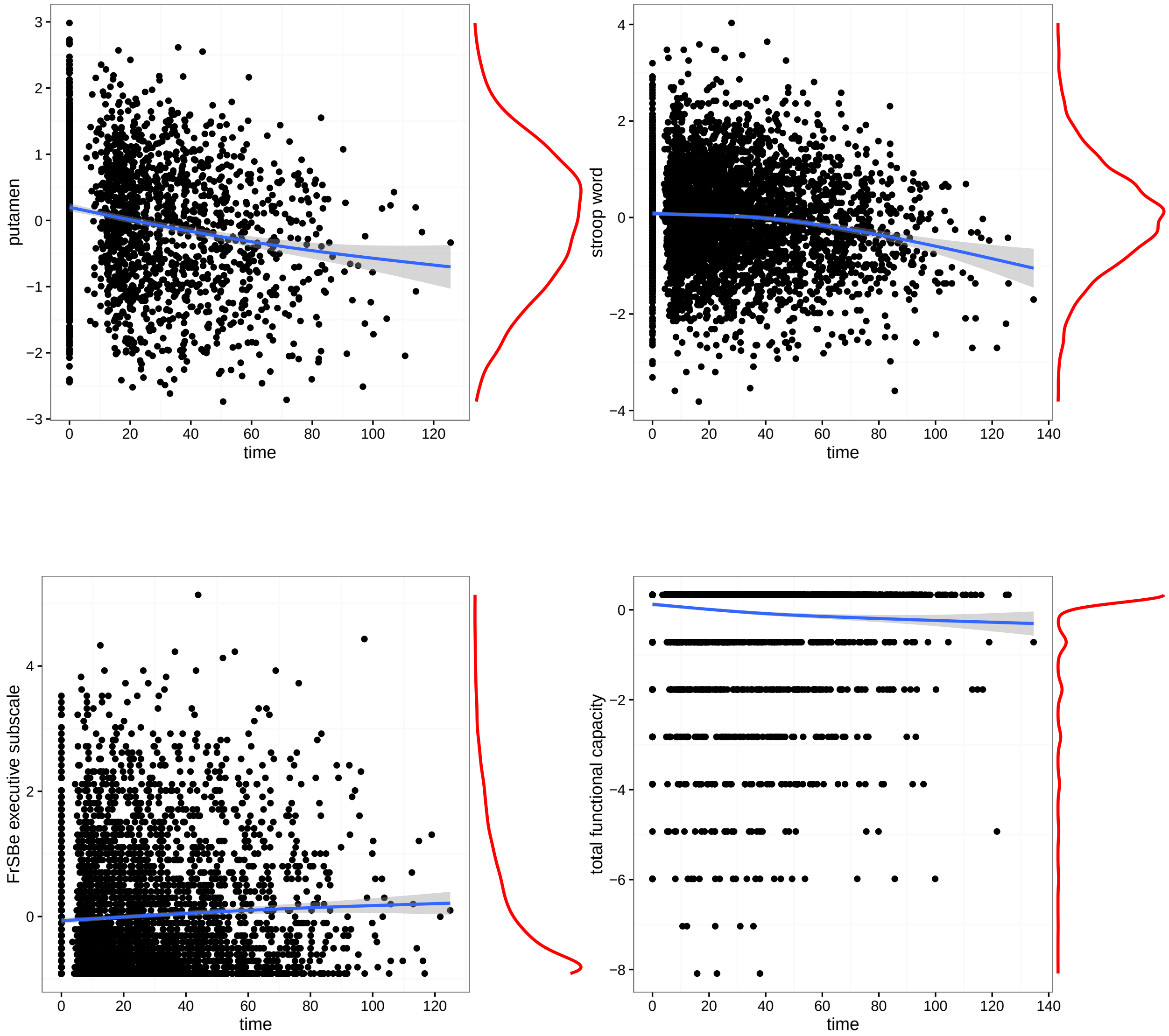


Figure 2.6: Scatter plot with LOESS curve for putamen (top left), stroop word (top right), FrBe executive subscale (bottom left) and total functional capacity (bottom right) from the study population.

3 Additional Data Analysis Results

Table 3.1: PREDICT-HD data analysis: Parameter estimation and 95% credible interval (in parenthesis) from QRJM at three quantiles (0.25, 0.5, and 0.75) with putamen, stroop word, FrBe executive subscale (FES), and total functional capacity (TFC) being the longitudinal biomarkers.

	putamen			stroop word		
	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
For longitudinal process						
int.	1.045	1.142	0.1.243	0.123	0.370	0.625
	(0.830, 1.266)	(0.864, 1.439)	(1.003, 1.511)	(-0.136, 0.378)	(0.114, 0.640)	(0.382, 0.857)
time (month)	-0.014	-0.014	-0.014	-0.007	-0.007	-0.007
	(-0.018, -0.010)	(-0.017, -0.010)	(-0.017, -0.010)	(-0.010, -0.004)	(-0.010, -0.003)	(-0.010, -0.004)
age_0	-0.023	-0.023	-0.017	-0.006	-0.007	-0.007
	(-0.028, -0.018)	(-0.030, -0.017)	(-0.029, -0.017)	(-0.013, 0.000)	(-0.013, 0.000)	(-0.013, -0.001)
For time to HD onset process						
assoct.	-1.038	-1.039	-1.045	-0.696	-0.709	-0.719
	(-1.213, -0.861)	(-1.217, -0.860)	(-1.224, -0.862)	(-0.832, -0.558)	(-0.846, -0.573)	(-0.853, -0.581)
eduyr	-0.060	-0.055	-0.050	-0.080	-0.069	-0.059
	(-0.094, -0.027)	(-0.093, -0.021)	(-0.090, -0.015)	(-0.106, -0.053)	(-0.097, -0.043)	(-0.088, -0.032)
male	-0.021	-0.020	-0.017	-0.173	-0.166	-0.157
	(-0.347, 0.283)	(-0.336, 0.290)	(-0.330, 0.303)	(-0.459, 0.111)	(-0.458, 0.111)	(-0.449, 0.124)

	FrBe executive subscale			total functional capacity		
	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
For longitudinal process						
int.	-0.533	-0.244	0.138	0.277	0.333	0.362
	(-0.731, -0.328)	(-0.493, -0.021)	(0.261, 0.494)	(0.218, 0.355)	(0.293, 0.372)	(0.330, 0.394)
time (month)	0.007	0.006	0.006	-0.028	-0.016	-0.007
	(0.003, 0.011)	(0.003, 0.010)	(0.002, 0.010)	(-0.033, -0.024)	(-0.020, -0.012)	(-0.012, -0.004)
age_0	0.002	0.002	0.001	-0.001	-0.000	0.000
	(-0.003, 0.007)	(-0.003, 0.008)	(-0.006, 0.008)	(-0.002, 0.001)	(-0.001, -0.001)	(-0.001, 0.001)
For time to HD onset process						
assoct.	0.524	0.434	0.376	-0.374	-0.478	-0.488
	(0.378, 0.667)	(0.306, 0.435)	(0.261, 0.494)	(-0.438,-0.305)	(-0.569, -0.389)	(-0.596, -0.378)
eduyr	-0.077	-0.082	-0.091	-0.097	-0.093	-0.093
	(-0.116, -0.041)	(-0.119, -0.052)	(-0.125, -0.060)	(-0.135, -0.064)	(-0.134, -0.058)	(-0.138, -0.054)
male	-0.263	-0.301	-0.291	0.104	0.057	0.031
	(-0.560, 0.051)	(-0.611, 0.004)	(-0.602, 0.006)	(-0.207, 0.405)	(-0.260, 0.354)	(-0.286, 0.329)

A JAGS Model File

JAGS model file to fit QRJM in simulation study.

```
model{
      k1 <- (1-2*qt)/(qt*(1-qt))
      k2 <- 2/(qt*(1-qt))
      for (i in 1:I){
        # prior for random effects
        u[i, 1:2] ~ dmnorm(zero[], precision[,])
        # longitudinal process
        for (j in 1:J[i]){
            er[i,j] ~ dexp(sigma)
            mu[i,j] <- beta + u[i,1] + u[i,2]*t[j] + time[1]*X[i,1]
                    + time[2]*X[i,2]*t[j] + k1*er[i,j]
            prec[i,j] <- sigma/(k2*er[i,j])
            y[i,j] ~ dnorm(mu[i,j], prec[i,j])
        } #end of j loop
        # time-to-event process
        A[i] <- assoct.*u[i,2] + assoct.*time[2]*X[i,2]
        B[i] <- assoct.*(time[1]*X[i,1] + u[i,1] + beta) + inprod(gamma, W[i,])
        S[i] <- exp(-c*exp(B[i])*(exp(A[i]*Ti[i])-1)/A[i])
        h[i] <- c*exp(inprod(gamma, W[i,]) + assoct.*(beta +
            time[1]*X[i,1] + time[2]*X[i,2]*Ti[i] + u[i,1] + u[i,2]*Ti[i]))
        L[i] <- pow(h[i], event[i])*S[i]/1E+08
        phi[i] <- -log(L[i])
        zeros[i] ~ dpois(phi[i])
    }#end of i loop
    precision[1:2, 1:2] ~ dwish(Omega[,], 3)
    Sigma[1:2,1:2] <- inverse(precision[,])
    Omega[1,1] <- 1
    Omega[2,2] <- 1
    Omega[1,2] <- 0
    Omega[2,1] <- 0
    # priors for other parameters
    assoct. ~ dnorm(0, 0.001)
    int. ~ dnorm(0, 0.001)
    time[1] ~ dnorm(0, 0.001)
    time[2] ~ dnorm(0, 0.001)
    gamma[1] ~ dnorm(0, 0.001)
    gamma[2] ~ dnorm(0, 0.001)
    sigma ~ dgamma(0.001, 0.001)
    c ~ dunif(0.01, 10)
}
```

References

Bland, J. M. and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455.

Farcomeni, A. and Viviani, S. (2015). Longitudinal quantile regression in the presence of informative dropout through longitudinal–survival joint modeling. *Statistics in Medicine*, 34(7):1199–1213.

Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics*, 8(1):140–154.

Geraci, M. and Bottai, M. (2014). Linear quantile mixed models. *Statistics and Computing*, 24(3):461–479.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, 91(1):74–89.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R. and Basset, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Koenker, R. and Hallock, K. (2001). Quantile regression: An introduction. *Journal of Economic Perspectives*, 15(4):43–56.

Koenker, R. and Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448):1296–1310.

Kotz, S., Kozubowski, T., and Podgorski, K. (2012). *The Laplace Distribution And Generalizations: A Revisit with Applications to Communications, Economics, Engineering, And Finance*. Springer Science & Business Media.

Kozumi, H. and Kobayashi, G. (2011). Gibbs sampling methods for bayesian quantile regression. *Journal of Statistical Computation and Simulation*, 81(11):1565–1578.

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Paulsen, J. S., Hayden, M., Stout, J. C., Langbehn, D. R., Aylward, E., Ross, C. A., Guttman, M., Nance, M., Kieburtz, K., Oakes, D., et al. (2006). Preparing for preventive clinical trials: the predict-hd study. *Archives of Neurology*, 63(6):883–890.

Paulsen, J. S., Langbehn, D. R., Stout, J. C., Aylward, E., Ross, C. A., Nance, M., Guttman, M., Johnson, S., MacDonald, M., Beglinger, L. J., et al. (2008). Detection of huntington's disease decades before diagnosis: the predict-hd study. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(8):874–880.

Paulsen, J. S., Long, J. D., Ross, C. A., Harrington, D. L., Erwin, C. J., Williams, J. K., Westervelt, H. J., Johnson, H. J., Aylward, E. H., Zhang, Y., et al. (2014). Prediction of manifest huntington's disease with clinical and imaging measures: a prospective observational study. *The Lancet Neurology*, 13(12):1193–1201.

Pepe, M. S., Feng, Z., and Gu, J. W. (2008). Comments on 'evaluating the added predictive ability of a new marker: From area under the roc curve to reclassification and beyond'by mj pencina et al., statistics in medicine. *Statistics in Medicine*, 27(2):173–181.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, page 125. Vienna.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829.

Rizopoulos, D., Hatfield, L. A., Carlin, B. P., and Takkenberg, J. J. M. (2014). Combining dynamic predictions from joint models for longitudinal and time-to-event data using bayesian model averaging. *Journal of the American Statistical Association*, 109(508):1385–1397.

Self, S. and Pawitan, Y. (1992). Modeling a marker of disease progression and onset of disease. In *AIDS Epidemiology*, pages 231–255. Springer.

Taylor, J. M. G., Park, Y., Ankerst, D. P., Proust-Lima, C., Williams, S., Kestin, L., Bae, K., Pickles, T., and Sandler, H. (2013). Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics*, 69(1):206–213.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834.

Tsiatis, A. A., Degruttola, V., and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association*, 90(429):27–37.

Yang, L., Yu, M., and Gao, S. (2016). Prediction of coronary artery disease risk based on multiple longitudinal biomarkers. *Statistics in Medicine*, 35(8):1299–1314.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447.

Zheng, Y., Cai, T., and Pepe, M. S. (2013). Adopting nested case–control quota sampling designs for the evaluation of risk markers. *Lifetime Data Analysis*, 19(4):568–588.

4 Journal Article 2

Title of Journal Article

Bayesian Quantile Regression Joint Models of Longitudinal and Recurrent Event Data

Journal proposed for article submission: Statistical Methods in Medical Research

Bayesian Quantile Regression Joint Models of Longitudinal and Recurrent Event Data

Abstract

Recurrent event outcomes (e.g., multiple heart failures, cancer recurrences, hospital readmissions, etc.) are commonly encountered in longitudinal biomedical studies along with some longitudinal continuous outcome(s). In this paper, we propose a new joint models (JM) framework that models a longitudinal continuous outcome using linear quantile mixed model (LQMM) jointly with a Cox proportional hazards model (PHM) for the recurrent event outcome. The recurrent event outcome is allowed to be right-censored. In contrast to conventional mean regression based JM, our quantile regression (QR) based JM provide a more flexible, distribution-free way to study covariate effects at different conditional quantiles of the longitudinal outcome. Meanwhile, it's association with the risk of event recurrence can also be examined. The model becomes extremely useful in application when higher (e.g. blood pressure) or lower (e.g. CD4 count) tail of the outcome is more relevant to clinical interest. We develop a Gibbs sampling algorithm for model inference, which is based on the location-scale representation of the asymmetric Laplace distribution for LQMM. The

Bayesian inferential algorithm can be easily implemented in existing software. We assess the performance of our Bayesian algorithm through extensive simulation study and apply the proposed model to the joint analysis of longitudinal systolic blood pressure and recurrent coronary heart diseases data from the Atherosclerosis Risk In Communities Study.

Key words: Bayesian; Joint models; Linear quantile mixed model; Recurrent events.

4.1 Introduction

In traditional joint models (JM) of longitudinal and time-to-event data, a linear mixed model (LMM) (Laird and Ware, 1982) is commonly used for the longitudinal continuous outcome; while possible violation to the normality assumption in the error term is not considered. Moreover, an LMM only models covariate effects on the conditional mean of the outcome; however, in many clinical settings it is more desirable to make inference at lower or higher quantiles of the outcome. For example, researchers used quantile regression (QR) to study the risk factors of lower birth weight, in which they found several effects on the lower quantiles were significantly different from the mean effects (Koenker and Hallock, 2001). In contrast to linear regression, QR is a more flexible tool that relaxes the distributional assumption, and provides a way to study covariate effects on various conditional quantiles of the outcome (Koenker, 2005). So far, there is little work has been done to connect QR method with the JM framework. To our knowledge, Farcomeni and Viviani (2015) is the first one to incorporate a linear quantile mixed model (LQMM) into a JM of longitudinal and terminal event data, in which they developed a Monte Carlo Expectation Maximization (MCEM) algorithm for parameter estimation.

Joint analysis of longitudinal and time-to-event outcomes have been studied by many authors. However, majority of the literature focuses on JM of longitudinal and a single time-to-event (e.g. death) data. For example, Self and Pawitan (1992); Tsiatis et al. (1995);

Wulfsohn and Tsiatis (1997) developed the JM methods for survival analysis with a time dependent covariate measured with error. Tsiatis and Davidian (2004) gives an excellent review of such JM method. In contrast, joint analysis of longitudinal and recurrent event outcomes has received less attention so far. To our knowledge, Henderson et al. (2000) developed a shared random effects JM for longitudinal data and recurrent events. Kim et al. (2012) considered a JM of longitudinal and recurrent event data with informative terminal event. Efendi et al. (2013) proposed a JM of longitudinal and recurrent event outcomes that accommodates overdispersion. Furthermore, there is little work has consider using LQMM in the JM of longitudinal and recurrent event data.

In this paper, we propose a new version of JM for longitudinal and recurrent event data, in which we adopt the LQMM in modeling the longitudinal continuous outcome and use the Cox proportional hazards model (PHM) for recurrent events. The model takes the format of shared random effects JM, similar as in Wulfsohn and Tsiatis (1997) and Rizopoulos (2011), in which the longitudinal outcome is treated as a time-dependent covariate in the recurrent event model and the dependence is measured by some association parameter. We develop a Gibbs sampling algorithm for model inference, which is based on the fact that minimizing the original QR loss function is equivalent to maximizing the likelihood function of the so-called asymmetric Laplace distribution (ALD) (Yu and Moyeed, 2001). Moreover, the ALD can be further reparameterized using a location-scale mixture representation that leads to a combination of normal and exponential distributions (Kotz et al., 2012; Kozumi and Kobayashi, 2011; Luo et al., 2012). The proposed Bayesian estimating algorithm can be directly implemented in existing software. In application, we use the data from the Atherosclerosis Risk in Communities Study (ARIC) (The ARIC investigators, 1989), in which we investigate various covariate effects on different quantiles of the systolic blood pressure (SBP) and its association with the recurrences of coronary heart disease (CHD). QR based JM is able to add much more flexibility in modeling and provide additional insight into the data by investigating covariate effects on various conditional quantiles of SBP and its association with the risk of CHD recurrence, which are not possible using traditional mean regression based methods.

The rest of this paper is organized as follows. In Section 4.2, we give details of the proposed statistical model and the Bayesian algorithm for model inference. In Section 4.3, we present simulation study to validate the proposed methods. In Section 4.4, we apply the proposed methods to the ARIC data. We conclude the paper with a discussion in Section 4.5.

4.2 Methods

4.2.1 Bayesian Linear Quantile Mixed Model

Let $Y_i(t_{ij})$ be the longitudinal outcome for subject i measured at time t_{ij} where $i = 1, \cdots, N$ and $j = 1, \cdots, n_i$. Consider the linear mixed effects model:

$$Y_i(t) = \boldsymbol{X}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i + \varepsilon_i(t), \tag{4.1}$$

where $\boldsymbol{\beta}$ is a $p-$dimensional vector of fixed effects, $\boldsymbol{X}_i(t)$ contains the corresponding fixed covariates, $\boldsymbol{u}_i$ is a $k-$dimensional vector of random effects for subject i, and $\boldsymbol{Z}_i(t)$ are the corresponding random covariates.

An LQMM assumes that the conditional quantile of the outcome is a linear function of the covariates, i.e.,

$$Q_{Y_i(t)|\boldsymbol{X}_i(t),\boldsymbol{Z}_i(t)}(\tau) = \boldsymbol{X}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i, \tag{4.2}$$

where the τth quantile of a random variable Y is defined as $Q_Y(\tau) = F_Y^{-1}(\tau) = \inf\{y : F_Y(y) \geq \tau\}$, for $\tau \in [0, 1]$. Parameter estimations can then be obtained by minimizing the following loss function,

$$\hat{\boldsymbol{\beta}}_\tau = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i,t} \left[\rho_\tau\left(Y_i(t) - \boldsymbol{X}_i^\top(t)\boldsymbol{\beta} - \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i\right)\right],$$

where $\rho_\tau(\cdot)$ is defined as $\rho_\tau(Y) = Y(\tau - I(Y < 0))$.

Above minimization problem can be rephrased as a maximum-likelihood problem by assuming the random error $\varepsilon_i(t)$ in (4.1) follows ALD with location parameter equals 0, scale parameter σ and skewness parameter τ (Koenker and Machado, 1999; Yu and Moyeed, 2001):

$$Y_i(t) = \boldsymbol{X}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i + \varepsilon_i(t), \varepsilon_i(t) \sim ALD(0, \sigma, \tau).$$

This becomes clear when writing out the conditional likelihood function:

$$\ell(Y_i(t)|\boldsymbol{\beta}_\tau, \boldsymbol{u}_i, \sigma) = \frac{\tau(1-\tau)}{\sigma}\exp\left[-\rho_\tau\left(\frac{Y_i(t) - \boldsymbol{X}_i^\top(t)\boldsymbol{\beta}_\tau - \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i}{\sigma}\right)\right].$$

In Bayesian quantile regression context a Gibbs sampler algorithm for model inference can be developed when we utilize a location-scale mixture representation of the ALD (Kotz et al., 2012). Under such parameterization the random error is represented as $\varepsilon_i(t) = \kappa_1 e_i(t) + \kappa_2\sqrt{\sigma e_i(t)}v_i(t)$ with $v_i(t) \sim \mathcal{N}(0,1), e_i(t) \sim \exp(1/\sigma)$ and

$$\kappa_1 = \frac{1-2\tau}{\tau(1-\tau)} \qquad \kappa_2^2 = \frac{2}{\tau(1-\tau)}.$$

This reparameterization leads to the following linear mixed model,

$$Y_i(t) = \boldsymbol{X}_i^\top(t)\boldsymbol{\beta}_\tau + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i + \kappa_1 e_i(t) + \kappa_2\sqrt{\sigma e_i(t)}v_i(t),$$

or equivalently,

$$\ell(Y_i(t)|\boldsymbol{\beta}_\tau, \boldsymbol{u}_i, e_i(t), \sigma) = \frac{1}{\sqrt{2\pi\kappa_2^2\sigma e_i(t)}} \exp\left[-\frac{(Y_i(t) - \boldsymbol{X}_i^\top(t)\boldsymbol{\beta}_\tau - \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i - \kappa_1 e_i(t))^2}{2\kappa_2^2\sigma e_i(t)}\right]. \tag{4.3}$$

As discussed in Yu and Moyeed (2001), irrespective of the actual distribution of the data, Bayesian quantile regression using ALD distribution works quite well for different error distributions and the performance is quite robust and satisfactory.

4.2.2 Joint Models Using Longitudinal Quantile Regression

For subject i, let T_{ik}^* be the underlying true kth recurrent event time and C_i be the censoring time, which is assumed to be independent of both outcomes. Then $T_{ik} = \min(C_i, T_{ik}^*)$, for $k = 1, \cdots, m_i$, is the observed kth event time, where m_i is the total number of recurrent events for subject i. Let Δ_{ik} be the recurrent event indicator at time T_{ik} which is defined as $\Delta_{ik} = I(T_{ik}^* < C_i)$, and $I(\cdot)$ is the indicator function. A kth recurrent event is observed at time T_{ik} if $\Delta_{ik} = 1$, i.e. $T_{ik}^* < C_i$; other wise, $\Delta_{ik} = 0$.

Let $Y_i(t)$ be the continuous longitudinal outcome for subject i measured at time t. Note that we can only observe $Y_i(t)$ when $t \leq C_i$, and the complete longitudinal trajectory up to follow-up time t for subject i can be written as $\mathcal{Y}_i(t) = \{Y_i(s) : 0 \leq s \leq t\}$. We denote the true underlying longitudinal measurement with $m_i(t)$ and his/her complete history of true longitudinal process as $\mathcal{M}_i(t) = \{m_i(s) : 0 \leq s \leq t\}$. We propose a new JM that uses longitudinal quantile mixed model (LQMM) as follows:

$$\begin{cases} Y_i(t) = m_i(t) + \varepsilon_i(t) = \boldsymbol{X}_i^\top(t)\boldsymbol{\beta}_\tau + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i + \varepsilon_i(t), \varepsilon_i(t) \sim ALD(0, \sigma, \tau) \\ r_i(t|\mathcal{M}_i(t), \boldsymbol{W}_i; \boldsymbol{\gamma}_\tau, \alpha_\tau) = r_{i0}(t)\exp(\boldsymbol{W}_i^\top\boldsymbol{\gamma}_\tau + \alpha_\tau(\boldsymbol{X}_i^\top(t)\boldsymbol{\beta}_\tau + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i)) \end{cases} \tag{4.4}$$

where in the LQMM for the longitudinal process, $\boldsymbol{X}_i(t)$ are $p-$dimensional fixed effect covariates and $\boldsymbol{Z}_i(t)$ are $k-$dimensional covariates associated with the $k-$dimensional multivariate normal random effects $\boldsymbol{u}_i$. The submodel for recurrent event process takes the format of Cox proportional hazards model (PHM) where $r_{i0}(\cdot)$ is the baseline intensity function and $\boldsymbol{W}_i$ are $q-$dimensional fixed effect covariates that are only associated with event time (not the longitudinal outcome). In Equation (4.4), individual heterogeneity is captured by $\boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i$, which is the deviation of subject i from the population average. Meanwhile, these two models are linked by treating the longitudinal outcome as a time dependent covariate in the recurrent event process, and the degree of associations is measured by parameter α.

Assume a total number of m_i events are observed for subject i within the censoring time C_i. The likelihood function for recurrent event data can be written as:

$$\begin{aligned}\ell(\boldsymbol{T}_i, \boldsymbol{\Delta}_i; \boldsymbol{\theta}) &= \prod_{k=1}^{m_i} \left[r_i(T_{ik}; \boldsymbol{\theta}|\mathcal{M}_i(T_{ik}), \boldsymbol{W}_i)^{\Delta_{ik}} \exp\left(-\int_{T_{ik-1}}^{T_{ik}} r_i(s; \boldsymbol{\theta}|\mathcal{M}_i(s), \boldsymbol{W}_i) ds \right) \right] \\ &= \prod_{k=1}^{m_i} \left[r_i(T_{ik}; \boldsymbol{\theta}|\mathcal{M}_i(T_{ik}), \boldsymbol{W}_i)^{\Delta_{ik}} \right] \exp\left(-\int_{0}^{T_{im_i}} r_i(s; \boldsymbol{\theta}|\mathcal{M}_i(s), \boldsymbol{W}_i) ds \right), \quad (4.5)\end{aligned}$$

where $r_i(\cdot)$ is given in (4.4) and T_{i0} is set to be 0.

For the baseline intensity $r_{i0}(t)$, a parametric form such as Weibull model can be used or it can be left unspecified. Specifically, we consider constant baseline intensity and piecewise-constant baseline intensity function in simulation study and data application respectively. A wider range of baseline intensity functions can also be used and the integration can be approximated with some numeric methods (e.g. Simpson's rule). Besides the Cox PHM, other functional forms can also be considered for the the recurrent events submodel. For example, accelerated failure time model when the proportionality assumption is violated and counting process approach is another nonparametric option.

In quantile regression, all parameter estimators are functions of the quantile. This is also true in the proposed JM. That is, parameter estimations in the recurrent events submodel, such as α and $\boldsymbol{\gamma}$, also change depending which τ is chosen. Quantile regression provides us the flexibility to conduct a study over the entire conditional distribution of the longitudinal outcome through fitting the model using a set of selected quantiles. Less varying values in the estimation indicates a relatively stable covariate effect on the outcome, and vice versa. If the interest lies only in assessing the effect on the lower or higher quantile of the longitudinal outcome and its association with the event process we may just fix the quantile and conduct the analysis. For simplicity we omit the quantile notation in all parameters in the following sections (e.g. $\boldsymbol{\theta}$ stands for $\boldsymbol{\theta}_\tau$ for all quantile-based parameters).

4.2.3 Complete Likelihood Function and Bayesian Inference

For subject i, the complete joint likelihood function of the longitudinal and recurrent event data is the product of three components: the conditional likelihood functions of the longitudinal and recurrent event outcomes (conditional on the unobserved random effects) and the density of the random effects:

$$L_i(\boldsymbol{\theta}; \boldsymbol{T}_i, \boldsymbol{\Delta}_i, \mathcal{Y}_i(C_i), \boldsymbol{u}_i) = \ell(\mathcal{Y}_i(C_i); \boldsymbol{\theta}|\boldsymbol{u}_i)\ell(\boldsymbol{T}_i, \boldsymbol{\Delta}_i; \boldsymbol{\theta}|\boldsymbol{u}_i) f(\boldsymbol{u}_i|\boldsymbol{\Sigma}), \tag{4.6}$$

where vector $\boldsymbol{\theta}$ represents a set of all the parameters from each distribution function in (4.6), $\ell(\boldsymbol{T}_i, \boldsymbol{\Delta}_i; \boldsymbol{\theta}|\boldsymbol{u}_i)$ is given in (4.5) and $\ell(\mathcal{Y}_i(C_i); \boldsymbol{\theta}|\boldsymbol{u}_i) = \prod_{0 \leq t \leq C_i} \ell(Y_i(t); \boldsymbol{\theta}|\boldsymbol{u}_i)$, where $\ell(Y_i(t), \boldsymbol{\theta}|\boldsymbol{u}_i)$ takes the format of (4.3), and random effects are assumed to be multivariate normal.

For parameter estimation, we take advantage of the location-scale mixture representation of the ALD that is described in Section 4.2.1 and propose a fully Bayesian inference approach for unknown parameters. Specifically, given the complete likelihood function in (4.6) and according to the Bayes theorem, the posterior distributions of the model parameters is given by

$$f(\boldsymbol{\theta}|\boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{\mathcal{Y}}, \boldsymbol{u}) \propto \prod_{i=1}^{N} L_i(\boldsymbol{T}_i, \boldsymbol{\Delta}_i, \mathcal{Y}_i(C_i), \boldsymbol{u}_i; \boldsymbol{\theta}) f(\boldsymbol{\theta}), \tag{4.7}$$

where N is the total number subjects, $\boldsymbol{T} = (\boldsymbol{T}_1, \boldsymbol{T}_2, \cdots, \boldsymbol{T}_N)$, $\boldsymbol{\mathcal{Y}} = (\mathcal{Y}_1(C_1), \mathcal{Y}_2(C_2), \cdots, \mathcal{Y}_N(C_N))$, $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \cdots, \boldsymbol{\Delta}_N)$, $\boldsymbol{u} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_N)$, and $f(\boldsymbol{\theta})$ is the product of the prior distributions:

$$f(\boldsymbol{\theta}) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma})\pi(\alpha)\pi(\sigma)\pi(\boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ is a $k \times k$ covariance matrix of the random effects. We may choose the following prior distributions: $\boldsymbol{\beta} \sim \mathcal{N}_p(\mathbf{0}, 10^3\mathbf{I}), \boldsymbol{\gamma} \sim \mathcal{N}_q(\mathbf{0}, 10^3\mathbf{I}), \alpha \sim \mathcal{N}(0, 10^3), \sigma \sim \mathcal{IG}(10^{-3}, 10^{-3}), \boldsymbol{\Sigma}^{-1} \sim Wishart(\mathbf{I}, k+1)$. We also consider Cholesky decomposition prior for $\boldsymbol{\Sigma}$ in our simulation studies and find similar results as Wishart prior gives (results not shown). In the simulation study, we find that the posterior inference is not sensitive to the prior choice.

By using fully Bayesian approach the uncertainty of the parameter estimates is fully captured in the posterior distributions and no asymptotic theory is needed to derive the standard error. It is straightforward to code the proposed JM and implement it in JAGS software (Plummer, 2003) and the JAGS model file for simulation study is provided separately in Web Supplement.

4.3 Simulation Study

In this section we aim to validate the proposed Bayesian inference algorithm through simulation study. We consider different simulation scenarios where the random error is generated from either standard normal distribution or skewed distributions with varied skewness parameter. Simulated data are then fitted using our proposed QRJM as well as the LMJM that assumes normality in the error term. We assess the model performance using bias and precision of the point estimates.

We simulate the data from Model (4.8), all the regression coefficients β_1, β_2, β_3, γ, and α are set to be 1. In the longitudinal process, we simulate X_{1i} and the random effect u_i from $\mathcal{N}(0,1)$ and X_{2i} from $Bernuolli(0.5)$. A maximum of six observations are generated for each subject at follow-up times $t = 0, 0.25, 0.5, 0.75, 1.0,$ and 1.25. To simulate event times, we set the baseline intensity $r_{0i}(t)$ to be constant 1 and generate W_i from $\mathcal{N}(0,1)$. The random censoring time C_i is generated from $2 + Beta(1,1)$ and the recurrent times T^*_{ik} are generated using calendar time. Finally, we set the observed recurrent times as $T_{ik} = min(C_i, T^*_{ik})$ and recurrent event indicators as $\Delta_{ik} = I(T_{ik} < C_i)$ for $k = 1, \cdots, m_i$. And we limit a maximum of five recurrent events for each subject.

$$\begin{cases} Y_i(t) = m_i(t) + \varepsilon_i(t) = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 t + u_i + \varepsilon_i(t) \\ r_i(t|W_i; \gamma, \alpha) = r_{0i}(t) \exp(\gamma W_i + \alpha m_i(t)) \end{cases} \tag{4.8}$$

We consider the following four scenarios in our simulation study. In each scenario, we simulate 200 data sets with sample size equals 250 or 500 in each.

- Scenario 1: $\varepsilon_i(t)$ follows ALD(0, 1, $\tau = 0.25$) (right-skewed);
- Scenario 2: $\varepsilon_i(t)$ follows standard normal distribution;
- Scenario 3: $\varepsilon_i(t)$ follows ALD(0, 1, $\tau = 0.50$) (symmetric at 0, heavy tails);
- Scenario 4: $\varepsilon_i(t)$ follows ALD(0, 1, $\tau = 0.75$) (left-skewed).

The simulation results are reported in Tables 4.1 and 4.2 as well as in Tables 4.4 and 4.5 of the Appendices. We report estimation bias, stand error (SE), mean squared error (MSE), and coverage probability (CP) from different model fittings. Table 4.1 summaries the results from simulation Scenario 1. Models under comparison including QRJM with $\tau = 0.25$ (true model), QRJM with $\tau = 0.50$ (i.e. median regression), and the conventional LMJM. It is seen that when random error is right-skewed under this scenario, our proposed Bayesian algorithm is able to recover the truth given the correct quantile value is set and it performs the best among the three models under comparison. Both the median regression and LMJM result in comparable estimates of regression coefficients in Scenario 1; however, for scale parameter σ and the constant baseline hazard r_0, both models give very biased point estimates with large MSE and low CP. The performance of LMJM becomes even worse in Scenario 4, where the data are generated from ALD(0, 1, $\tau = 0.75$). Parameter estimates (except for γ) are highly biased with large MSE and very low CP (Appendices Table 4.5). In contrast, the performance of QRJM with $\tau = 0.5$ stays relatively stable in Scenario 4 as it does in Scenario 1. Overall, except for the true model, median regression outperforms LMJM when the data are skewed. In Scenario 2, we are interested in study the performance of the proposed Bayesian inference algorithm when the underlying error distribution is not ALD. Under this scenario, the error term follows standard normal distribution, and median regression (QRJM with $\tau = 0.5$) performs comparably to the true model with just slightly larger bias and lower CP. Thus, our Bayesian inferential algorithm using the ALD distribution for QRJM is robust to model misspecification. Simulation results for Scenarios 3 can be found in Appendices Tables 4.4. Results from Scenario 3 shows that when data are symmetric with heavy tail, LMJM performs slightly worse but comparably to the true model. In addition, as expected, we observe smaller bias and MSE when the sample size increases from 250 to 500 for most parameter estimations in all simulation scenarios.

In summary, the proposed Gibbs sampling algorithm performs well in recovering the true model parameters and is robust to model misspecification. Median regression model is robust to data skewness in terms of estimating regression coefficients. And the traditional LMJM is sensitive to deviation from normality assumption in the error term.

Table 4.1: Simulation result for Scenario 1 in which random error is generated from ALD$(0, 1, \tau = 0.25)$.

		QRJM ($\tau = 0.25$)				QRJM ($\tau = 0.50$)				LMJM			
		Bias	SE	MSE	CP	Bias	SE	MSE	CP	Bias	SE	MSE	CP
$n = 250$	Coefficients for longitudinal process												
	β_1	0.014	0.091	0.008	0.980	0.025	0.102	0.012	0.960	0.036	0.112	0.013	0.955
	β_2	−0.002	0.164	0.029	0.920	0.007	0.174	0.031	0.930	0.022	0.182	0.034	0.955
	β_3	0.033	0.068	0.005	0.940	0.046	0.083	0.009	0.890	0.058	0.095	0.012	0.890
	σ	−0.000	0.031	0.001	0.950	−0.321	0.021	0.103	0.000	–	–	–	–
	Coefficients for recurrent event process												
	γ	0.001	0.073	0.005	0.955	0.002	0.078	0.005	0.970	0.004	0.081	0.007	0.935
	r_0	0.032	0.134	0.018	0.945	−0.786	0.055	0.622	0.000	−0.915	0.032	0.838	0.000
	α	−0.007	0.071	0.005	0.950	−0.028	0.080	0.008	0.905	−0.030	0.090	0.009	0.920
$n = 500$	Coefficients for longitudinal process												
	β_1	−0.001	0.064	0.004	0.920	0.009	0.071	0.006	0.920	0.010	0.078	0.007	0.930
	β_2	−0.003	0.116	0.011	0.970	0.011	0.121	0.012	0.980	0.006	0.126	0.013	0.955
	β_3	0.020	0.048	0.003	0.950	0.026	0.058	0.004	0.950	0.029	0.067	0.005	0.935
	σ	0.001	0.022	0.001	0.970	−0.320	0.015	0.103	0.000	–	–	–	–
	Coefficients for recurrent event process												
	γ	0.007	0.052	0.004	0.920	0.007	0.056	0.004	0.920	0.007	0.058	0.004	0.915
	r_0	−0.017	0.093	0.008	0.940	−0.810	0.036	0.657	0.000	−0.929	0.020	0.863	0.000
	α	0.003	0.051	0.003	0.950	−0.001	0.059	0.004	0.940	0.004	0.068	0.004	0.940

Table 4.2: Simulation result for Scenario 2 in which random error is generated from $\mathcal{N}(0, 1)$.

		LMJM				QRJM ($\tau = 0.5$)			
		Bias	SE	MSE	CP	Bias	SE	MSE	CP
	Coefficients for longitudinal process								
	β_1	−0.015	0.076	0.005	0.950	−0.010	0.076	0.006	0.960
	β_2	−0.002	0.148	0.026	0.920	0.000	0.149	0.027	0.910
	β_3	0.004	0.038	0.001	0.970	0.003	0.038	0.002	0.920
$n = 250$	σ	0.009	0.047	0.002	0.960	–	–	–	–
	Coefficients for recurrent event process								
	γ	0.002	0.054	0.003	0.960	-0.009	0.053	0.003	0.930
	r_0	0.014	0.090	0.009	0.940	0.046	0.091	0.011	0.875
	α	0.010	0.048	0.002	0.930	-0.022	0.047	0.003	0.875
	Coefficients for longitudinal process								
	β_1	−0.006	0.053	0.003	0.920	0.000	0.054	0.003	0.930
	β_2	0.001	0.106	0.012	0.930	0.006	0.106	0.012	0.940
	β_3	0.010	0.026	0.001	0.920	0.009	0.027	0.001	0.920
$n = 500$	σ	0.003		0.000	0.960	–	–	–	–
	Coefficients for recurrent event process								
	γ	0.003	0.038	0.002	0.940	−0.007	0.037	0.002	0.930
	r_0	−0.009	0.063	0.005	0.910	0.022	0.063	0.006	0.900
	α	0.009	0.034	0.002	0.890	−0.014	0.033	0.002	0.850

4.4 Application

4.4.1 The Atherosclerosis Risk in Communities (ARIC) Study

We apply the proposed QRJM to a data derived from the Atherosclerosis Risk in Communities Study (ARIC). ARIC is a prospective epidemiological study conducted in four diverse U.S. communities (i.e. Forsyth County in North Carolina, Jackson County in Mississippi, Minneapolis suburbs in Minnesota, and Washington County in Maryland). The aims of ARIC are to investigate the causes of atherosclerosis and its clinical outcomes, and variation in cardiovascular risk factors, medical care, and disease by race, gender, location, and date (The ARIC investigators, 1989). In the ARIC cohort study component, a cohort sample of approximately 4000 individuals aged 45 to 64 years were randomly selected from each of the four ARIC field centers as the representatives of the community to receive extensive follow-up study. The cohort recruitment was started in 1987 and the first screening examination was conducted in $1987-1989$ when the baseline data were collected. Additional three cohort follow-up examinations were conducted at approximately three-year intervals, in $1990-1992$, $1993-1995$, and $1996-1998$, and the longitudinal measures were created. In addition, follow-up also occurs semi-annually, by telephone, to maintain contact and to assess health status of the cohort.

One of the objectives of the cohort study is to investigate the trends in rates of hospitalized myocardial infarction (MI) and coronary heart diseases (CHD) in those communities. It is reported in previous studies of ARIC data that risk factors for CHD differ significantly by race group. Wattanakit et al. (2005) and Rodriguez et al. (2014) also showed that systolic blood pressure (SBP) was an important risk factor for CHD in the ARIC cohort. The range of SBP can be roughly divided into three intervals: ≤ 120 mm Hg (normal), 120 to 140 mm Hg (prehypertension), and ≥ 140 mm Hg (high blood pressure). However, few studies have considered the time change rate of SBP among hypertensive patients, especially the time effect on different quantiles of SBP, and its association with the risk of recurrent CHD. To fill the gap, we aim to use the proposed QRJM to investigate the baseline covariate and time effects on different quantiles of SBP and to characterize the association between SBP trajectory and the risk of CHD recurrences.

Data used in this study is derived from one of the four study communities (center ID was de-identified to protect privacy), in which we include only white hypertensive participants (with SBP $>$ 140 mm Hg and DBP $>$ 90 mm Hg at baseline or self-reported history of physician-diagnosed hypertension or taking anti-hypertensive medicine). Participants who had prevalent CHD (defined by Q waves on the electrocardiogram, self-reported history of MI diagnosis, coronary artery bypass graft, or coronary angioplasty) before the first examination are excluded from the analysis. The resulting study cohort consists of 657 participants. Repeated measures of SBP were collected from the four longitudinal examination cycles that started at 1987 and ended at 1998. Out of the total 657 participants, 440 (67%) individuals had four complete SBP measures, 113 (17%), 54 (8%), and 50 (8%) had three, two and only one measure respectively. The LOESS curve in Figure 4.1 shows no obvious time trend for SBP in approximately 11-year follow-up period. Follow-up for recurrent CHD continued through 2010. The median follow-up time is 21 years with the maximum as 24 years. 242 (36%) deaths occurred during the follow-up and 115, 31, and 17 patients experienced one, two or more than two CHD events. Baseline characteristics of the study cohort is presented in Appendices Table 4.6.

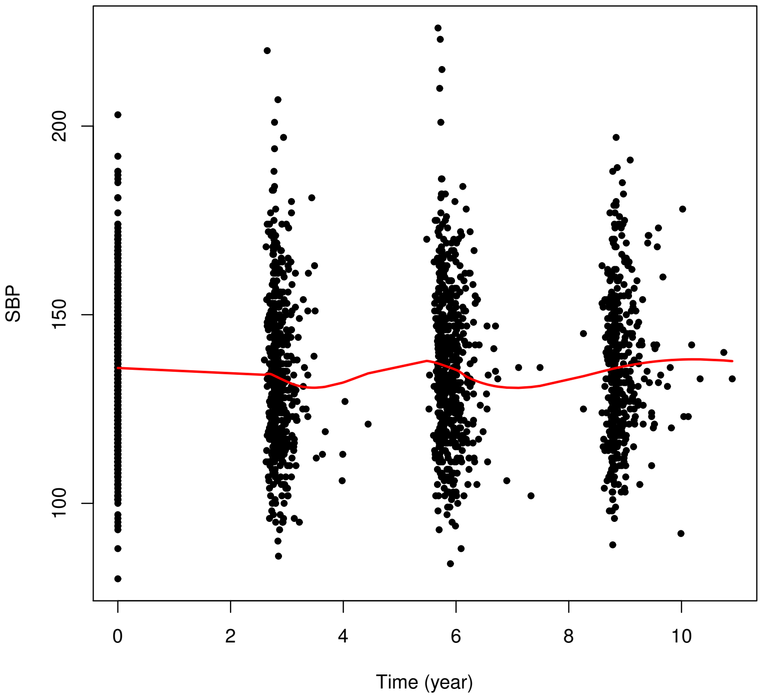


Figure 4.1: Scatter plot with LOESS curve of longitudinal SBP measures in the study cohort.

In data analysis, follow-up time is converted from days to years and the first examination date is set to time 0; baseline age is centered by subtracting the overall mean, and total-cholesterol is standardized to have mean 0 and standard deviation 1. We consider the following QRJM:

$$\begin{cases} sbp_i(t) = m_i(t) + \varepsilon_i(t) = \beta_0 + \beta_1 age_{0i} + \beta_2 chol_i + \beta_3 I_{hyper-med_i} + \beta_4 t + u_{i1} + u_{i2} t + \varepsilon_i(t) \\ r_i(t|\mathcal{M}_i(t); \boldsymbol{\gamma}, \alpha) = r_0(t) v_i \exp(\gamma_1 I_{male_i} + \gamma_3 I_{smoke_i} + \gamma_4 I_{diabetes_i} + \alpha m_i(t)) \end{cases}$$

where we assume $\varepsilon_i(t) \sim ALD(0, \sigma, \tau)$. age_0 is the baseline age at the first examination, $chol$ stands for the total-cholesterol level (mg/dL), I_{med} is the variable indicating whether an individual had taken hypertension lowering medication, t is the follow-up time in year, and u_{i1} and u_{i2} are subject-specific random intercept and slope to account for the within subject correlation and between subject variation. In the recurrent events submodel, we specify a piecewise constant baseline intensity function with three time intervals, where λ_k is the hazard rate for time interval $[t_{k-1}, t_k)$, that is $I_k(t) = 1$ if $t \in [t_{k-1}, t_k)$ and 0 otherwise. Knots t_1 and t_2, used to define piecewise constant time intervals, are selected as the 33.3% and 66.7% percentiles of the ordered follow-up time; while $t_0 = 0$ and t_3 is the maximum of follow-up time. We also include a frailty term v_i in the recurrent events model that introduces correlation of multiple CHD events within the same individual (Hougaard, 2012). Other model covariates include indicator variables for male (I_{male}), ever smoke (I_{smoke}), and diabetes mellitus ($I_{diabetes}$). In the QRJM, the true underlying longitudinal measure of SBP is treated as a time dependent covariate in the recurrent CHD process and α is the association parameter governing the dependence between these two processes. Two chains with diverse initial values are initiated in the Bayesian inference algorithm and the chains are considered to converge if the potential scale reduction factors (PSRF)(Brooks and Gelman, 1998) for all parameters are below 1.1.

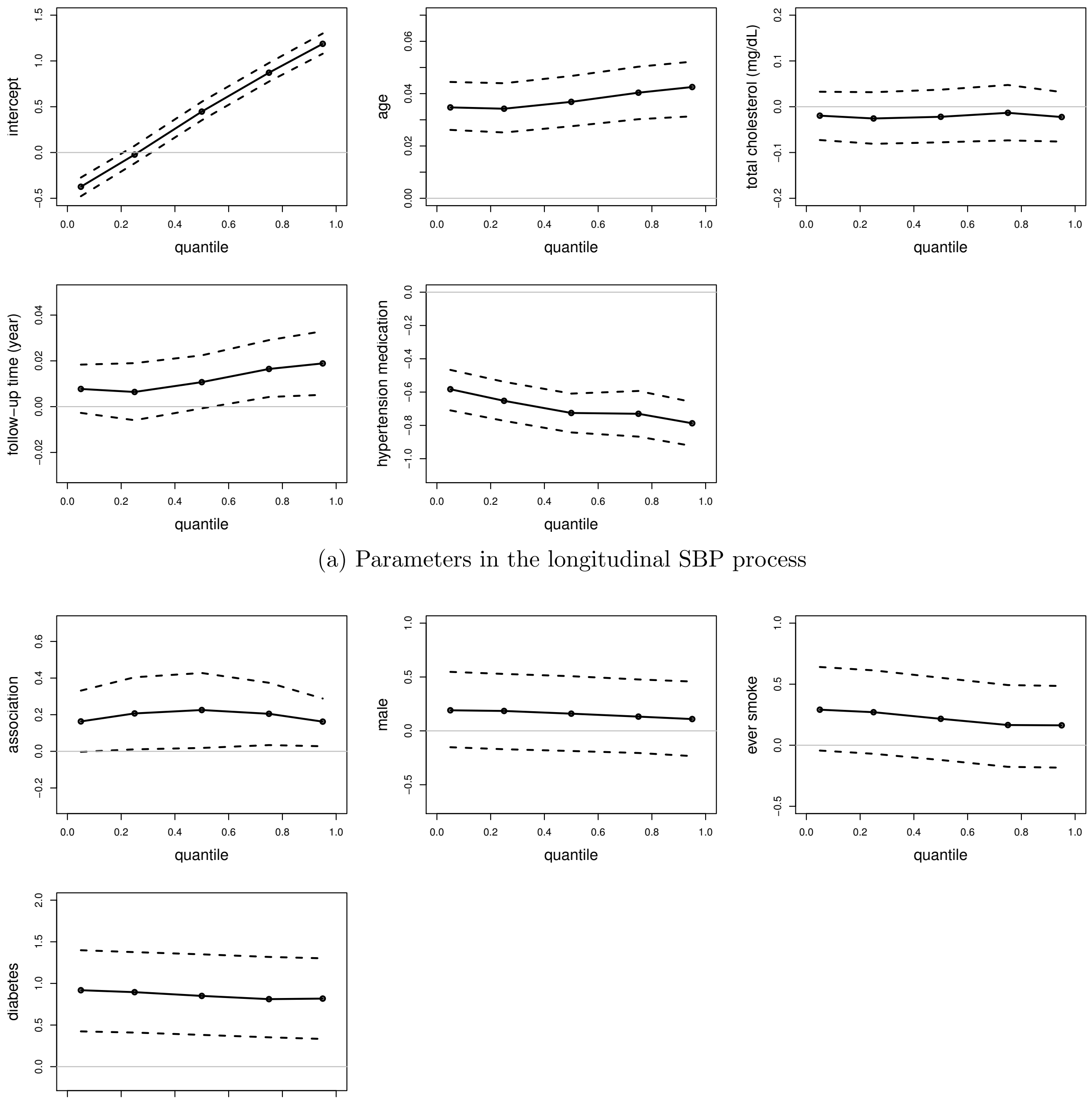
4.4.2 Inference Results for ARIC data

Inference results from five different conditional quantiles (0.05, 0.25, 0.50, 0.75, and 0.95) of SBP are shown in Table 4.3 as well as visualized in Figure 4.2. In the longitudinal SBP process, older participants generally have higher SBP level and the effect of baseline age is consistently positive across all five selected quantiles of SBP. For example, one year increase in baseline age is associated with 0.037 (95% CI: (0.028, 0.047)) unit increase in the median (i.e., $\tau = 0.50$) of standardized SBP in the study cohort when controlling for other covariates. Total cholesterol level is negatively associated with SBP; however, the effects are not significant across all five quantiles. In general, people who took hypertension medications have significantly lower SBP and the effect of taking hypertension medications is larger at higher levels of SBP. Moreover, it is interesting to see that follow-up time has a significantly positive effect on higher quantiles of SBP (i.e., $\tau = 0.75$ and 0.95) while for lower quantiles ($\tau = 0.05$, 0.25, and 0.50) the effect is not significant. This finding can be an important indication that among the hypertension patients who originally have excessively higher SBP deteriorate even faster than those with lower SBP.

In the recurrent CHD process, we see all positive association between the five conditional quantiles of SBP and the risk of CHD recurrences, which coincides with our expectation as well as previous findings from ARIC data. However, the degree of association between these two processes varies among the conditional quantiles and is found to be strongest at the conditional median of SBP (relative risk:1.25, 95% CI: (1.02, 1.53)) among the five selected quantiles. For other regression covariates, diabetic patients are at significantly higher risk of having recurrent CHD compared with non-diabetic. For example, when controlling for other factors, the risk of having additional CHD event is 2.3 times higher (exp(0.85), 95% CI: (1.46, 3.85)) for people with diabetes than those who are diabetes free at $\tau = 0.5$. We also observe the posterior effect of diabetes decreases as quantile increases. This indicates that the effect of diabetes on the risk of recurrent CHD is less important for patients with higher SPB. Although male patients and ever smokers are also at higher risk of experiencing recurrent CHD, the relative risks are statistically insignificant compared with female and never smokers respectively.

Table 4.3: ARIC data analysis: Parameter estimation and 95% credible interval (in parenthesis) from QRJM at five quantiles.

	$\tau = 0.05$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.95$
Coefficients for longitudinal SBP process					
Intercept	−0.374	−0.023	0.447	0.872	1.187
	(−0.478, −0.274)	(−0.118, 0.074)	(0.352, 0.554)	(0.775, 0.978)	(1.079, 1.300)
Age_0	0.035	0.034	0.037	0.040	0.043
	(0.026, 0.044)	(0.025, 0.044)	(0.028, 0.047)	(0.030, 0.050)	(0.031, 0.052)
Total cholesterol (mg/dL)	−0.020	−0.026	−0.022	−0.013	−0.022
	(−0.073, 0.033)	(−0.081, 0.032)	(−0.078, 0.037)	(−0.073, 0.047)	(−0.076, 0.032)
Hypertension medicine	−0.583	−0.652	−0.725	−0.730	−0.787
	(−0.710, −0.467)	(−0.773, −0.538)	(−0.842, −0.609)	(−0.868, −0.593)	(−0.924, −0.660)
Follow- up time (yr)	0.008	0.006	0.011	0.016	0.019
	(−0.003, 0.018)	(−0.006, 0.019)	(−0.001, 0.022)	(0.004, 0.029)	(0.005, 0.033)
Coefficients for recurrent CHD process					
Male	0.191	0.185	0.160	0.132	0.110
	(−0.152, 0.548)	(−0.170, 0.528)	(−0.187, 0.507)	(−0.205, 0.477)	(−0.234, 0.458)
Ever smoke	0.291	0.271	0.216	0.165	0.163
	(−0.044, 0.641)	(−0.070, 0.613)	(−0.121, 0.552)	(−0.177, 0.493)	(−0.184, 0.485)
Diabetes	0.918	0.895	0.850	0.811	0.818
	(0.424, 1.399)	(0.409, 1.376)	(0.381, 1.349)	(0.352, 1.318)	(0.333, 1.301)
Association	0.163	0.207	0.226	0.205	0.162
	(−0.003, 0.332)	(0.011, 0.405)	(0.019, 0.428)	(0.034, 0.374)	(0.028, 0.288)



(a) Parameters in the longitudinal SBP process

(b) Parameters in the recurrent CHD events process

Figure 4.2: ARIC data analysis: Posterior mean (solid line) and point-wise 95% credible interval (dashed lines) of parameter estimation against different quantiles.

4.5 Discussion

In the application of conventional JM methodologies for longitudinal and recurrent event data, we usually encountered two limitations: first, the normality assumption of the random error in the LMM was not realistic, and no obvious transformation of the longitudinal outcome to produce residual normality was applicable. This limitation is confirmed by our simulation study where LMJM tends to provide biased point estimates as well as lower coverage probabilities for model parameters when the longitudinal data are non-normal. Second, LMJM models only the conditional mean of the outcome; however, in our (and other) clinical research applications, it is more desirable to consider the tails of the biomarker distribution.

Our work on QRJM that uses an LQMM for the longitudinal process provides a more flexible way for simultaneously modeling conditional quantile a of longitudinal outcome and the risk of event recurrences. In the application of ARIC data, we illustrate this flexibility by jointly modeling repeated SBP measurements and the risk of developing recurrent CHD. Our results reveal some findings that may not be observed using linear regression based method. For example, patients who originally have higher SBP deteriorate faster than those with lower SBP; while there is no significant increasing time trend for lower quantiles of SBP. Diabetes is a strong risk factor of CHD recurrences and after controlling for different quantiles of SBP, we find the effect of diabetes diminishes among higher SBP groups. Thus, QRJM is able to provide us with more informative insight into the disease progression and the association between the two disease processes in terms of various quantile-based estimations.

Our novel extension of traditional JM finds practical importance in many clinical fields. Besides our example in cardiovascular study, potential applications would include, but not limited to, cancer studies (e.g., prostate-specific antigen level and the risk of prostate cancer recurrences), hospital care studies (e.g., extracellular fluid volume and hospital readmissions among gastrointestinal cancer patients), etc. In those studies, higher or lower level of the longitudinal biomarker is often associated with worse medical condition in patients and leads to higher risk of disease. If treatment effect on tails of the outcome is significantly different from the average effect, treating different groups of patients using the exact same way won't be as effective as we tailor the treatment strategy according to patient-specific situation.

In this work, we developed a Gibbs sampling algorithm to fit the proposed model, where the likelihood function of the longitudinal quantile regression is written under the location-scale representation of the ALD distribution. The proposed algorithm is straightforward to implement in existing Bayesian software. The current version of Gibbs sampler, which is implemented in JAGS software, uses a piecewise constant baseline intensity function in the recurrent events submodel. However, other choices can also be considered and the integration of the baseline intensity function can be approximated using Simpson's rule. Moreover, other functional forms for both outcomes can also be considered to extend the proposed method. For example, in the longitudinal process, nonlinear QR (Koenker and Park, 1996) models can be used in stead of linear QR. In the recurrent events submodel, accelerated failure time model can be considered when the proportionality assumption is violated and counting process approach is another nonparametric option.

Appendices

A Additional Simulation Results

Table 4.4: Simulation result for Scenario 3 in which random errors are generated from ALD$(0, 1, \tau = 0.50)$.

		QRJM ($\tau = 0.5$)				LMJM			
		Bias	SE	MSE	CP	Bias	SE	MSE	CP
	Coefficients for longitudinal process								
	β_1	0.007	0.087	0.007	0.970	0.010	0.094	0.009	0.970
	β_2	-0.008	0.162	0.028	0.925	-0.005	0.167	0.029	0.935
	β_3	0.026	0.062	0.004	0.945	0.032	0.073	0.006	0.935
$n = 250$	σ	-0.002	0.031	0.001	0.935	–	–	–	–
	Coefficients for recurrent event process								
	γ	0.006	0.070	0.005	0.960	0.005	0.075	0.006	0.930
	r_0	0.019	0.124	0.016	0.950	0.026	0.142	0.018	0.965
	α	0.016	0.065	0.004	0.940	0.015	0.073	0.006	0.950
	Coefficients for longitudinal process								
	β_1	-0.001	0.061	0.004	0.920	-0.000	0.066	0.005	0.925
	β_2	-0.002	0.113	0.011	0.965	-0.001	0.117	0.011	0.965
	β_3	0.015	0.044	0.002	0.950	0.018	0.051	0.003	0.935
$n = 500$	σ	0.002	0.022	0.001	0.940	–	–	–	–
	Coefficients for recurrent event process								
	γ	0.008	0.050	0.003	0.915	0.008	0.054	0.004	0.920
	r_0	-0.016	0.087	0.008	0.940	-0.018	0.099	0.009	0.945
	α	0.003	0.047	0.003	0.940	0.003	0.054	0.003	0.975

Table 4.5: Simulation result for Scenario 4 in which random errors are generated from ALD$(0, 1, \tau = 0.75)$.

		QRJM ($\tau = 0.75$)				QRJM ($\tau = 0.5$)				LMJM			
		Bias	SE	MSE	CP	Bias	SE	MSE	CP	Bias	SE	MSE	CP
$n = 250$	Coefficients for longitudinal process												
	β_1	0.006	0.091	0.007	0.970	0.033	0.104	0.013	0.950	0.142	0.111	0.030	0.800
	β_2	-0.008	0.165	0.029	0.930	0.015	0.176	0.031	0.950	0.130	0.193	0.050	0.940
	β_3	0.028	0.068	0.005	0.940	0.048	0.083	0.010	0.890	0.146	0.092	0.029	0.665
	σ	-0.002	0.031	0.001	0.940	-0.323	0.021	0.105	0.000	–	–	–	–
	Coefficients for recurrent event process												
	γ	0.010	0.074	0.005	0.940	-0.047	0.082	0.008	0.930	-0.000	0.079	0.007	0.925
	r_0	0.018	0.137	0.017	0.965	4.468	0.636	20.354	0.000	8.599	0.431	73.958	0.000
	α	0.016	0.070	0.004	0.955	-0.020	0.083	0.008	0.930	-0.133	0.067	0.022	0.515
$n = 500$	Coefficients for longitudinal process												
	β_1	-0.005	0.064	0.005	0.940	0.012	0.072	0.006	0.900	0.111	0.076	0.019	0.685
	β_2	-0.001	0.116	0.015	0.920	0.007	0.122	0.014	0.960	0.101	0.133	0.026	0.895
	β_3	0.027	0.048	0.003	0.900	0.029	0.058	0.004	0.935	0.111	0.064	0.015	0.590
	σ	0.003	0.022	0.001	0.940	-0.317	0.015	0.101	0.000	–	–	–	–
	Coefficients for recurrent event process												
	γ	0.005	0.052	0.002	0.955	0.006	0.056	0.004	0.940	-0.004	0.056	0.004	0.910
	r_0	-0.003	0.096	0.010	0.935	4.377	0.436	19.469	0.000	8.773	0.267	76.972	0.000
	α	0.001	0.051	0.003	0.940	-0.001	0.060	0.004	0.920	-0.104	0.049	0.013	0.400

B Summary Table of Study Cohort Characteristics

Table 4.6: Baseline characteristics of study cohort with stratification by SBP level

		SBP groups (mm Hg)			
Characteristics[†]	Total ($n = 657$)	< 120 ($n = 133$, 20.2%)	[120, 140) ($n = 217$, 33.0%)	≥ 140 ($n = 307$, 46.7%)	p-value[*]
Age	56.4 (5.8)	55.0 (5.7)	55.7 (6.0)	57.4 (5.4)	<0.001
SBP	135.9 (18.5)	110.5 (6.9)	129.2 (5.7)	151.6 (11.2)	<0.001
Cholesterol (mg/dL)	215.9 (41.7)	215.1 (42.1)	214.0 (42.0)	217.6 (41.7)	0.60
Gender (male)	341 (51.9)	64 (48.1)	117 (53.9)	160 (52.1)	0.57
Ever smoke (yes)	379 (57.5)	81 (60.9)	126 (58.1)	172 (56.0)	0.63
Hypertension medication (yes)	445 (67.7)	132 (99.2)	196 (90.3)	117 (38.1)	<0.001
Diabetes (yes)	90 (13.7)	15 (11.3)	27 (12.4)	48 (15.6)	0.38

[†] mean (sd) for continuous variables and frequency (percentage) for categorical variables.

[*] Comparing three SBP groups; ANOVA test for continuous variables and χ^2 test for categorical variables.

C JAGS Model File

JAGS model file to fit QRJM of longitudinal and recurrent event data in simulation study.

```
model{
      k1 <- (1 - 2 * qt) / (qt * (1 - qt))
      k2 <- 2 / (qt * (1 - qt))

      # prior of random effects
      for (i in 1:I){ # I: unique subject id
        # prior for random effects
          u[i] ~ dnorm(0, tau)
      } # end of loop i

      # longitudinal process, BQR mixed model using ALD representation
      for (j in 1:N_l){ # N_l: number of longitudinal observations
          er[j] ~ dexp(sigma)
          mu[j] <- beta1 * X1_l[j] + beta2[X2_l[j]] + beta3 * t[j] + u[id_l[j]]+ k1 * er[j]
          prec[j] <- sigma / (k2 * er[j])
          y[j] ~ dnorm(mu[j], prec[j])
      } #end of j loop

      # recurrent events part, baseline hazard is set to constant c
      for(k in 1:I){
        for (l in (s[k]+1):s[k+1]){
          m1[l] <- beta1 * X1[k] + beta2[X2[k]] + beta3 * Ri1[l] + u[id_r[l]]
          m2[l] <- beta1 * X1[k] + beta2[X2[k]] + beta3 * Ri2[l] + u[id_r[l]]
          res[l] <- (exp(gamma * W[k] + alpha * m2[l]) - exp(gamma * W[k] + alpha * m1[l])) / (alpha * beta3)
          S[l] <- exp(- c * res[l])
          risk[l] <- c * exp(gamma * W[k] + alpha * m2[l])
          L[l] <- pow(risk[l], event[l]) * S[l] / 1E+08
          zeros[l] ~ dpois(-log(L[l]))
        } # end of l loop
      }#end of k loop

      # priors for other parameters
      alpha ~ dnorm(0, 0.001)
      beta1 ~ dnorm(0, 0.001)
      beta2[1] <- 0
      beta2[2] ~ dnorm(0, 0.001)
      beta3 ~ dnorm(0, 0.001)
      gamma ~ dnorm(0, 0.001)
      sigma ~ dgamma(0.001, 0.001)
      c ~ dunif(0.01, 10)
      tau <- pow(var, -2)
      var ~ dunif(0, 1000)
  }
```

References

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational And Graphical Statistics*, 7(4):434–455.

Efendi, A., Molenberghs, G., Njagi, E. N., and Dendale, P. (2013). A joint model for longitudinal continuous and time-to-event outcomes with direct marginal interpretation. *Biometrical Journal*, 55(4):572–588.

Farcomeni, A. and Viviani, S. (2015). Longitudinal quantile regression in the presence of informative dropout through longitudinal–survival joint modeling. *Statistics in Medicine*, 34(7):1199–1213.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480.

Hougaard, P. (2012). *Analysis of Multivariate Survival Data.* Springer Science & Business Media.

Kim, S., Zeng, D., Chambless, L., and Li, Y. (2012). Joint models of longitudinal data and recurrent events with informative terminal event. *Statistics in Biosciences*, 4(2):262–281.

Koenker, R. (2005). *Quantile Regression.* Cambridge University Press.

Koenker, R. and Hallock, K. (2001). Quantile regression: An introduction. *Journal of Economic Perspectives*, 15(4):43–56.

Koenker, R. and Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448):1296–1310.

Koenker, R. and Park, B. J. (1996). An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics*, 71(1):265–283.

Kotz, S., Kozubowski, T., and Podgorski, K. (2012). *The Laplace Distribution And Generalizations: A Revisit with Applications to Communications, Economics, Engineering, And Finance*. Springer Science & Business Media.

Kozumi, H. and Kobayashi, G. (2011). Gibbs sampling methods for bayesian quantile regression. *Journal of Statistical Computation and Simulation*, 81(11):1565–1578.

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Luo, Y., Lian, H., and Tian, M. (2012). Bayesian quantile regression for longitudinal data models. *Journal of Statistical Computation and Simulation*, 82(11):1635–1649.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, volume 124, page 125.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829.

Rodriguez, C. J., Swett, K., Agarwal, S. K., Folsom, A. R., Fox, E. R., Loehr, L. R., Ni, H., Rosamond, W. D., and Chang, P. P. (2014). Systolic blood pressure levels among adults with hypertension and incident cardiovascular events: the atherosclerosis risk in communities study. *JAMA Internal Medicine*, 174(8):1252–1261.

Self, S. and Pawitan, Y. (1992). Modeling a marker of disease progression and onset of disease. In *AIDS Epidemiology*, pages 231–255. Springer.

The ARIC investigators (1989). The atherosclerosis risk in community (ARIC) study: Design and objectwes. *American Journal of Epidemiology*, 129(4):687–702.

Tsiatis, A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834.

Tsiatis, A., Degruttola, V., and Wulfsohn, M. (1995). Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association*, 90(429):27–37.

Wattanakit, K., Folsom, A. R., Chambless, L. E., and Nieto, F. J. (2005). Risk factors for cardiovascular event recurrence in the atherosclerosis risk in communities (aric) study. *American Heart Journal*, 149(4):606–612.

Wulfsohn, M. and Tsiatis, A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1):330–339.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447.

5 Journal Article 3

Title of Journal Article

Bayesian Quantile Regression Joint Models: Dynamic Predictions of Recurrent Event Probability

Journal proposed for article submission: Statistics in Medicine

Bayesian Quantile Regression Joint Models: Dynamic Predictions of Recurrent Event Probability

Abstract

Joint models (JM) of longitudinal and time-to-event outcomes have received increasing interest recently. A novel use of JM is to make dynamic predictions of event probability from observed longitudinal and time-to-event data. In contrast to the extensive literature on JM of longitudinal and single time-to-event (e.g. death) data, less attention has been received for the JM of longitudinal and recurrent event data. In this work, we develop a Gibbs sampling algorithm for making subject-specific dynamic predictions of the risk of event recurrence based the JM of longitudinal and recurrent event outcomes. In our JM, differently from traditional JM, we adopt a linear quantile mixed model (LQMM) instead of the frequently used linear mixed model (LMM) for the longitudinal outcome. Compared with LMM, as a quantile regression based model, LQMM is more robust against non-normality or outliers in the data. Moreover, LQMM is more flexible than LMM in that it allows investigation of covariate effects on any conditional quantile of the outcome. In the proposed Bayesian algorithm, predictions are calculated based on the entire longitudinal trajectory as well as the recurrent events history and can be dynamically updated when new data from either (or both) outcome is available. In addition, implemented through the MCMC technique, the uncertainty of the predictive inference is fully captured in the posterior distribution and no asymptotic theory is needed to derive the standard error. It is straightforward to code and implement the proposed Bayesian predictive algorithm in existing software. We assess the performance of our model through extensive simulation studies and apply it to dynamically predict the probability of recurrent coronary heart diseases for the Atherosclerosis Risk In Communities Study cohort.

Key words: Bayesian; Dynamic predictions; Joint models; Linear quantile mixed model; Recurrent events.

5.1 Introduction

Joint analysis of longitudinal and time-to-event outcomes has been studied by many authors. However, majority of the literature focuses on JM of longitudinal and a single time-to-event (e.g. death) outcomes. For example, Self and Pawitan (1992); Tsiatis et al. (1995); Wulfsohn and Tsiatis (1997) developed the JM methods for survival analysis with a time dependent covariate measured with error. Despite the popularity of repeated time-to-event or recurrent event data, such as hospital readmissions, multiple cardiovascular diseases (e.g. stroke, heart failure, etc.), and cancer recurrences, etc., joint analysis of longitudinal and recurrent event outcomes has received less attention so far. To our knowledge, Henderson et al. (2000) developed a shared random effects JM for longitudinal and recurrent event data. Kim et al. (2012) considered a JM of longitudinal and recurrent event data with informative terminal event. And Efendi et al. (2013) proposed a JM of longitudinal data and recurrent events that accommodates overdispersion. Moreover, disease recurrence is always one of the important clinical outcomes in longitudinal biomedical studies, which can be used to monitor disease progression and health condition of the patients. Predictions of future event probability based on historical data attracts increasing interest recently. Accurate predictions of disease probability can play an important role in disease intervention and prevention. The JM framework offers a novel way of making such personalized dynamic predictions of future event-free probability (Rizopoulos, 2011; Taylor et al., 2013). However, so far, there is little work has been done on the dynamic predictions of event recurrences under the JM framework.

In this paper, we propose a Gibbs sampling algorithm for making subject-specific predictions of the risk of event recurrence based on a new JM framework. Differently from conventional JM, we develop a quantile regression joint models (QRJM) framework that uses linear quantile mixed model (LQMM) in modeling the longitudinal continuous outcome; while the Cox proportional hazard model (PHM) is used for recurrent event outcome. There are several advantages of quantile regression (QR) based methods over the mean regression models (e.g. linear mixed model or LMM). First of all, LMM assumes normal error in the data; however, it is common that this normality assumption is violated in reality and no suitable transformation can be found. This is especially true when working with longitudinal data as the skewness changes over time. Moreover, LMM only models covariate effects on the conditional mean of the outcome; however, in many clinical settings it is more desirable to make inference at lower or higher quantiles of the outcome. For example, researchers used quantile regression QR to study risk factors of lower birth weight, in which they found several effects on the lower quantiles were significantly different from the mean effects (Koenker and Hallock, 2001). This feature of QR find its great importance in biomedical studies, where individuals with extremer biomarker measurements are often at higher risk of disease or death. To our knowledge, Farcomeni and Viviani (2015) is the first one to incorporate an LQMM into a JM of longitudinal and terminal event data. However, there is little work has been do to use LQMM in the joint analysis with recurrent event data so far.

The development of the Gibbs sampler is based on the fact that minimizing the original QR loss function is equivalent to maximizing the likelihood function of the asymmetric Laplace distribution (ALD) (Yu and Moyeed, 2001). It is relatively easy to make subject-specific predictions from the posterior samples of the fixed effects and the posterior predictive distributions of the random effects, which are direct results of our sampling algorithm. In addition, implemented through the MCMC technique, the uncertainty of the predictive inference is fully captured in the posterior distribution. We conduct extensive simulation studies to validate the proposed JM in model inference as well in making predictions. After that we apply the proposed algorithm to the Atherosclerosis Risk in Communities Study (ARIC) data (The ARIC investigators, 1989), in which we investigate various covariate effects on different quantiles of the systolic blood pressure (SBP), its association with the recurrence of coronary heart disease (CHD), and make dynamic predictions of future recurrent CHD probability using historical SBP measures and recurrent CHD data. Moreover, such predictions can be dynamically updated whenever new data from either longitudinal or time-to-event outcome are available.

The rest of this paper is organized as follows. In Section 5.2, we give details of the proposed statistical model and the Bayesian algorithms used for model inference and dynamic predictions. In Section 5.3, we present simulation studies to validate the proposed methods. In Section 5.4.1, we apply the proposed methods to the ARIC data to make dynamic predictions of CHD risk. We conclude the paper with a discussion in Section 5.5.

5.2 Methods

5.2.1 Joint Models Using Longitudinal Quantile Regression

For subject i, let T_{ik}^{*} be the underlying true kth recurrent event time and C_i be the censoring time, which is assumed to be independent of both outcomes. Then $T_{ik} = \min(C_i, T_{ik}^{*})$, for $k = 1, \cdots, K$, is the observed kth event time, where K is the total number of recurrent events for subject i. Let Δ_{ik} be the recurrent event indicator at time T_{ik} which is defined as $\Delta_{ik} = I(T_{ik}^{*} < C_i)$, and $I(\cdot)$ is the indicator function. A kth recurrent event is observed at time T_{ik} if $\Delta_{ik} = 1$, i.e. $T_{ik}^{*} < C_i$; other wise, $\Delta_{ik} = 0$.

Let $Y_i(t)$ be the continuous longitudinal outcome for subject i measured at time t. Note that we can only observe $Y_i(t)$ when $t \leq C_i$, and the complete longitudinal trajectory up to follow-up time t for subject i can be written as $\mathcal{Y}_i(t) = \{Y_i(s) : 0 \leq s \leq t\}$. We denote the true underlying longitudinal measurement with $m_i(t)$ and his/her complete history of true longitudinal process as $\mathcal{M}_i(t) = \{m_i(s) : 0 \leq s \leq t\}$. We propose a new JM that uses longitudinal quantile mixed model (LQMM) as follows:

$$\begin{cases} Y_i(t) = m_i(t) + \varepsilon_i(t) = \boldsymbol{X}_i^{\top}(t)\boldsymbol{\beta}_{\tau} + \boldsymbol{Z}_i^{\top}(t)\boldsymbol{u}_i + \varepsilon_i(t), \varepsilon_i(t) \sim ALD(0, \sigma, \tau) \\ r_i(t|\mathcal{M}_i(t), \boldsymbol{W}_i; \boldsymbol{\gamma}_{\tau}, \alpha_{\tau}) = r_{i0}(t)\exp(\boldsymbol{W}_i^{\top}\boldsymbol{\gamma}_{\tau} + \alpha_{\tau}(\boldsymbol{X}_i^{\top}(t)\boldsymbol{\beta}_{\tau} + \boldsymbol{Z}_i^{\top}(t)\boldsymbol{u}_i)) \end{cases} \tag{5.1}$$

where in the LQMM for the longitudinal process, $\boldsymbol{X}_i(t)$ are $p-$dimensional fixed effect covariates and $\boldsymbol{Z}_i(t)$ are $k-$dimensional multivariate normal covariates associated with the $k-$dimensional random effects $\boldsymbol{u}_i$. The submodel for recurrent event process takes the format of Cox proportional hazards model (PHM) where $r_{i0}(\cdot)$ is the baseline intensity function and $\boldsymbol{W}_i$ are $q-$dimensional fixed effect covariates that are only associated with event time (not the longitudinal outcome). In Equation (5.1), individual heterogeneity is captured by $\boldsymbol{Z}_i^{\top}(t)\boldsymbol{u}_i$, which is the deviation of subject i from the population average. Meanwhile, these two models are linked by treating the longitudinal outcome as a time dependent covariate in the recurrent event process, and the degree of associations is measured by parameter α.

In quantile regression, all parameter estimators are functions of the quantile. This is also true in the proposed JM. That is, parameter estimations in the recurrent events submodel, such as α and $\boldsymbol{\gamma}$, also change depending which τ is chosen. Quantile regression provides us the flexibility to conduct a study over the entire conditional distribution of the longitudinal outcome through fitting the model using a set of selected quantiles. Less varying values in the estimation indicates a relatively stable covariate effect on the outcome, and vice versa. If the interest lies only in assessing the effect on the lower or higher quantile of the longitudinal outcome and its association with the event process we may just fix the quantile and conduct the analysis.

5.2.2 Bayesian Linear Quantile Mixed Model

Consider the linear mixed effects model:

$$Y_i(t) = \boldsymbol{X}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i + \varepsilon_i(t), \tag{5.2}$$

where $\boldsymbol{\beta}$ is a $p-$dimensional vector of fixed effects, $\boldsymbol{X}_i(t)$ contains the corresponding fixed covariates, $\boldsymbol{u}_i$ is a $k-$dimensional vector of random effects for subject i, and $\boldsymbol{Z}_i(t)$ are the corresponding random covariates.

An LQMM assumes that the conditional quantile of the outcome is a linear function of the covariates, i.e.,

$$Q_{Y_i(t)|\boldsymbol{X}_i(t),\boldsymbol{Z}_i(t)}(\tau) = \boldsymbol{X}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i, \tag{5.3}$$

where the τth quantile of a random variable Y is defined as $Q_Y(\tau) = F_Y^{-1}(\tau) = \inf\{y : F_Y(y) \geq \tau\}$, for $\tau \in [0,1]$. Parameter estimations can then be obtained by minimizing the following loss function,

$$\hat{\boldsymbol{\beta}}_\tau = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i,t} \left[\rho_\tau \left(Y_i(t) - \boldsymbol{X}_i^\top(t)\boldsymbol{\beta} - \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i\right)\right],$$

where $\rho_\tau(\cdot)$ is defined as $\rho_\tau(Y) = Y(\tau - I(Y < 0))$.

Above minimization problem can also be rephrased as a maximum-likelihood problem by assuming the random error $\varepsilon_i(t)$ in (5.2) follows the asymmetric Laplace distribution (ALD) with location parameter equals 0, scale parameter σ and skewness parameter τ (Koenker and Machado, 1999; Yu and Moyeed, 2001):

$$Y_i(t) = \boldsymbol{X}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i + \varepsilon_i(t), \varepsilon_i(t) \sim ALD(0, \sigma, \tau).$$

This becomes clear when writing out the conditional likelihood function of the outcome variable:

$$\ell(Y_i(t)|\boldsymbol{\beta}_\tau, \boldsymbol{u}_i, \sigma) = \frac{\tau(1-\tau)}{\sigma} \exp\left[-\rho_\tau\left(\frac{Y_i(t) - \boldsymbol{X}_i^\top(t)\boldsymbol{\beta}_\tau - \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i}{\sigma}\right)\right].$$

In Bayesian quantile regression context a Gibbs sampling algorithm for model inference is developed when we utilize a location-scale mixture representation of the ALD (Kotz et al., 2012). Under such parameterization the random error is represented as $\varepsilon_i(t) = \kappa_1 e_i(t) + \kappa_2\sqrt{\sigma e_i(t)} v_i(t)$ with $v_i(t) \sim N(0,1), e_i(t) \sim \exp(1/\sigma)$ and

$$\kappa_1 = \frac{1-2\tau}{\tau(1-\tau)} \qquad \kappa_2^2 = \frac{2}{\tau(1-\tau)}.$$

This re-parameterization leads to the following linear mixed model,

$$Y_i(t) = \boldsymbol{X}_i^\top(t)\boldsymbol{\beta}_\tau + \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i + \kappa_1 e_i(t) + \kappa_2\sqrt{\sigma e_i(t)}v_i(t),$$

or equivalently,

$$\ell(Y_i(t)|\boldsymbol{\beta}_\tau, \boldsymbol{u}_i, e_i(t), \sigma) = \frac{1}{\sqrt{2\pi\kappa_2^2\sigma e_i(t)}} \exp\left[-\frac{(Y_i(t) - \boldsymbol{X}_i^\top(t)\boldsymbol{\beta}_\tau - \boldsymbol{Z}_i^\top(t)\boldsymbol{u}_i - \kappa_1 e_i(t))^2}{2\kappa_2^2\sigma e_i(t)}\right]. \quad (5.4)$$

As discussed in Yu and Moyeed (2001), irrespective of the actual distribution of the data, Bayesian quantile regression using ALD distribution works quite well for different error distributions and the performance is quite robust and satisfactory.

5.2.3 The Recurrent Events Submodel

Assume a total number of K events are observed for subject i within the censoring time C_i, the likelihood function for recurrent event data can be written as:

$$\begin{aligned}\ell(\boldsymbol{T}_i, \boldsymbol{\Delta}_i; \boldsymbol{\theta}) &= \prod_{k=1}^{K}\left[r_i(T_{ik}; \boldsymbol{\theta}|\mathcal{M}_i(T_{ik}), \boldsymbol{W}_i)^{\Delta_{ik}} \exp\left(-\int_{T_{ik-1}}^{T_{ik}} r_i(s; \boldsymbol{\theta}|\mathcal{M}_i(s), \boldsymbol{W}_i)ds\right)\right] \\ &= \prod_{k=1}^{K}\left[r_i(T_{ik}; \boldsymbol{\theta}|\mathcal{M}_i(T_{ik}), \boldsymbol{W}_i)^{\Delta_{ik}}\right] \exp\left(-\int_{0}^{T_{iK}} r_i(s; \boldsymbol{\theta}|\mathcal{M}_i(s), \boldsymbol{W}_i)ds\right), \quad (5.5)\end{aligned}$$

where $r_i(\cdot)$ is given in (5.1).

For the baseline intensity $r_{i0}(t)$, a parametric form such as Weibull model can be used or it can be left unspecified. Specifically, we consider constant baseline intensity and piecewise-constant baseline intensity function in simulation study and data application respectively. As mentioned previously, under QR model all estimators are quantile dependent, however, for simplicity we omit the quantile notation in all parameters in the following sections (e.g. $\boldsymbol{\theta}$ stands for $\boldsymbol{\theta}_\tau$ for all quantile-based parameters).

5.2.4 Complete Likelihood Function and Bayesian Inference

For subject i in the sample, the complete joint likelihood of the longitudinal and recurrent event data is the product of three components: the conditional likelihood functions (conditional on the unobserved random effects) of the longitudinal and recurrent event outcomes and the density of the random effects:

$$L_i(\boldsymbol{\theta}; \boldsymbol{T}_i, \boldsymbol{\Delta}_i, \mathcal{Y}_i(C_i), \boldsymbol{u}_i) = \ell(\mathcal{Y}_i(C_i); \boldsymbol{\theta}|\boldsymbol{u}_i)\ell(\boldsymbol{T}_i, \boldsymbol{\Delta}_i; \boldsymbol{\theta}|\boldsymbol{u}_i) f(\boldsymbol{u}_i|\boldsymbol{\Sigma}), \tag{5.6}$$

where vector $\boldsymbol{\theta}$ represents a set of all the parameters from each distribution function in (5.6), $\ell(\boldsymbol{T}_i, \boldsymbol{\Delta}_i; \boldsymbol{\theta}|\boldsymbol{u}_i)$ is given in (4.5) and $\ell(\mathcal{Y}_i(C_i); \boldsymbol{\theta}|\boldsymbol{u}_i) = \prod_{0 \le t \le C_i} \ell(Y_i(t); \boldsymbol{\theta}|\boldsymbol{u}_i)$, where $\ell(Y_i(t), \boldsymbol{\theta}|\boldsymbol{u}_i)$ takes the format of (5.4).

For parameter estimation, we take advantage of the location-scale mixture representation of the ALD that is described in Section 5.2.2 and propose a fully Bayesian inference approach for unknown parameters. Specifically, given the complete likelihood function in (5.6) and according to the Bayes theorem, the posterior distributions of the model parameters are given by

$$f(\boldsymbol{\theta}|\boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{\mathcal{Y}}, \boldsymbol{u}) \propto \prod_{i=1}^{N} L_i(\boldsymbol{T}_i, \boldsymbol{\Delta}_i, \mathcal{Y}_i(C_i), \boldsymbol{u}_i; \boldsymbol{\theta}) f(\boldsymbol{\theta}), \tag{5.7}$$

where N is the total number subjects, $\boldsymbol{T} = (\boldsymbol{T}_1, \boldsymbol{T}_2, \cdots, \boldsymbol{T}_N)$, $\boldsymbol{\mathcal{Y}} = (\mathcal{Y}_1(C_1), \mathcal{Y}_2(C_2), \cdots, \mathcal{Y}_N(C_N))$, $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \cdots, \boldsymbol{\Delta}_N)$, $\boldsymbol{u} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_N)$, and $f(\boldsymbol{\theta})$ is the product of the prior distributions:

$$f(\boldsymbol{\theta}) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma})\pi(\alpha)\pi(\sigma)\pi(\boldsymbol{\Sigma}),$$

where $\mathbf{\Sigma}$ is a $k \times k$ covariance matrix of the random effects. We may choose the following prior distributions: $\boldsymbol{\beta} \sim \mathcal{N}_p(\mathbf{0}, 10^3\mathbf{I}), \boldsymbol{\gamma} \sim \mathcal{N}_q(\mathbf{0}, 10^3\mathbf{I}), \alpha \sim \mathcal{N}(0, 10^3), \sigma \sim \mathcal{IG}(10^{-3}, 10^{-3}), \mathbf{\Sigma}^{-1} \sim Wishart(\mathbf{I}, k+1)$. We also consider Cholesky decomposition prior for $\mathbf{\Sigma}$ in our simulation studies and find similar results as Wishart prior gives (results not shown). In the simulation study, we find that the posterior inference is not sensitive to the prior choice.

5.2.5 Dynamic Predictions of Event-Free Probability

It is always of clinical interest to predict the likelihood that a patient will (or will not) have additional event within a certain time window in the future. The JM of longitudinal and recurrent events framework provides a convenient way to achieve the prediction conditional on the complete bivariate outcomes of the patient. Following previous notations, assume a patient i is followed up to time t, let $\mathcal{Y}_i(t)$ be the observed complete longitudinal measurements, $\mathcal{M}_i(t)$ be the true underlying longitudinal measurements up to time t, and $\mathcal{T}_{it-} = \{T_{ik} : 1 \leq k \leq K_i, T_{iK_i} < t\}$ be the recurrent times before time t. The predicted event-free probability (which is one minus the risk of event) at time m ($m > t$) given previous event times and longitudinal measurements up to censoring time t is:

$$p_i(m|t) = Pr(T_{iK_i+1} \geq m | T_{iK_i+1} > t, \mathcal{T}_{it-}, \mathcal{Y}_i(t); \boldsymbol{\theta}).$$

With further derivation:

$$\begin{aligned} p_i(m|t) &= \int Pr(T_{iK_i+1} \geq m | T_{iK_i+1} > t, \mathcal{T}_{it-}, \mathcal{Y}_i(t), \boldsymbol{u}_i; \boldsymbol{\theta}) \cdot Pr(\boldsymbol{u}_i | T_{iK_i+1} > t, \mathcal{T}_{it-}, \mathcal{Y}_i(t); \boldsymbol{\theta}) d\boldsymbol{u}_i \\ &= \int Pr(T_{iK_i+1} \geq m | T_{iK_i+1} > t, \mathcal{T}_{it-}, \boldsymbol{u}_i; \boldsymbol{\theta}) \cdot Pr(\boldsymbol{u}_i | T_{iK_i+1} > t, \mathcal{T}_{it-}, \mathcal{Y}_i(t); \boldsymbol{\theta}) d\boldsymbol{u}_i \\ &= \int \frac{Pr(T_{iK_i+1} \geq m | \mathcal{M}_i(m, \boldsymbol{u}_i; \boldsymbol{\theta}), \mathcal{T}_{it-}; \boldsymbol{\theta})}{Pr(T_{iK_i+1} > t | \mathcal{M}_i(t, \boldsymbol{u}_i; \boldsymbol{\theta}), \mathcal{T}_{it-}; \boldsymbol{\theta})} \cdot Pr(\boldsymbol{u}_i | T_{iK_i+1} > t, \mathcal{T}_{it-}, \mathcal{Y}_i(t); \boldsymbol{\theta}) d\boldsymbol{u}_i. \quad (5.8) \end{aligned}$$

To estimate $p_i(m|t)$, we can take the advantage of the proposed Gibbs sampling algorithm discussed in Section 5.2.4 and use MCMC technique to calculate the posterior mean of the prediction. Specifically, we are going to estimate

$$\begin{aligned} E_{\boldsymbol{\theta}|\mathcal{D}_N}[p_i(m|t)] &= Pr(T_{iK_i+1} \geq m | T_{iK_i+1} > t, \mathcal{T}_{it-}, \mathcal{Y}_i(t)) \\ &= \int Pr(T_{iK_i+1} \geq m | T_{iK_i+1} > t, \mathcal{T}_{it-}, \mathcal{Y}_i(t); \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}_N) d\boldsymbol{\theta}, \qquad (5.9) \end{aligned}$$

where the first part of the equation is given in (5.8).

A Monte Carlo (MC) estimate of $p_i(m|t)$ can be obtained using the following procedure:

1. Draw $\boldsymbol{\theta}^{(p)}$ from the posterior distributions $Pr(\boldsymbol{\theta}|\mathcal{D}_N)$ for $p = 1, \cdots, P$;
2. For each of the P draws of $\boldsymbol{\theta}^{(p)}$, make Q draws of $\boldsymbol{u}_i^{(q)}, q = 1, \cdots, Q$, from the posterior distribution of random effects $Pr(\boldsymbol{u}_i|\mathcal{D}_N, \boldsymbol{\theta}^{(p)})$ and approximate $p_i(m|t)^{(p)}$ by

$$\frac{1}{Q} \sum_{q=1}^{Q} \frac{Pr(T_{iK_i+1} \geq m | \mathcal{M}_i(m, \boldsymbol{u}_i^{(q)}; \boldsymbol{\theta}^{(p)}), \mathcal{T}_{it-}; \boldsymbol{\theta}^{(p)})}{Pr(T_{iK_i+1} > t | \mathcal{M}_i(t, \boldsymbol{u}_i^{(q)}; \boldsymbol{\theta}^{(p)}), \mathcal{T}_{it-}; \boldsymbol{\theta}^{(p)})};$$

3. After collecting all P $p_i(m|t)^{(p)}$, approximate $p_i(m|t)$ by $\frac{1}{P}\sum_{p=1}^{P} p_i(m|t)^{(p)}$.

In above estimation procedure, P is the total number of MC iterations, $f(\boldsymbol{\theta}|\mathcal{D}_N)$ are the posterior distributions of $\boldsymbol{\theta}$ given in (5.7), and $Pr(\boldsymbol{u}_i|\mathcal{D}_N, \boldsymbol{\theta}^{(p)})$ (i.e., $f(\boldsymbol{u}_i|T_{iK_i+1} > t, \mathcal{T}_{it-}, \mathcal{Y}_i(t), \boldsymbol{\theta}^{(p)})$) is the posterior distribution of the random effects for subject i. The standard error can be computed using the sample variance.

In Step 2 of above algorithm, the posterior predictive values of the random effects are directly results from the MCMC iterations if the subject is within the training data. For out-of-sample subjects who don not belong to the original study population, we can use the inference results from the training data to run additional MCMC iterations to obtain such predictions and the rest of the algorithm follows. Since for each individual there are only a few random effects (two in our current model) to estimate, a short MCMC with 200 iterations should be sufficient for converge (Taylor et al., 2013).

It is relatively easy to make subject-specific predictions of event-free probability from the posterior samples of the fixed effects and the posterior predictive distributions of the random effects, which are direct results of our sampling algorithm. In addition, the uncertainty of the predictive inference is fully captured in the posterior distribution and no asymptotic theory is needed to derive the standard error. It is straightforward to code the proposed JM and implement the algorithm in JAGS software (Plummer, 2003) and the JAGS model file is provided separately in Appendices.

5.2.6 Predictive Accuracy

Predictive accuracy of a model can be evaluated from different perspectives, such as discrimination, calibration, and reclassification, etc. Discrimination measures a model's ability in identifying events versus non-events. Calibration quantifies the closeness of the predictions and the observed values. While reclassification assesses the improvement of a model in prediction after adding new predictor(s). Here we mainly focus on the discriminative ability of our model. Area under the receiver operating characteristic curve (AUC) is a commonly used statistics to evaluate the discriminative ability in prediction, while above average risk difference (AARD) measures the difference in the risk rates comparing events versus non-events at the level of population average risk, and mean risk difference (MRD) is the average difference between TPR and FPR across the risk scale (Pepe et al., 2008). In this work, we use all these three measurements as summary statistics to evaluate the predictive performance of our model.

Following Zheng et al. (2013) and Yang et al. (2016), at a given time t, a future time $t + \Delta t$

and a threshold c, the true positive rate (TPR) and false positive rate (FPR) of the predictive results can be defined as follows:

$$\mathrm{TPR}_t^{\Delta t}(c) = Pr(\mathbf{1} - \boldsymbol{p}(t + \Delta t|t) \geq c|\boldsymbol{T} \leq t + \Delta t),$$

$$\mathrm{FPR}_t^{\Delta t}(c) = Pr(\mathbf{1} - \boldsymbol{p}(t + \Delta t|t) \geq c|\boldsymbol{T} > t + \Delta t),$$

where $\boldsymbol{p}(t + \Delta t|t)$ is a vector of predicted event-free probabilities at time $t + \Delta t$ based on the longitudinal measurements up to time t:

$$p_i(t + \Delta t|t) = S_i(t + \Delta t|\mathcal{Y}_i(t), \boldsymbol{u}_i; \boldsymbol{\theta}), i = 1, \cdots, N.$$

The estimate of $\boldsymbol{p}(t + \Delta t|t)$ is denoted by $\hat{\boldsymbol{p}}(t + \Delta t|t)$ and the estimators of TPR and FPR can be written as:

$$\widehat{TPR}_t^{\Delta t}(c) = \frac{\sum_{i=1}^N (1 - \hat{p}_i(t + \Delta t|t))I(1 - \hat{p}_i(t + \Delta t|t) \geq c)}{\sum_{i=1}^N (1 - \hat{p}_i(t + \Delta t|t))},$$

$$\widehat{FPR}_t^{\Delta t}(c) = \frac{\sum_{i=1}^N \hat{p}_i(t + \Delta t|t)I(1 - \hat{p}_i(t + \Delta t|t) \geq c)}{\sum_{i=1}^N \hat{p}_i(t + \Delta t|t)}.$$

And by definition:

$$\widehat{AUC}_t^{\Delta t} = \int \widehat{TPR}_t^{\Delta t} \left\{(\widehat{FPR}_t^{\Delta t})^{-1}(u)\right\} du,$$

$$\widehat{AARD}_t^{\Delta t} = \widehat{TPR}_t^{\Delta t}(\hat{\rho}) - \widehat{FPR}_t^{\Delta t}(\hat{\rho}),$$

$$\widehat{MRD}_t^{\Delta t} = \int_c \widehat{TPR}_t^{\Delta t}(c)dc - \int_c \widehat{FPR}_t^{\Delta t}(c)dc.$$

And in AARD, $\hat{\rho} = \frac{\sum_{i=1}^N (1 - \hat{p}_i(t + \Delta t|t))}{N}$ is the average risk in the study population at time $t + \Delta t$.

5.3 Simulation Study

We conduct simulation studies to validate the proposed QRJM in making subject-specific dynamic predictions of future event-free probability. In the simulation studies, we assess the predicted event-free probability given in (5.9), by comparing it with the "gold standard" calculated from the true (simulated) values of random effects, fixed effects, and other model parameters.

We simulate the data from Model (5.10), all the regression coefficients β_1, β_2, β_3, γ, and α are set to be 1. In the longitudinal process, we simulate X_{1i} and the random effect u_i from $N(0,1)$, and X_{2i} from $Bernuolli(0.5)$. A maximum of six observations are generated for each subject at follow-up times $t = 0, 0.25, 0.5, 0.75, 1.0,$ and 1.25. To simulate recurrent times, we set the baseline intensity $r_{0i}(t)$ to be constant 1 and generate W_i from $N(0,1)$. The random censoring time C_i is generated from $2 + Beta(1,1)$ and the recurrent times T^*_{ik} are generated using calendar time. Finally, we set the observed recurrent times as $T_{ik} = min(C_i, T^*_{ik})$ and recurrent event indicators as $\Delta_{ik} = I(T_{ik} < C_i)$ for $k = 1, \cdots, m_i$. And we limit a maximum of five recurrent events for each subject.

$$\begin{cases} Y_i(t) = m_i(t) + \varepsilon_i(t) = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 t + u_i + \varepsilon_i(t) \\ r_i(t|W_i;\gamma,\alpha) = r_{0i}(t)\exp(\gamma W_i + \alpha m_i(t)) \end{cases} \tag{5.10}$$

We consider the following four scenarios in simulation study. For each scenario, we simulate 200 data sets with sample size equals to 500 in each. In the simulated data, around 90% of the subjects have at least two events.

- Scenario 1: $\varepsilon_i(t)$ follows ALD with $\tau = 0.25$ (right-skewed);
- Scenario 2: $\varepsilon_i(t)$ follows ALD with $\tau = 0.50$ (symmetric at 0, heavy tail);

- Scenario 3: $\varepsilon_i(t)$ follows ALD with $\tau = 0.75$ (left-skewed);
- Scenario 4: $\varepsilon_i(t)$ follows standard normal distribution.

Before making predictions, 400 (80%) out of the total 500 samples are randomly selected to draw model inference and the rest 100 subjects are used to make out-of-sample dynamic predictions of event-free probability. Inference results for all four scenarios can be found in Table 5.4 in Appendices. From the results we can see in Scenarios 1 to 3, our Bayesian algorithm performs pretty well in recovering the true parameter values with small bias and MSE and high coverage probability. In Scenario 4, where the simulated data are normal, median regression (QRJM with $\tau = 0.5$) also performs well and the result is comparable to the true model (not shown), thus QRJM is robust to model misspecification.

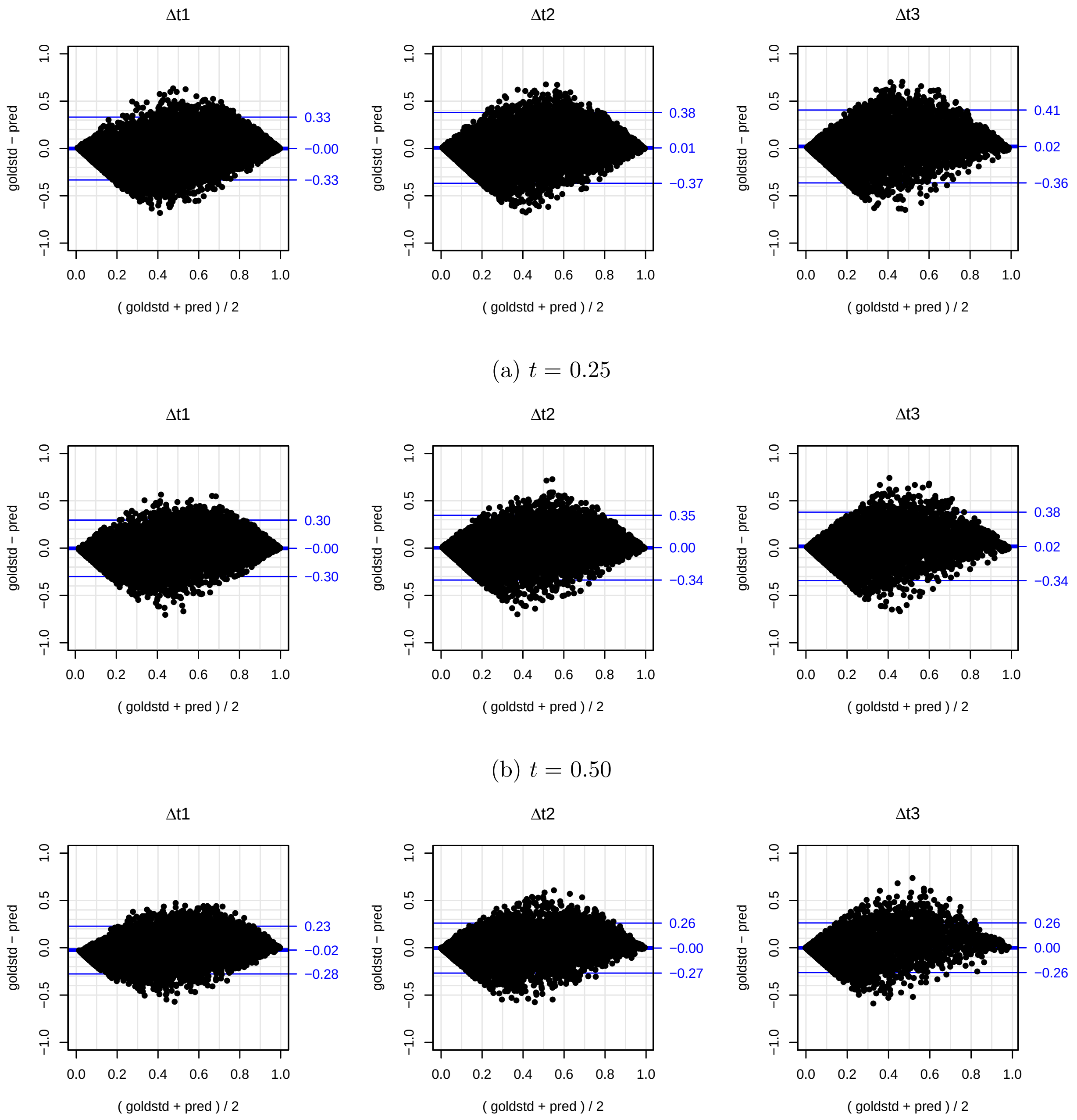
In the prediction part, we use the simulated data and true parameter values to calculate the event-free probability and use it as the "gold standard". To assess the prediction results from different models, we compare the predicted values with the gold standard in terms of bias and MSE. Bland-Altman plot (Bland and Altman, 1986), a commonly used method to compare the agreement of two measurement methods, is used to visualize the prediction results. To make the predictions "dynamic", we choose different combinations of follow-up time (i.e., t) and the prediction time interval (i.e., Δt) to mimic expected real-world scenario.

Table 5.1 summarizes the prediction results from three follow-up time points (t =0.25, 0.50, and 1.00) combined with three prediction time window (Δt = 0.25, 0.50, and 1.00) for Scenario 1. For fixed follow-up time t, with increasing prediction time window Δt, the accuracy and precision of the predicted values, compared with the gold standard, decrease gradually, i.e. large Δt results in larger bias and MSE. This is expected due to increased uncertainty in predictions for time points further into the future. When Δt is fixed and follow-up time t increases, we see improvement in the predictions, which is indicated by decreased MSE with longer follow-up time. This makes sense to us since with longer follow-up time we tend to have more longitudinal measurements as well as event data per subject. As a result, predictions become more precise with additional information from both outcomes. The same message can be found in the Bland-Altman plots. In Figure 5.1, as we fix the follow-up time t and look horizontally, the variation of the plot increases. This is indicated by larger agreement intervals between the predictions and gold standard. While we fix the prediction time window Δt, the agreement becomes stronger with large follow-up time. Moreover, all Bland-Altman plots are horizontally spindle-shaped, suggesting that it is easier to predict a event-free probability near 0 or 1 than that near 0.5.

Comparing different models, Table 5.1 also suggests that when data are generated from skewed distribution, the predictions from models other than the true model have larger MSE and bias thus are less reliable. When data are generated as normal, QRJM with $\tau = 0.50$ performs comparably with the true model. In summary, the best predictions of event-free probability are obtained using the QRJM model and as expected, applying the exact quantile that generated the outcome data performs best and QRJM is robust to model misspecification. Additional simulation results can be found in Appendices Section A.

Table 5.1: Simulation study: MSE and bias of the difference between predicted event-free probability and the gold standard (Scenario 1).

		QRJM ($\tau = 0.25$)		QRJM ($\tau = 0.5$)		LMJM	
t	Δt	MSE	Bias	MSE	Bias	MSE	Bias
0.25	0.25	0.028	0.001	0.035	0.067	0.033	0.023
	0.50	0.035	−0.006	0.045	0.079	0.043	0.024
	1.00	0.037	−0.021	0.048	0.074	0.046	0.015
0.5	0.25	0.022	0.002	0.029	0.067	0.026	0.011
	0.50	0.029	−0.005	0.039	0.078	0.036	0.007
	1.00	0.033	−0.018	0.043	0.077	0.040	−0.005
1.00	0.25	0.018	0.011	0.025	0.078	0.020	0.019
	0.50	0.023	−0.005	0.033	0.079	0.022	−0.004
	1.00	0.026	−0.016	0.036	0.078	0.022	−0.009



(a) $t = 0.25$

(b) $t = 0.50$

(c) $t = 1.00$

Figure 5.1: Bland-Altman plot (bias and 95% limits of agreement) of model predictions versus gold standard based on increasing follow-up time (t= 0.25, 0.50, and 1.00) and three different prediction time intervals ($\Delta t1 < \Delta t2 < \Delta t3$) under Scenario 1.

Table 5.2: Simulation study: MSE and bias of the difference between predicted event-free probability and the gold standard (Scenario 4).

		QRJM ($\tau = 0.5$)		LMJM	
t	Δt	MSE	Bias	MSE	Bias
0.25	0.25	0.015	−0.003	0.014	−0.001
	0.50	0.019	−0.007	0.018	−0.003
	1.00	0.020	−0.014	0.019	−0.010
0.5	0.25	0.012	0.001	0.011	0.002
	0.50	0.015	−0.004	0.014	−0.002
	1.00	0.016	−0.009	0.014	−0.007
1.00	0.25	0.009	0.006	0.008	0.005
	0.50	0.010	−0.004	0.009	−0.003
	1.00	0.010	−0.010	0.010	−0.009

5.4 Application

5.4.1 The Atherosclerosis Risk in Communities (ARIC) Study

In this section, we apply the proposed joint models to the data from the Atherosclerosis Risk in Communities Study (ARIC). ARIC is a prospective epidemiological study conducted in four diverse U.S. communities (i.e. Forsyth County in North Carolina, Jackson County in Mississippi, Minneapolis suburbs in Minnesota, and Washington County in Maryland). The aims of ARIC study are to investigate the causes of atherosclerosis and its clinical outcomes, and variation in cardiovascular risk factors, medical care, and disease by race, gender, location, and date (The ARIC investigators, 1989). There are two components in ARIC study, the cohort component and the community surveillance component. In the cohort component, a cohort sample of approximately 4000 individuals aged $45 - 64$ years were randomly selected from each of the four ARIC field centers as the representative of the community to receive extensive follow-up study. The data used in this work is from the cohort component.

One of the objectives of the cohort study is to investigate the trends in rates of hospitalized myocardial infarction (MI) and coronary heart diseases (CHD) in those communities. From previous studies it is known that risk factors for CHD differ significantly by race group. Wattanakit et al. (2005); Rodriguez et al. (2014) also showed that systolic blood pressure (SBP) was an important risk factor for CHD events in the ARIC data. However, few studies have considered the time change rate of SBP among hypertensive patients, especially the time effect on different quantiles of SBP, and its association with the risk of recurrent CHD events. Thus, in this study, we aim to use longitudinal quantile regression to study the baseline covariate effects and time change rate on different quantiles of SBP and to characterize the association between SBP trajectory and recurrence of CHD events.

Data used in this work is derived from one of the four study communities (center ID is de-identified in the data), in which we include only white hypertensive participants (with SBP $>$ 140 mm Hg and DBP $>$ 90 mm Hg at baseline or self-reported history of physician-diagnosed hypertension or taking anti-hypertensive medicine). Participants who had prevalent CHD (defined by Q waves on the electrocardiogram, self-reported history of MI diagnosis, coronary artery bypass graft, or coronary angioplasty) before the first examination are excluded from the analysis. The resulting study cohort consists of 657 participants. Repeated measures of SBP were collected from the four longitudinal examination cycles that started at 1987 and ended at 1998 (i.e. $1987 - 1989$, $1990 - 1992$, $1993 - 1995$, and $1996 - 1998$). Out of the total 657 participants, 440 (67%) individuals had four complete SBP measures, 113 (17%), 54 (8%), and 50 (8%) had three, two and only one measure respectively. The LOESS curve in

Figure 5.2 shows no obvious time trend for SBP in approximately 11-year follow-up period. Follow-up for recurrent CHD events continued through 2010. The median follow-up time is 21 years with the maximum as 24 years. 242 (36%) deaths occurred during the follow-up and 115, 31, and 17 patients experienced one, two or more than two CHD events. Summary of the baseline characteristics of the study cohort can be found in Appendices Table 5.5.

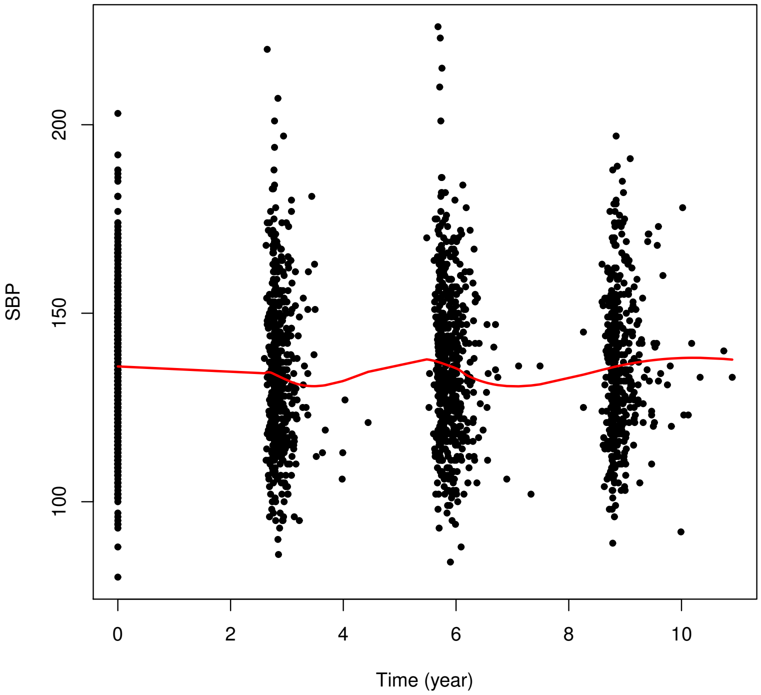


Figure 5.2: Scatter plot with LOESS curve of longitudinal SBP measures in the study cohort.

In data analysis, follow-up time is converted from days to years and the first examination date is set to time 0; baseline age is centered by subtracting the overall mean, and total-cholesterol is standardized to have mean 0 and standard deviation 1. We consider the following JM:

$$\begin{cases} sbp_i(t) = m_i(t) + \varepsilon_i(t) = \beta_0 + \beta_1 age_{0i} + \beta_2 chol_i + \beta_3 I_{hyper-med_i} + \beta_4 t + u_{i1} + u_{i2} t + \varepsilon_i(t) \\ r_i(t|\mathcal{M}_i(t); \boldsymbol{\gamma}, \alpha) = r_0(t) v_i \exp(\gamma_1 I_{male_i} + \gamma_3 I_{smoke_i} + \gamma_4 I_{diabetes_i} + \alpha m_i(t)) \end{cases}$$

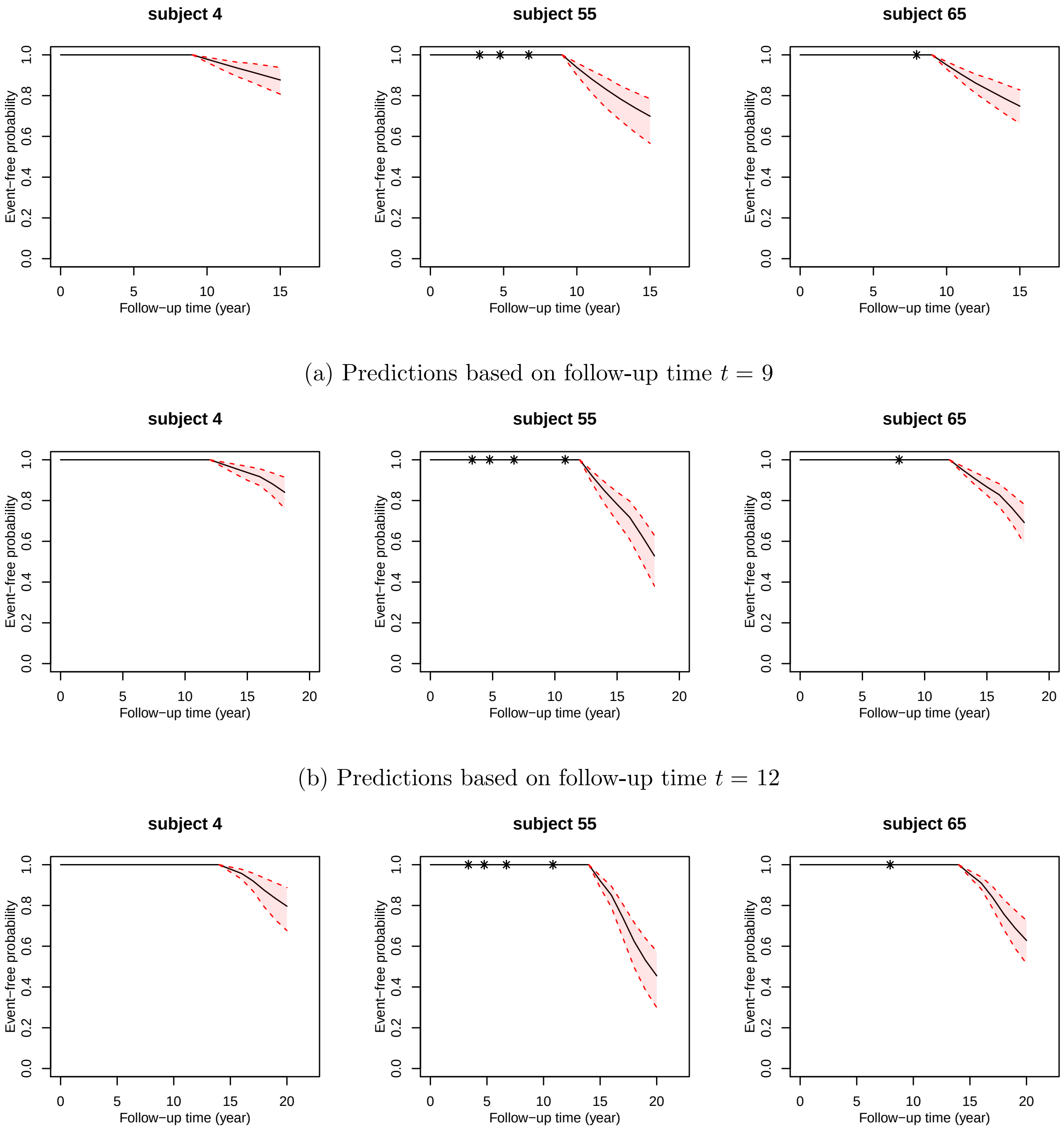
where we assume $\varepsilon_i(t) \sim ALD(0, \sigma, \tau)$. age_0 is the baseline age at the first examination, *chol* stands for the total-cholesterol level (mg/dL), $I_{hyper-med}$ is the variable indicating whether an individual had taken hypertension lowering medication, t is the follow-up time, and u_{i1} and u_{i2} are subject-specific random intercept and slope to account for the within subject correlation and between subject variation. In the recurrent events submodel, we specify a piecewise constant baseline intensity function with three time intervals, where λ_k is the hazard rate for time interval $[t_{k-1}, t_k)$, that is $I_k(t) = 1$ if $t \in [t_{k-1}, t_k)$ and 0 otherwise. Knots t_1 and t_2, used to define piecewise constant time intervals, are selected as the 33.3% and 66.7% percentiles of the ordered follow-up time; while $t_0 = 0$ and t_3 is the maximum of follow-up time. We also include a frailty term v_i in the recurrent events model that introduces correlation of multiple CHD events within the same individual. Other model covariates include indicator variables for male (I_{male}), ever smoke (I_{smoke}), and diabetes mellitus ($I_{diabetes}$). In above QRJM, the true underlying longitudinal measure of SBP is treated as a time dependent covariate in the recurrent event process and α is the association parameter governing the dependence between these two processes. Two chains with diverse initial values are initiated in the Bayesian inference algorithm and the chains are considered to converge if the potential scale reduction factors (PSRF) for all parameters are below 1.1.

We select 80% of the study cohort (i.e. 526 subjects) to draw parameter inference, based on which we then make predictions of CHD event probability for the rest 131 individuals. Inference results and interpretation can be found in Appendices Table 5.6 and will not be discussed further here. In stead, we focus on presenting the predictions result in the following subsection.

5.4.2 Dynamic Predictions of Recurrent CHD Event Risk

To demonstrate how the model works in making subject-specific dynamic predictions of recurrent CHD event risk (in terms of event-free probability), we select three subjects (4, 55, and 65) with distinct data features as the representatives. Figure 5.3 has three panels with increasing follow-up time in years. For each follow-up time t, we make predictions of CHD event-free probability for 1, 2, and 3 years in the future. There are some interesting findings that we would like to highlight: (i) patients who had more CHD events are predicted to have higher risk of CHD recurrence in the future. For example, subject 55 had three CHD events (the most among the three subjects) before year 9 and his/her predicted event-free probability is the lowest (or the highest risk) among the three, followed by subject 65, who had only one event. And subject 4, who didn't have any event occur at year 9, is predicted to only have small probabilities of CHD event in the near future. (ii) With longer follow-up time but no additional event, the predicted event-free probability improves. For example, in subject 55, there is no additional event occurs from $t = 12$ to $t = 14$ and the predicted event-free probabilities based on $t = 14$ are higher than those from $t = 12$ for the same $t + \Delta t$ values. (iii) Wider credible interval for larger prediction time window. This makes sense as the prediction uncertainty is higher for further time in the future.

Furthermore, we make the predictions for the whole testing data with three conditional quantiles (0.25, 0.50, and 0.75) of SBP. We summarize the predictive accuracy and present it in Table 5.3 using AUC, AARD and MRD. First, for the same follow-up time t, predictive accuracy is best with smallest Δt and decreases when Δt increases. This is probably because in this data set majority of the patients didn't have any event observed during the study follow-up and the predicted event-free probability didn't decrease significantly with larger Δt, thus it is more difficult for the model to differentiate those who will versus who will not have CHD event in the longer future since the predictions are close to each other. Second, with longer follow-up time, the predictive accuracy improves accordingly. This finding is consistent with what we observed in the simulation study and it probably dues to the effect of additional longitudinal and recurrent events information used in making predictions. In additional, we also calculate the AUC for the predictions made from the traditional LMJM model. Compared with LMJM, QRJM has better predictive accuracy at some quantiles, but not all. This is because some conditional quantiles of the SBP can be more informative in predicting future recurrent events than the conditional mean while some are not.



(a) Predictions based on follow-up time $t = 9$

(b) Predictions based on follow-up time $t = 12$

(c) Predictions based on follow-up time $t = 14$

Figure 5.3: ARIC data analysis: Dynamic predictions of CHD event-free probability, based on various follow-up time and prediction time window, with 95% credible interval from QRJM at $\tau = 0.5$ for selected subjects ($*$ indicates CHD event).

Table 5.3: ARIC data analysis: AUC, AARD and MRD of the predictions of CHD event-free probability from QRJM and AUC from LMJM.

t	Δt	AUC (τ)			AARD (τ)			MRD (τ)			AUC (LMJM)
(year)		0.25	0.50	0.75	0.25	0.50	0.75	0.25	0.50	0.75	
9	1	0.726	0.713	0.712	0.357	0.327	0.327	0.035	0.032	0.034	0.717
	2	0.685	0.671	0.670	0.286	0.255	0.255	0.028	0.024	0.025	0.676
	3	0.669	0.654	0.654	0.257	0.227	0.228	0.027	0.022	0.023	0.659
12	1	0.770	0.756	0.754	0.434	0.402	0.400	0.056	0.053	0.053	0.761
	2	0.721	0.703	0.703	0.345	0.303	0.304	0.044	0.039	0.039	0.710
	3	0.699	0.680	0.680	0.307	0.266	0.267	0.042	0.035	0.035	0.687
14	1	0.797	0.784	0.784	0.487	0.463	0.464	0.071	0.068	0.069	0.789
	2	0.748	0.731	0.732	0.394	0.355	0.357	0.059	0.054	0.054	0.738
	3	0.714	0.695	0.697	0.331	0.288	0.293	0.059	0.049	0.050	0.704

5.5 Discussion

In this work, we develop a Bayesian algorithm to make subject-specific dynamic predictions of recurrent event probability based on a new version of JM that uses LQMM for the longitudinal outcome , i.e. QRJM. Derivation of the Bayesian algorithm is based on the location-scale representation of the ALD for the longitudinal quantile regression. The Bayesian algorithm, which is straightforwardly implemented in JAGS software, uses a piecewise constant baseline hazard function in the recurrent events submodel. However, other functional forms can also be considered and the integration of the hazard function can be approximated using numerical integration such as Simpson's rule. Moreover, our predictions of recurrent event probability are based on the entire longitudinal trajectory as well as the recurrent events history of a subject and can be dynamically updated when new data from either (or both) outcome is available. In the real data application, we illustrate the flexibility of the QRJM and its advantages over the LMJM by jointly modeling the risk of CHD recurrences and longitudinal SBP measurements. QRJM is able to provide more insight into the disease progression and the association between the two disease processes in terms of various quantile-based estimations and dynamic predictions.

The significance of current work can be interpreted from two perspectives. First of all, the idea of personalized dynamic predictions of recurrent event probability finds its practical importance in disease control and prevention. Event prediction using commonly collected biomarkers can provide clinicians with continuously updated "disease progression" information potentially allowing them to make appropriately timed intervention decisions for individual subjects. Utilizing the dynamic predictions based on the proposed QRJM framework, we obtain subject-specific predictions of event risk that would actually allow physicians to target or tailor medical treatment based on a specific patient profile. In addition, our novel extension of traditional JM with LQMM adds more flexibility to the modeling framework and allows us to investigate specific subgroup of patients of interest. One one hand, patients with extremer biomarker measurements are usually at higher risk of disease thus deserve additional attention in health care. On the other hand, if treatment effect on tails of the outcome is significantly different from the average effect, treatment strategy should also be adjust accordingly to achieve better clinical outcome.

In summary, the QRJM proposed in this paper is a good alternative to the LMJM when either the normality assumption of the errors term is concerned or the conditional quantiles are more relevant to research question. And the dynamic predictions algorithm can be a highly potent tool in personalized disease control and prevention. The current version of QRJM uses LQMM and Cox PHM for the longitudinal and recurrent event processes respectively. However, other functional forms for both outcomes can also be considered to extend the proposed method. For example, in the longitudinal process, nonlinear QR (Koenker and Park, 1996) or even nonparametric QR (Le et al., 2005) models can be used to free the assumption of linearity. In the recurrent events submodel, accelerated failure time model can be considered when the proportionality assumption is violated and counting process approach is another nonparametric option.

Appendices

A Additional Simulation Results

Table 5.4: Simulation study: Inference results for data generated from various error distributions*.

	$ALD(0, 1, 0.25)$				$ALD(0, 1, 0.50)$				$ALD(0, 1, 0.75)$				$\mathcal{N}(0, 1)$			
	Bias	SE	MSE	CP	Bias	SE	MSE	CP	Bias	SE	MSE	CP	Bias	SE	MSE	CP
	Coefficients for longitudinal process															
β_1	-0.001	0.064	0.004	0.920	-0.001	0.061	0.004	0.920	-0.005	0.064	0.005	0.940	0.000	0.054	0.003	0.930
β_2	-0.003	0.116	0.011	0.970	-0.002	0.113	0.011	0.965	-0.001	0.116	0.015	0.920	0.006	0.106	0.012	0.940
β_3	0.020	0.048	0.003	0.950	0.015	0.044	0.002	0.950	0.027	0.048	0.003	0.900	0.009		0.009	0.027
0.001	0.920															
σ	0.001	0.022	0.001	0.970	0.002	0.022	0.001	0.940	0.003	0.022	0.001	0.940	–	–	–	–
	Coefficients for recurrent event process															
γ	0.007	0.052	0.004	0.920	0.008	0.050	0.003	0.915	0.005	0.052	0.002	0.955	-0.007	0.037	0.002	0.930
r_0	-0.017	0.093	0.008	0.940	-0.016	0.087	0.008	0.940	-0.003	0.096	0.010	0.935	0.022	0.063	0.006	0.900
α	0.003	0.051	0.003	0.950	0.003	0.047	0.003	0.940	0.001	0.051	0.003	0.940	-0.014	0.033	0.002	0.850

*data are fitted with true model for ALD distributed error and with ALD(0, 1, 0.50) for standard normal error.

B Study Cohort Characteristics

Table 5.5: Baseline characteristics of study cohort with stratification by SBP level

		SBP groups (mm Hg)			
Characteristics†	Total ($n = 657$)	< 120 ($n = 133$, 20.2%)	[120, 140) ($n = 217$, 33.0%)	≥ 140 ($n = 307$, 46.7%)	p-value*
Age	56.4 (5.8)	55.0 (5.7)	55.7 (6.0)	57.4 (5.4)	<0.001
SBP	135.9 (18.5)	110.5 (6.9)	129.2 (5.7)	151.6 (11.2)	<0.001
Cholesterol (mg/dL)	215.9 (41.7)	215.1 (42.1)	214.0 (42.0)	217.6 (41.7)	0.60
Gender (male)	341 (51.9)	64 (48.1)	117 (53.9)	160 (52.1)	0.57
Ever smoke (yes)	379 (57.5)	81 (60.9)	126 (58.1)	172 (56.0)	0.63
Hypertension medication (yes)	445 (67.7)	132 (99.2)	196 (90.3)	117 (38.1)	<0.001
Diabetes (yes)	90 (13.7)	15 (11.3)	27 (12.4)	48 (15.6)	0.38

†mean (sd) for continuous variables and frequency (percentage) for categorical variables.

*Comparing three SBP groups; ANOVA test for continuous variables and χ^2 test for categorical variables.

C Inference Results for ARIC Data

Inference results for five different quantiles (0.05, 0.25, 0.50, 0.75, and 0.95) are shown in Table 5.6. In the longitudinal SBP process, older participants generally have higher SBP level and this effect of baseline age is consistently positive across five quantiles of SBP. For example, one year increase in baseline age is associated with 0.036 (95% CI: (0.026, 0.047)) unit increase in the median (i.e., $\tau = 0.50$) of standardized SBP in the study cohort when controlling for other covariates. Total cholesterol level is negatively associated with SBP; however, the effects are not significant for all five quantiles. In general, people who took hypertension medications have significantly lower SBP and the effect of taking hypertension medications increases as the SBP level increases. It is interesting to see that time has a significantly positive effect on higher quantile of SBP (i.e., $\tau = 0.75$ and 0.95) while for lower quantiles ($\tau = 0.05$, 0.25, and 0.50) the effect is not significant. This can be an important indication that among the hypertension patients who originally have higher SBP deteriorate even faster than those with lower SBP.

In the recurrent CHD event process, we see all positive association between the five conditional quantiles of SBP and the risk of CHD event, which coincide with our expectation as well as previous studies using ARIC data. While the degree of association varies among the conditional quantiles and is strongest on the conditional median of SBP (relative risk:1.25, 95% CI: (1.02, 1.53)) in our case. For other regression covariates, diabetic patients are at significantly higher risk of having recurrent CHD event compared with non-diabetic. For example, when controlling for other factors, at $\tau = 0.5$ the risk of having additional CHD event is 2.6 times higher (exp(0.95), 95% CI: (1.57, 4.45)) for people with diabetes than those who are free of the disease. Although, males and ever smokers are also at higher risk of experiencing CHD events, the effects are not statistically significant compared with females and never smokers respectively.

Table 5.6: ARIC data analysis: Parameter estimation and 95% credible interval (in parenthesis) from the QRJM at five quantiles with SBP as the longitudinal biomarker.

	$\tau = 0.05$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.95$
Coefficients for longitudinal SBP process					
Intercept	−0.374	−0.023	0.447	0.872	1.187
	(−0.478, −0.274)	(−0.118, 0.074)	(0.352, 0.554)	(0.775, 0.978)	(1.079, 1.300)
Age$_0$	0.035	0.034	0.037	0.040	0.043
	(0.026, 0.044)	(0.025, 0.044)	(0.028, 0.047)	(0.030, 0.050)	(0.031, 0.052)
Total cholesterol (mg/dL)	−0.020	−0.026	−0.022	−0.013	−0.022
	(−0.073, 0.033)	(−0.081, 0.032)	(−0.078, 0.037)	(−0.073, 0.047)	(−0.076, 0.032)
Hypertension medicine	−0.583	−0.652	−0.725	−0.730	−0.787
	(−0.710, −0.467)	(−0.773, −0.538)	(−0.842, −0.609)	(−0.868, −0.593)	(−0.924, −0.660)
Follow- up time (yr)	0.008	0.006	0.011	0.016	0.019
	(−0.003, 0.018)	(−0.006, 0.019)	(−0.001, 0.022)	(0.004, 0.029)	(0.005, 0.033)
Coefficients for recurrent CHD process					
Male	0.191	0.185	0.160	0.132	0.110
	(−0.152, 0.548)	(−0.170, 0.528)	(−0.187, 0.507)	(−0.205, 0.477)	(−0.234, 0.458)
Ever smoke	0.291	0.271	0.216	0.165	0.163
	(−0.044, 0.641)	(−0.070, 0.613)	(−0.121, 0.552)	(−0.177, 0.493)	(−0.184, 0.485)
Diabetes	0.918	0.895	0.850	0.811	0.818
	(0.424, 1.399)	(0.409, 1.376)	(0.381, 1.349)	(0.352, 1.318)	(0.333, 1.301)
Association	0.163	0.207	0.226	0.205	0.162
	(−0.003, 0.332)	(0.011, 0.405)	(0.019, 0.428)	(0.034, 0.374)	(0.028, 0.288)

D JAGS Model File for Bayesian Inference

JAGS model file to fit QRJM of longitudinal and recurrent event data.

```
model{
      k1 <- (1-2*qt)/(qt*(1-qt))
      k2 <- 2/(qt*(1-qt))

      # prior of random effects
      for (i in 1:I){ # I: unique subject id
        # prior for random effects
        u[i] ~ dnorm(0, tau)
    } # end of loop i

    # longitudinal process, BQR mixed model using ALD representation
    for (j in 1:N_l){ # N_l: number of longitudinal observations
        er[j] ~ dexp(sigma)
        mu[j] <- beta1*X1_l[j] + beta2[X2_l[j]] + beta3*t[j] + u[id_l[j]]
         + k1*er[j]
        prec[j] <- sigma/(k2*er[j])
        y[j] ~ dnorm(mu[j], prec[j])
    } #end of j loop

  # recurrent events part, baseline hazard is set to constant c
    for(k in 1:I){
      for (l in (s[k]+1):s[k+1]){
        m1[l] <- beta1*X1[k]+beta2[X2[k]]+beta3*Ri1[l]+u[id_r[l]]
        m2[l] <- beta1*X1[k]+beta2[X2[k]]+beta3*Ri2[l]+u[id_r[l]]
        res[l] <- (exp(gamma*W[k]+alpha*m2[l])
                     -exp(gamma*W[k]+alpha*m1[l]))/(alpha*beta3)
        S[l] <- exp(-c*res[l])
        risk[l] <- c*exp(gamma*W[k] + alpha*m2[l])
        L[l] <- pow(risk[l], event[l])*S[l]/1E+08
        zeros[l] ~ dpois(-log(L[l]))
      } # end of l loop
    }#end of k loop

  # priors for other parameters
    alpha ~ dnorm(0, 0.001)
    beta1 ~ dnorm(0, 0.001)
    beta2[1] <- 0
    beta2[2] ~ dnorm(0, 0.001)
    beta3 ~ dnorm(0, 0.001)
    gamma ~ dnorm(0, 0.001)
    sigma ~ dgamma(0.001, 0.01)
    c ~ dunif(0.01, 10)
    tau <- pow(var, -2)
    var ~ dunif(0, 1000)
}
```

References

Bland, J. M. and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310.

Efendi, A., Molenberghs, G., Njagi, E. N., and Dendale, P. (2013). A joint model for longitudinal continuous and time-to-event outcomes with direct marginal interpretation. *Biometrical Journal*, 55(4):572–588.

Farcomeni, A. and Viviani, S. (2015). Longitudinal quantile regression in the presence of informative dropout through longitudinal–survival joint modeling. *Statistics in Medicine*, 34(7):1199–1213.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480.

Kim, S., Zeng, D., Chambless, L., and Li, Y. (2012). Joint models of longitudinal data and recurrent events with informative terminal event. *Statistics in biosciences*, 4(2):262–281.

Koenker, R. and Hallock, K. (2001). Quantile regression: An introduction. *Journal of Economic Perspectives*, 15(4):43–56.

Koenker, R. and Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448):1296–1310.

Koenker, R. and Park, B. J. (1996). An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics*, 71(1):265–283.

Kotz, S., Kozubowski, T., and Podgorski, K. (2012). *The Laplace Distribution And Generalizations: A Revisit with Applications to Communications, Economics, Engineering, And Finance*. Springer Science & Business Media.

Le, Q. V., Sears, T., and Smola, A. J. (2005). Nonparametric quantile regression. Technical report, Technical report, National ICT Australia, June 2005.

Pepe, M., Feng, Z., and Gu, J. (2008). Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by MJ Pencina et al., statistics in medicine. *Statistics in Medicine*, 27(2):173–181.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, volume 124, page 125.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829.

Rodriguez, C. J., Swett, K., Agarwal, S. K., Folsom, A. R., Fox, E. R., Loehr, L. R., Ni, H., Rosamond, W. D., and Chang, P. P. (2014). Systolic blood pressure levels among adults with hypertension and incident cardiovascular events: the atherosclerosis risk in communities study. *JAMA internal medicine*, 174(8):1252–1261.

Self, S. and Pawitan, Y. (1992). Modeling a marker of disease progression and onset of disease. *AIDS Epidemiology*, pages 231–255.

Taylor, J. M., Park, Y., Ankerst, D. P., Proust-Lima, C., Williams, S., Kestin, L., Bae, K., Pickles, T., and Sandler, H. (2013). Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics*, 69(1):206–213.

The ARIC investigators (1989). The atherosclerosis risk in community (ARIC) study: Design and objectwes. *American journal of epidemiology*, 129(4):687–702.

Tsiatis, A., Degruttola, V., and Wulfsohn, M. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, 90(429):27–37.

Wattanakit, K., Folsom, A. R., Chambless, L. E., and Nieto, F. J. (2005). Risk factors for cardiovascular event recurrence in the atherosclerosis risk in communities (ARIC) study. *American heart journal*, 149(4):606–612.

Wulfsohn, M. and Tsiatis, A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1):330–339.

Yang, L., Yu, M., and Gao, S. (2016). Prediction of coronary artery disease risk based on multiple longitudinal biomarkers. *Statistics in medicine*, 35(8):1299–1314.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447.

Zheng, Y., Cai, T., and Pepe, M. S. (2013). Adopting nested case–control quota sampling designs for the evaluation of risk markers. *Lifetime Data Analysis*, 19(4):568–588.

References

Brown, E. R., Ibrahim, J. G., and DeGruttola, V. (2005). A flexible b-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, 61(1):64–73.

Collins, F. S. and Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine*, 372(9):793–795.

Efendi, A., Molenberghs, G., Njagi, E. N., and Dendale, P. (2013). A joint model for longitudinal continuous and time-to-event outcomes with direct marginal interpretation. *Biometrical Journal*, 55(4):572–588.

Elashoff, R. M., Li, G., and Li, N. (2008). A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics*, 64(3):762–771.

Farcomeni, A. and Viviani, S. (2015). Longitudinal quantile regression in the presence of informative dropout through longitudinal–survival joint modeling. *Statistics in Medicine*, 34(7):1199–1213.

FDA (2013). Paving the way for personalized medicine. FDA's role in a new era of medical product development.

Fu, L. and Wang, Y.-G. (2012). Quantile regression for longitudinal data with a working correlation model. *Computational Statistics & Data Analysis*, 56(8):2526–2538.

Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics*, 8(1):140–154.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480.

Jung, S.-H. (1996). Quasi-likelihood for median regression models. *Journal of the American Statistical Association*, 91(433):251–257.

Kim, S., Zeng, D., Chambless, L., and Li, Y. (2012). Joint models of longitudinal data and recurrent events with informative terminal event. *Statistics in biosciences*, 4(2):262–281.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R. and Basset, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Koenker, R. and Hallock, K. (2001). Quantile regression: An introduction. *Journal of Economic Perspectives*, 15(4):43–56.

Koenker, R. and Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448):1296–1310.

Kotz, S., Kozubowski, T., and Podgorski, K. (2012). *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Springer Science & Business Media.

Kozumi, H. and Kobayashi, G. (2011). Gibbs sampling methods for bayesian quantile regression. *Journal of Statistical Computation and Simulation*, 81(11):1565–1578.

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Luo, Y., Lian, H., and Tian, M. (2012). Bayesian quantile regression for longitudinal data models. *Journal of Statistical Computation and Simulation*, 82(11):1635–1649.

Pepe, M., Feng, Z., and Gu, J. (2008). Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by MJ Pencina et al., statistics in medicine. *Statistics in Medicine*, 27(2):173–181.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, volume 124, page 125.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829.

Rizopoulos, D. and Ghosh, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in medicine*, 30(12):1366–1380.

Self, S. and Pawitan, Y. (1992). Modeling a marker of disease progression and onset of disease. *AIDS Epidemiology*, pages 231–255.

Taylor, J. M., Park, Y., Ankerst, D. P., Proust-Lima, C., Williams, S., Kestin, L., Bae, K., Pickles, T., and Sandler, H. (2013). Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics*, 69(1):206–213.

Tsiatis, A., Degruttola, V., and Wulfsohn, M. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, 90(429):27–37.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834.

Wulfsohn, M. and Tsiatis, A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1):330–339.

Yang, L., Yu, M., and Gao, S. (2016). Prediction of coronary artery disease risk based on multiple longitudinal biomarkers. *Statistics in medicine*, 35(8):1299–1314.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447.

Zheng, Y., Cai, T., and Pepe, M. S. (2013). Adopting nested case–control quota sampling designs for the evaluation of risk markers. *Lifetime Data Analysis*, 19(4):568–588.