Title

A latent class based imputation method under Bayesian quantile regression framework using asymmetric Laplace distribution for longitudinal medication usage data with intermittent missing values

Journal

Journal of Biopharmaceutical Statistics, 30(1)

ISSN

1054-3406

Authors

Lee, Minjae
Rahbar, Mohammad H
Gensler, Lianne S
et al.

Publication Date

2020-01-02

DOI

10.1080/10543406.2019.1684306

Peer reviewed

HHS Public Access

Author manuscript

J Biopharm Stat. Author manuscript; available in PMC 2021 January 01.

Published in final edited form as:

J Biopharm Stat. 2020 ; 30(1): 160–177. doi:10.1080/10543406.2019.1684306.

A latent class based imputation method under Bayesian quantile regression framework using asymmetric Laplace distribution for longitudinal medication usage data with intermittent missing values

MinJae Lee,
Division of Clinical and Translational Sciences, Department of Internal Medicine, University of Texas McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX, USA

Mohammad H. Rahbar,
The University of Texas Health Science Center at Houston, TX, USA

Lianne S. Gensler,
University of California, San Francisco, CA, USA

Matthew Brown,
Queensland University of Technology, Australia

Michael Weisman,
Cedars-Sinai Medical Center in Los Angeles, CA, USA

John D. Reveille
The University of Texas Health Science Center at Houston, TX, USA

Abstract

Evaluating the association between diseases and the longitudinal pattern of pharmacological therapy has become increasingly important. However, in many longitudinal studies, self-reported medication usage data collected at patients' follow up visits could be missing for various reasons. These pieces of missing or inaccurate/untenable information complicate determining the trajectory of medication use and its complete effects for patients. Although longitudinal models can deal with specific types of missing data, inappropriate handling of this issue can lead to a biased estimation of regression parameters especially when missing data mechanisms are complex and depend upon multiple sources of variation. We propose a latent class based multiple imputation (MI) approach using a Bayesian quantile regression (BQR) that incorporates cluster of unobserved heterogeneity for medication usage data with intermittent missing values. Findings from our simulation study indicate that the proposed method performs better than traditional MI methods under certain scenarios of data distribution. We also demonstrate applications of the proposed method to data from the Prospective Study of Outcomes in Ankylosing Spondylitis (AS) (PSOAS) cohort when assessing an association between longitudinal nonsteroidal anti-inflammatory drugs

MinJae.Lee@uth.tmc.edu.

(NSAIDs) usage and radiographic damage in AS, while the longitudinal NSAID index data are intermittently missing.

Keywords

Multiple Imputation; Intermittent Missing; Bayesian Quantile Regression; Latent Class; Asymmetric Laplace Distribution; Prospective Study of Outcomes in Ankylosing Spondylitis (PSOAS)

1 Introduction

Longitudinal cohort studies provide an opportunity for assessing the effect of pharmacological therapy on diseases. For example, nonsteroidal anti-inflammatory drugs (NSAIDs) are commonly used to manage inflammation and chronic pain in patients with ankylosing spondylitis (AS) [1]. For effect optimization, patients may need to take higher doses; this may carry a greater risk for side effects, most notably gastrointestinal and cardiovascular adverse events [2]. Therefore, it is important to explore how the level of use varies over time among patients, and assess its association with disease severity or progression. However, most commonly used measurement tools for medication usage rely on participant self-report and these self-reported medication usage data collected at patients' follow up visits could be missing for various reasons. Besides missed/skipped study visits, limited ability to self-report can also lead to missing or inaccurate/untenable information, which complicates determining the trajectory of medication use and its complete effects for patients. Due to these types of gaps and biases in participant reporting, statistical modeling to assess these measurements is challenging. These issues become even more critical in the assessment of change, since they may mask the longitudinal signal. In the Prospective Study of Outcomes in Ankylosing Spondylitis (PSOAS) [3, 4], which motivated our study, 69.3% of patients had missing NSAIDs intake data (NSAID index) for at least one time point among their follow up visits. An intermittent missing data pattern (i.e., a missing value was followed by an observed value) [5] was found for the majority of study patients in PSOAS cohort.

Compared to naive or ad-hoc methods for handling missing data including single imputation approaches that may introduce substantial bias or invalid study conclusion, it has been shown that model-based imputation techniques such as multiple imputation (MI) methods [6] provide more valid statistical inference [7, 8]. MI approaches have been widely used in medical research to better understand treatment effects in clinical trials [9, 10]. However, many of these were likelihood-based methods that focus on estimation of means assuming normality of the data distribution, which might be invalid for the non-normal data. In many observational studies, medication usage data based on dose and frequency, are usually not normally distributed. For the variables arising from this type of data source, applying transformation may not be a perfect solution to deal with complex data distributions. Specifically, NSAID usage data in PSOAS cohort contain a large number of zeros, and also transformation is not a solution to make data normal. Quantile regression models [11] have been used in longitudinal analyses as an important alternative to mean-based regression

models because of their flexibility for modeling non-normal data and heterogeneous conditional distributions. Its implementation for MI have been introduced by Wei *et al.* [12] for handling missing covariates in quantile regression for independent samples and Liu *et al.* [13] proposed a quantile regression in the presence of monotone missingness with sensitivity analysis. Bayesian methods can be easily adapted to treat missing data as additional unknown quantities for which a posterior distribution is estimated. A transition of the Bayesian approach to quantile regression can be implemented through modeling the error distribution using asymmetric Laplace distribution (ALD). Various estimation approaches of Bayesian quantile regression (BQR) have been developed [14, 15, 16] and extended to deal with missing data [17, 18]. To assess the effectiveness of medications on disease over time, it would be important to characterize its longitudinal pattern that could vary across individuals. This may lead to further investigation to identify groups of patients with similar medication usage patterns over time. The findings from this work can also be effectively used when we impute the intermittent missing data over follow up time points. For example, there could be an attempt to impute missing pieces of medication usage data for those with a persistent longitudinal medication usage, i.e., patients who reported no medication for all visits, or had consistently high (or low) level of medication intake over time. However, it would not be easy to impute missing values for patients who frequently changed their level of medication intake over time, unless we incorporate the longitudinal trajectory of data into imputation modeling. Group-based trajectory modeling has been developed to identify clusters of individuals using their longitudinal patterns, which also allows irregular spacing of measurements and missing data [19, 20, 21]. Assuming a heterogeneous population, latent class models can be also considered; a Bayesian two-part model has been recently introduced to analyze longitudinal medical expenditure data [22]. The latent class based multiple imputation approaches have been developed for missing categorical data [23, 24], and specifically those in large-scale assessment surveys [25, 26]. Two-stage multiple imputation [27], which accounts for qualitatively different types of missing categorical data (e.g. refusal vs. don’t know) has been also proposed. However, these were mainly for categorical data in a cross-sectional analysis. To our knowledge there are no published studies that have developed an imputation approach which specifically accommodates missing data under the joint latent class BQR modeling for multi-level (or longitudinal) designs. The development of a statistical approach to handle missing data while simultaneously controlling for unobserved heterogeneity, which also helps identify the cluster of longitudinal trajectories, is necessitated, especially for the longitudinal studies that involve effects of self-reported drug treatment on disease severity/progression. We adopt the idea of latent class framework to establish appropriate implementation of imputation techniques, through BQR model under a mixed effects structure, which helps avoid distributional assumptions and misspecification of error distributions.

In this paper, the objective is to propose a specific multiple imputation strategy for intermittent missing data, that incorporates latent class into BQR model, such that we can provide a better understanding of data associations, in a situation where we are interested in identifying different longitudinal data patterns over time, that may lead to different risks of disease outcomes. The focus here is to assess, through simulation studies, the performance of our proposed approach by comparing to other imputation methods and illustrate its

application to real life data from PSOAS, to achieve realistic situations while specifically evaluating the longitudinal association between NSAIDs intake and radiographic damage, assessed by modified Stoke Ankylosing Spondylitis Spine Score (mSASSS) [28].

2 Statistical Approach

2.1 Linear Quantile Regression Model

Let z_{ij}^* be the measurement for the i-th subject at time j. Suppose we define a linear regression model $z_{ij}^* = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \epsilon_{ij}$, $i = 1,\cdots,n;\ j = 1,\cdots,n_i$, where $\boldsymbol{x}_{ij}$ is a $p \times 1$ vector of covariates that can include the time of measurement, $\boldsymbol{\beta}$ is an unknown $p \times 1$ vector of regression parameters, and the random errors ϵ_{ij} are correlated within the subject to reflect the serial correlations of repeated measurements within each individual. If the τ-th conditional quantile of ϵ_{ij} given $\boldsymbol{x}_{ij}$ is assumed to be zero, a quantile regression model related to the τ-th quantile of variable, $q_\tau(z_{ij}^*)$, conditional on x_{ij}, has the form $q_\tau(z_{ij}^*) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_\tau$, $0 < \tau < 1$, where $\boldsymbol{\beta}_\tau$ is a vector of quantile-specific regression parameters corresponding to the coefficient $\boldsymbol{\beta}$ in the linear regression model above. We can define the objective function for longitudinal data z_{ij}^* as

$$Q_n(\boldsymbol{\beta}_\tau) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\rho_\tau(z_{ij}^* - \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_\tau). \quad (1)$$

The loss function $\rho_\tau(u) = u\{\tau - I(u \le 0)\}$, with $I(\cdot)$ being an indicator function, represents the contribution of residuals.

2.2 Longitudinal Mixed effect Model based on Bayesian Quantile Regression

Bayesian inference depends on prior and likelihood function. A transition of a Bayesian approach to quantile regression was implemented by modeling the error distribution using asymmetric Laplace distribution (ALD). ALD has good performance on data generated from various error distributions [29, 30] and theoretic justification. Since ALD can be defined as a mixture of normals based on the results of Laplace distribution with τ [31, 32] and it includes the common loss function that is used for quantile regression in its kernel, we can use it for Bayesian estimation in quantile regression models. The mean of the ALD can be determined as a linear function of mixed effect components that allows us to model longitudinal (or multilevel) data with multiple sources of variation. In this study, we propose the method based on ALD distribution for Bayesian modeling that utilizes Markov Chain Monte Carlo (MCMC) computational techniques. Assuming a random variable z^* has an asymmetric Laplace distribution, denoted ALD(μ, σ, τ), we define a probability density function:

$$f(z^*|\mu,\sigma,\tau) = \frac{\tau(1-\tau)}{\sigma}exp\left\{-\rho_\tau\left(\frac{z^*-\mu}{\sigma}\right)\right\},$$

where μ is the location parameter, σ is the scale parameter and $0 < \tau < 1$ is the skewness parameter. A loss function ρ_τ assigns weights τ and $1 - \tau$ to observation less and greater than μ, respectively, i.e., $P(z^* \leq \mu) = \tau$, regardless of the scale parameter σ.

ALD can be also represented as a scale mixture of normal distribution, i.e., $z_{ij}^* = \mu_{ij} + \kappa_1 e_{ij} + \zeta_{ij}\sqrt{\kappa_2 \sigma e_{ij}}$, where $\zeta_{ij} \sim N(0,1)$, $e_{ij} > 0$ is following an exponential distribution with mean σ (i.e. $e_{ij} \sim EXP(\sigma)$), ζ_{ij} and e_{ij} are independent. Scalars $\kappa_1 = \frac{1-2\tau}{\tau(1-\tau)}$ and $\kappa_2 = \frac{2}{\tau(1-\tau)}$ are dependent on τ. Based on this mixture, we define a linear mixed effect model that can be expressed as a linear function of the set of variables, $\boldsymbol{x}_{ij}$ and $\boldsymbol{v}_{ij}$ for fixed and random component, respectively:

$$z_{ij}^* = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \boldsymbol{v}_{ij}^T \boldsymbol{\gamma}_i + \kappa_1 e_{ij} + \zeta_{ij}\sqrt{\kappa_2 \sigma e_{ij}}, \quad (2)$$

where $\boldsymbol{\beta}$ is a vector for fixed effect variables, a random effect parameter for subject i, $\boldsymbol{\gamma}_i$ is following q-dimensional multivariate normal (MVN) distribution, $\boldsymbol{\gamma}_i | C_i \sim MVN_q(\mathbf{0}, \boldsymbol{\Sigma})$. For example, $\boldsymbol{\Sigma}$ is a 2x2 matrix for the random intercept-slope model (q=2). $\boldsymbol{\beta}$, $\boldsymbol{\gamma}_i$, and e_{ij} are mutually independent each other. Then, a likelihood for z_{ij}^*, following ALD distribution, can be expressed as

$$L(\boldsymbol{\beta}, \sigma | z_{ij}^*, \tau) = \frac{\tau(1-\tau)}{\sigma^N} exp\left\{ -\sum_{i=1}^{n} \sum_{j=1}^{n_i} \rho_\tau \left(\frac{z_{ij}^* - \boldsymbol{x}_{ij}^T \boldsymbol{\beta} - \boldsymbol{v}_{ij}^T \boldsymbol{\gamma}_i - \kappa_1 e_{ij}}{\sigma} \right) \right\}.$$

Considering σ as a nuisance parameter, the maximization of the likelihood above is equivalent to the minimization of the objective function $Q_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \rho_\tau \left(z_{ij}^* - \boldsymbol{x}_{ij}^T \boldsymbol{\beta} - \boldsymbol{v}_{ij}^T \boldsymbol{\gamma}_i \right)$. At any value of τ in (0,1), we can also define a normal distribution for z_{ij}^* as follows:

$$z_{ij}^* | e_{ij} \sim N(\boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \boldsymbol{v}_{ij}^T \boldsymbol{\gamma}_i + \kappa_1 e_{ij}, \kappa_2 \sigma e_{ij}),$$

where $\boldsymbol{\beta}$ is a τ-specific parameter of fixed effect variables. A linear mixed effect model (2) then leads the probability density function at τ determined as

$$f(z_{ij}^* | \beta, \gamma_i, e_{ij}, \sigma) = \frac{1}{\sqrt{2\pi\kappa_2 \sigma e_{ij}}} exp\left\{ \frac{(z_{ij}^* - \boldsymbol{x}_{ij}^T \boldsymbol{\beta} - \boldsymbol{v}_{ij}^T \boldsymbol{\gamma}_i - \kappa_1 e_{ij})^2}{-2\kappa_2 \sigma e_{ij}} \right\}.$$

A conditional distribution of z_{ij}^* can be defined as follows:

$$\prod_{i=1}^{n}\prod_{j=1}^{n_i} f\left(z_{ij}^{*}|\boldsymbol{\beta},\boldsymbol{\gamma}_i,e_{ij},\sigma\right)=\left(2\pi\kappa_2\sigma\right)^{-N/2}\left(\prod_{i=1}^{n}\prod_{j=1}^{n_i}e_{ij}\right)^{-1/2}$$
$$\times exp\left\{\frac{-1}{2\pi\kappa_2\sigma}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\frac{\left(z_{ij}^{*}-\boldsymbol{x}_{ij}^{T}\boldsymbol{\beta}-\boldsymbol{v}_{ij}^{T}\boldsymbol{\gamma}_i-\kappa_1 e_{ij}\right)^2}{e_{ij}}\right\}.$$

For the situation where a variable z_{ij}^{*} ranges between 0 and 1 (e.g. NSAID index data in PSOAS cohort), we can assume that the variable z_{ij}^{*} $(0 < z_{ij}^{*} < 1)$ has a logit-normal distribution and apply a normal approximation on the logit scale.

2.3 Bayesian Quantile Regression with Latent Class

A model (2) can be extended to allow for the latent classes by introducing a latent categorical variable, C_i that takes the values k $(k = 1, \cdots, K)$, when patient i belongs to class k. Given $f(z_{ij}^{*}|C_i = k,\gamma_i) = N(z_{ij}^{*};\mu_{ijk}+\kappa_1 e_{ij},\kappa_2\sigma_k e_{ij})$, where $\mu_{ijk} = \boldsymbol{x}_{ij}^{T}\boldsymbol{\beta}_k + \boldsymbol{v}_{ij}^{T}\boldsymbol{\gamma}_i$, a model with a latent class can be expressed as:

$$z_{ij}^{*} = \boldsymbol{x}_{ij}^{T}\boldsymbol{\beta}_k + \boldsymbol{v}_{ij}^{T}\boldsymbol{\gamma}_i + \kappa_1 e_{ij} + \zeta_{ij}\sqrt{\kappa_2\sigma_k e_{ij}}, \tag{3}$$

where parameter $\boldsymbol{\beta}_k$ is for fixed effect variables specific to class k, and $\boldsymbol{\gamma}_i|C_i$ is a random effect parameter for subject i, following $\mathrm{MVN}_q(\mathbf{0}, \boldsymbol{\Sigma}_k)$ with a class k-specific parameter $\boldsymbol{\Sigma}_k$.

For latent class modeling, we define a class indicator C_i which has a categorical distribution (Cat) taking the value k with probability $\boldsymbol{\pi}_{ik}$. When latent class models are used to assess the pattern of outcome measures, often latent class models are conducted without covariates, or relationships between class membership it covariates is assessed separately after class membership was estimated. However, this could make a latent class model misspecified. Covariates that influence, in theory, the latent class membership can be included in the model since the latent class models with covariates may allow us to classify an individual in one of the latent classes, on the basis of some individual characteristics. Use of covariates in latent class models can also help understand how different levels on the covariates predict subgroup membership [33, 34].

A probability $\boldsymbol{\pi}_{ik}$ can be determined through generalized logit model as follows with s-dimensional covariate vector, $\boldsymbol{w}_i$, which can include variables that affect the probability of belonging to a given class:

$$C_i \sim \mathrm{Cat}(\pi_{i1},\pi_{i2},\cdots,\pi_{iK}),\ \pi_{ik} = \frac{exp(\boldsymbol{w}_i^{T}\boldsymbol{\delta}_k)}{exp(\sum_{h=1}^{K}\boldsymbol{w}_i^{T}\boldsymbol{\delta}_h)}, \tag{4}$$

where C_i is an integer variable ranging from 1 to K and $\boldsymbol{\delta}_h$ $(h=1, \cdots, \mathrm{K};\ \boldsymbol{\delta}_1=\mathbf{0}$ for identifiability) is a vector of regression parameters for a covariate vector $\boldsymbol{w}_i$. We assume that

the number of classes, K is known. Bayesian model-selection strategies such as deviance information criterion (DIC) [35] can be used for determining the optimal value of K.

2.4 Multiple Imputation Process: MCMC Algorithm for Bayesian Quantile Regression with Latent Class

Given the fact that inferential machinery which is available for Bayesian parameter estimation extends to models with missing data, we can specify an appropriate joint model for the observed and missing data to model parameters in a usual way through MCMC integration. We use Gibbs sampling algorithm for estimations and imputations based on Bayesian quantile regression models. Further details regarding the rationale for the posterior inference in Bayesian quantile regression with asymptotic Laplace distribution was discussed by Kozumi *et al.* [36], and Gibbs sampling methods for BQR were introduced by Yang *et al.* [37]. The imputation through Gibbs sampling methodology requires the generation of a Markov chain from the posterior density. Under the Bayesian framework, we can derive the posterior distribution of the parameter through the prior distribution of parameters. Informative prior distributions of conjugate form can be assumed for the parameters. Specifically, we assume the prior conditional distribution of $\boldsymbol{\beta}_k$ as MVN($\boldsymbol{b}_0$, $\boldsymbol{B}_0$) = MVN($\mathbf{0}$, $I_{p\times p}$) and the prior distribution of $\mathbf{g}_i|C_i$ as MVN($\mathbf{0}$,$\boldsymbol{\Sigma}_k$) = MVN($\mathbf{0}$,$\phi_k^2 I_{q\times q}$), where $\phi_k^2 \sim \text{IG}(\nu_0, \omega_0)$ = IG(10, 0.1) (IG: inverse Gaussian distribution). In the prior distributions, these hyperparameters $\boldsymbol{\beta}_k$, $\boldsymbol{\gamma}_i$, $\boldsymbol{B}_0$, $\boldsymbol{\Sigma}_k$ are initially assumed to be known constant vectors and matrices. The prior distributions for other parameters are determined as: $\sigma_k \sim$ IG(c_0, d_0) = IG(1, 1), $e_{ij}|C_i \sim$ EXP($1/\sigma_k$) and $\boldsymbol{\delta}_k \sim$MVN($\boldsymbol{0}$,$\frac{9}{4} I_{s\times s}$). Based on this prior specification, z_{ij} is simulated from the Bayesian predictive distribution. Gelman [38] indicated that reasonable choices of prior distributions will have minor effects on posterior inferences with well-identified parameters and large sample sizes, but it is important to study the sensitivity of posterior inferences. To check the dependence on prior distributions we conducted a sensitivity analysis by comparing posterior inferences under different choices of prior distribution (e.g. informative prior distributions), and we found that posterior inferences were not sensitive to the choice of priors.

Our goal is to impute the missing values by taking independent draws from the distribution $f(z_{ij}|\boldsymbol{\delta}_k, C_i, \boldsymbol{\beta}_k, \sigma_k, \boldsymbol{\phi}_k, \boldsymbol{\gamma}_i, e_{ij})$. Quantile regression in Gibbs sampling is fitted separately, at each quantile level, based on a grid of l_n quantile levels (e.g. l_n =9), $0 < u_1 < \cdots < u_{l_n} < 1$. We also checked the sensitivity due to the choices of l_n and confirmed that the results of imputations were not dependent on the way the grid of l_n levels was defined. We carry the Gibbs sampling for a quantile level $u \in (u_l, u_{l+1})$ as follows:

Step (1) Initialize the parameters $\boldsymbol{\theta}^{(0)} = \left(\boldsymbol{\delta}_k^{(0)}, C_i^{(0)}, \boldsymbol{\beta}_k^{(0)}, \sigma_k^{(0)}, \boldsymbol{\phi}_k^{(0)}, \boldsymbol{\gamma}_i^{(0)}, e_{ij}^{(0)}\right)$ from the model with missing data.

Step (2) *Imputation Step* (update imputed values): Given $\boldsymbol{\theta}^{(r)}$ at the r-th iteration, sample $z_{ij}^{(r)}$ from $N(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}_k^{(r)} + \boldsymbol{v}_{ij}^T \boldsymbol{\gamma}_i^{(r)} + \kappa_1 e_{ij}^{(r)}, \kappa_2 \sigma^{(r)} e_{ij}^{(r)})$ for the missing

observations. $\boldsymbol{\beta}_k$ is u-th quantile-specific parameter estimator for class k, $\kappa_1 = \frac{1-2u}{u(1-u)}$ and $\kappa_2 = \frac{2}{u(1-u)}$.

Step (3) *Posterior Step* (update parameter estimates): Given the complete sample data $z_{ij}^{(r+1)}$, simulate the posterior parameter estimates $\boldsymbol{\theta}^{(r+1)}$. These new estimates are then used in the next imputation step. Sampling steps are as follows:

a. sample $\boldsymbol{\delta}_k^{(r+1)} \sim f(\boldsymbol{\delta}_k | z_{ij}^{(r+1)}, C_i^{(r)}, \boldsymbol{\beta}_k^{(r)}, \sigma_k^{(r)}, \phi_k^{(r)}, \boldsymbol{\gamma}_i^{(r)}, e_{ij}^{(r)})$

b. sample $C_i^{(r+1)} \sim f(C_i | z_{ij}^{(r+1)}, \boldsymbol{\delta}_k^{(r+1)}, \boldsymbol{\delta}_k^{(r)}, \sigma_k^{(r)}, \phi_k^{(r)}, \boldsymbol{\gamma}_i^{(r)}, e_{ij}^{(r)})$

c. sample $\boldsymbol{\beta}_k^{(r+1)} \sim f\left(\boldsymbol{\beta}_k | z_{ij}^{(r+1)}, C_i^{(r+1)}, \boldsymbol{\delta}_k^{(r+1)}, \sigma_k^{(r)}, \phi_k^{(r)}, \boldsymbol{\gamma}_i^{(r)}, e_{ij}^{(r)}\right)$

d. sample $\sigma_k^{(r+1)} \sim f\left(\sigma_k | z_{ij}^{(r+1)}, \boldsymbol{\beta}_k^{(r+1)}, C_i^{(r+1)}, \boldsymbol{\delta}_k^{(r+1)}, \phi_k^{(r)}, \boldsymbol{\gamma}_i^{(r)}, e_{ij}^{(r)}\right)$

e. sample $\phi_k^{2(r+1)} \sim f\left(\phi_k^2 | z_{ij}^{(r+1)}, \sigma_k^{(r+1)}, \boldsymbol{\beta}_k^{(r+1)}, C_i^{(r+1)}, \boldsymbol{\delta}_k^{(r+1)}, \boldsymbol{\gamma}_i^{(r)}, e_{ij}^{(r)}\right)$

Repeat Step (c) - (e) for classes k=1, ⋯, K.

f. sample $\boldsymbol{\gamma}_i^{(r+1)} \sim f\left(\boldsymbol{\gamma}_i | z_{ij}^{(r+1)}, \phi_k^{(r+1)}, \sigma_k^{(r+1)}, \boldsymbol{\beta}_k^{(r+1)}, C_i^{(r+1)}, \boldsymbol{\delta}_k^{(r+1)}, e_{ij}^{(r)}\right)$

g. sample $e_{ij}^{(r+1)} \sim f\left(e_{ij} | z_{ij}^{(r+1)}, \boldsymbol{\gamma}_i^{(r+1)}, \phi_k^{(r+1)}, \sigma_k^{(r+1)}, \boldsymbol{\beta}_k^{(r+1)}, C_i^{(r+1)}, \boldsymbol{\delta}_k^{(r+1)}\right)$

Return to Step (a) and repeat until convergence.

Step (4) Repeat Step (2) and Step (3) until the algorithm converges.

For the Gibbs sampling, we generate 700 iterations for each simulated dataset and after discarding a burn-in of the first 200 realizations of the sequence, we take the imputed values from M iterations to form the M imputed datasets. For example, the 201st, 301st, 401st, 501st, and 601st iteration can be taken for M=5 datasets. The length of the burn-in can be monitored using trace plots and autocorrelation function (ACF) plots for each parameter, in order to assess if it is sufficient to achieve convergence. All details related to conditional distributions of parameters $\boldsymbol{\theta} = \left(\boldsymbol{\delta}_k, C_i, \quad \boldsymbol{\beta}_k, \sigma_k, \phi_k^2, \boldsymbol{\gamma}_i, e_{ij}\right)$ are discussed in Appendix A.

For each missing value for z_{ij}, u is randomly sampled from an uniform distribution $u \sim$ UNIF(0, 1), and the imputed value of z_{ij} from the Gibbs sampling at corresponding quantile level $u \in (u_l, u_{l+1})$ is chosen for the final M imputed datasets. We can conduct any longitudinal regression analyses using these M imputed datasets. For our motivating example PSOAS study, the association between a variable with imputed values (e.g. NSAID index) and an outcome variable (e.g. radiographic damage score, mSASSS) can be assessed using each imputed dataset. To obtain the parameter estimates of interest, we define the combined MI estimator as a mean of M estimates [6]. i.e., $\hat{\boldsymbol{\alpha}}_{MI} = M^{-1} \sum_{m=1}^{M} \hat{\boldsymbol{\alpha}}_m$. However, for the variance of estimator and related p-value which is determined by the normality of

estimated parameters $\hat{\alpha}_{MI}$, we adopt a bootstrap method by resampling the paired observations with replacement based on 500 bootstrap samples, rather than a traditional approach. Wang and Feng [39] discussed the asymptotic properties of their proposed multiple imputation procedure based on the conditional quantile function and suggested bootstrapping methods for the variance estimation, as the asymptotic variance of MI estimators takes complex forms and it is difficult to estimate directly.

3 Simulation Studies

We conducted simulation studies to investigate the performance of our developed MI methods through different scenarios. Based on the distribution of longitudinal data in PSOAS cohort, we assume a longitudinal variable z_{ij}^*, given $C_i = k$ and γ_i, defined through a linear mixed effect regression model,

$$z_{ij}^* = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_k + \boldsymbol{v}_{ij}^T \boldsymbol{\gamma}_i + \epsilon_{ij}, \tag{5}$$

where class k-specific ($k = 1, \cdots, 4$) parameter $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \beta_{2k}, \beta_{3k})^T = (0.01, 0.5, 0.3, 0.02)$, which is determined at each class separately for a set of variables $\boldsymbol{x}_{ij} = (1, x_{1ij}, x_{2ij}, t_{ij})$, and a random intercept parameter $\gamma_i \sim N(0, \phi_k^2)$. We generated $\boldsymbol{x}_1$ from zero inflated negative binomial distribution, ZINB($\lambda = 20$, $\vartheta = 1$, $p_{stl0} = 0.35$) (ϑ: dispersion parameter; p_{stl0}: probability of structural zero), $\boldsymbol{x}_2$ from Bernoulli (0.7) and a time variable $\boldsymbol{t}$ was generated ranging between 0 and 8, where $i = 1, \cdots, n$; $j = 1, \cdots, 8$. For the class membership modeling which is based in equation (4), we generated a variable $k = 1, \cdots, 4$ based on a multinomial logit regression model using a variable $\boldsymbol{w} \sim$ Bernoulli(0.6) with parameters $\delta_2 = -2.1$, $\delta_3 = 1.6$, $\delta_4 = 3.9$. We considered the following three different scenarios of data distribution to generate data $\boldsymbol{z}^*$.

Scenario 1 Normal distribution: The error term, $\epsilon_{ij} \sim N(0, \sigma^2)$, where $\sigma^2 = 1$

Scenario 2 Asymmetric (Exponential) distribution: $\epsilon_{ij} = exp(\xi_{ij}) - 1$ and $\xi_{ij} \sim N(0, \sigma^2)$, where $\sigma^2 = 1$

Scenario 3 Asymmetric (Exponential) distribution, heteroscedastic covariance structure (i.e., covariance depends on a set of covariates): $\epsilon_{ij} = exp(\xi_{ij}) - 1$ and $\xi_{ij} \sim N(0, 1/(1 + x_{2ij}))$.

In order to define an outcome variable that mimics the distribution of mSASSS in PSOAS cohort, we assumed the following Poisson regression model

$$log(y_{ij}) = \alpha_0 + \alpha_1 z_{ij}^* + \alpha_2 x_{1ij}, \tag{6}$$

where $\alpha_0 = -3$, $\alpha_1 = 1.1$, $\alpha_2 = 3.1$, and z_{ij}^* denotes complete data which were generated based on the aforementioned three scenarios. We then produced missing values for z_{ij}^* based on the intermittent missing data pattern of NSAID index data that was found in PSOAS cohort. We postulated a logistic regression model, $logit(\eta) = \varphi_0 + \varphi_1 x_{2ij}$, where $\varphi_0 = -2.4$, $\varphi_1 = 2.5$ or 4.1. Under this setting (i.e. missing at random (MAR)), we were able to achieve

desired missing percentage of 30 or 50%. For example, one of the simulated data with 30% overall missing rate, there were 18.2% of subjects with missing data for only 1 visit, 32.8% had 2 visits, 25.2% had 3 visits, 16.2% had 4 visits, etc., and 4.2% of patients who had complete measurements. For each scenario, three hundred simulation datasets with sample size of 500 (based on PSOAS cohort) or 200 were generated.

For multiple imputation, we applied BQR model without latent class, which ignores the group-based trajectory (MI-BQR_1), as well as that with latent class incorporated (MI-BQR_2). Other imputation methods were further considered, that included MCMC-based MI method (MI-MCMC) [40, 41, 42], specifically multivariate imputation by chained equations (MICE) method [42]. MICE has been widely used with advances in software development, since it is useful for large imputation procedures and very flexible in a broad range of settings [43]. We also conducted the analysis using observed data only (OBSVD). Using imputed data generated from these different MI methods (or the observed data only), we conducted longitudinal regression analyses for each scenario, based on model (6). In order to assess the performance of each estimator, we calculated bias and ratio of the mean squared error (MSE) of the omniscient estimator (OMNI), a gold standard that is based on the data without missing values, to that of each estimator. Throughout we refer to this ratio of MSEs as relative efficiency (RE), which will be used for comparing the performance of each method. We assume all REs are lesser than 1, i.e., the estimator of the each method has higher MSE than the omniscient estimator, a gold standard, and the method with higher RE produces more efficient estimators.

Tables 1–3 show the results of simulation study from data with 30% or 50% missing values that were generated under MAR mechanism, for each of aforementioned Scenario 1-Scenario 3. MSE and relative efficiency (100xRE) of parameter a_0, a_1 and a_2 were calculated for each simulated dataset and then averaged to be presented in tables. When 30% of data were missing (Table 1), overall, the proposed BQR approaches (MI-BQR_1 and MI-BQR_2) produced more efficient estimators (i.e. higher REs) than other methods that were used for comparison. Specifically, with data based on normal distribution, relative efficiency of our MI methods for a_1, the coefficients of variable with missing values, was between 89.3% and 90.7%, while a relative efficiency of MI-MCMC was 87.3% and that of OBSVD was 83.3%. Although when data are from non-normal distribution (i.e. Scenario 2 and Scenario 3), the magnitude of REs for the proposed method decreased slightly, we still found higher relative efficiencies, ranging from 85% to 86.9% for a_1, compared to other methods with relative efficiencies ranging from 81.3% to 81.9% for OBSVD, from 82.8% to 84% for MI-MCMC. We also obtained similar patterns in relative efficiencies for other two coefficients, a_0 and a_2.

Similar findings were observed in the presence of 50% of data missing, as shown in Table 2. Our MI methods provided higher REs compared to the other methods across all three scenarios. It also demonstrates that the proposed method was not sensitive to the choice of data distribution, as compared to MI-MCMC that assumes normality of data. For example, relative efficiency of a_1 for MI-MCMC under Scenario 3 was about 13% lower than that from Scenario 1 (i.e. from 74.7 for Scenario 1 to 64.98 for Scenario 3), while the proposed method provided consistent REs (<5% change) over all three scenarios.

The performance of the proposed method was further assessed when a sample size was relatively small (i.e. $n = 200$), through various aforementioned scenarios (Table 3). Although MSEs were getting larger as a sample size decreases, the proposed approaches outperformed other methods.

4 Application to Prospective Study of Outcomes in Ankylosing Spondylitis (PSOAS) data

Ankylosing spondylitis (AS) is a chronic inflammatory disease characterized by inflammatory spinal pain that usually begins in the second to fourth decades of life. It can result in chronic spinal and joint pain and stiffness, leading to functional impairment and diminished quality of life, and in some patients, complete spinal fusion. In the PSOAS cohort, participants meeting the modified New York (mNY) Classification Criteria for AS [44] were enrolled from one of the five study sites (Cedars-Sinai Medical Center in Los Angeles, California, the University of Texas McGovern Medical School at Houston (UTH), the NIH Clinical Center, University of California at San Francisco (UCSF), and the Princess Alexandra Hospital in Brisbane, Australia (PAH) and were followed for up to 15 years (through two cycles of NIH funding: 2002–2006 and 2007–2016). At each study visit, spaced 6 months apart, the patients underwent a comprehensive clinical evaluation for disease activity, spinal mobility and functional impairment. Self-reported outcomes were measured at 6-month intervals, and radiographic data including an anterior posterior (AP) pelvis X-ray, AP and lateral lumbosacral X-ray and lateral cervical spinal X-ray were collected every 2 years, in order to assess longitudinal radiographic damage which was defined by scoring the modified Stoke Ankylosing Spondylitis Spine Score (mSASSS) [28] and the Bath Ankylosing Spondylitis Radiology Index (BASRI) [45]. All medications and supplements taken by the patient, including NSAIDs and tumor necrosis factor inhibitors (TNFi), as well as laboratory test results of C-reactive protein (CRP) levels and erythrocyte sedimentation rate (ESR) were determined at each clinical visit.

One of the objectives of PSOAS, was to evaluate factors associated longitudinal radiographic severity as well as the rate of progression in AS patients. Specifically, we focused on evaluating the longitudinal association between NSAIDs usage and the extent of radiographic damage that is assessed by mSASSS values at each X-ray visit, while controlling for the potential confounders. We considered analysis cohort of 536 patients who were confirmed AS by mNY criteria and had at least 2 years of radiologic follow up data available (as of August 2016) to be able to determine patient's disease progression. However, we faced with a challenge in analyzing NSAID index data in relation to mSASSS. We found that 69.3% of patients had missing NSAID index data for at least their one visit; 12.1% of patients had missing NSAIDs intake data at just one visit, 8.24%, 11.3%, 7.28%, 9.2%, 9.2% and 12% of patients who had missing at 2, 3, 4, 5, 6 and >6 visits, respectively. As PSOAS study is based on dynamic cohort [46] where patients enter and leave over time based on their qualifying status, and also patient recruitment and follow-up are currently still ongoing, the number of visits (i.e. follow-up duration) differs by subjects. Since medication usage data were supposed to be collected every 6 months, we define a completeness of NSAID index data for each subject if there is no missing visits between the first and last

available visits. An important feature of NSAIDs usage data was that it has an intermittent missing pattern. We examined whether there was any other specific missing pattern (e.g. monotonic pattern) besides intermittent missing, but no other notable pattern was appeared. Additionally, we were able to find that 16% of patients never took NSAIDs (index <0.15%), 10% had continuously high (index ≥ 50%) NSAIDs use, 5% had continuously low (index < 50%) use of NSAIDs over time, and rest of them had variable NSAIDs use pattern.

Missing NSAIDs intake data across visits in PSOAS cohort were imputed based on three different approaches (MI-MCMC, MI-BQR$_1$, MI-BQR$_2$) that are described in section 3.

For imputation procedure for NSAID index data, BQR was modeled as a function of covariates that include time, sex, race, age, co-morbidities, education, smoking status, other medications use and mSASSS. In practice, it is important to use all available information to build imputation model [6, 47, 48]. The imputation model can also include the covariates and outcome of the potential analysis models even if they have limited predictive power [49]. Figure 1 shows one of the trace and autocorrelation function (ACF) plots for selected parameters of $\boldsymbol{\beta}$ and σ^2, which indicate good performance of imputed iteration.

For the latent class model comparisons, deviance information criterion (DIC) [35] was calculated for multiple models (from one- to five-class), and we found that four-class model was optimal with the lowest DIC value, i.e., 43863, 41798, 33639, 30553 and 32810 for one-, two-, three-, four-, and five-class, respectively. Group membership of each class was 30.44% (Group 1), 14.68% (Group 2), 28.84% (Group 3) and 26.03% (Group 4). These four classes corresponded to an increasing tendency to NSAID index; median values of NSAID index for each of four group were 1, 3, 25 and 100, respectively. The results from the latent class membership model, which was built based on the significant variables in classifying the patients, indicated that patients who had lower baseline disease activity score (Bath Ankylosing Spondylitis Disease Activity Index (BASDAI); the higher score, the worse the symptoms) or longer disease duration were more likely to be classified into Group 1. Though not statistically significant, patients who reported no TNFi use at the baseline visit, were less likely to be in Group 1.

Figure 2 displays how NSAID index missing data were imputed for each latent class based on the proposed method MI-BQR$_2$; a distribution of NSAID index data is shown, distinguishing the observed data (green area) from imputed data (gray area). More zeros were imputed in Group 1 compared to other three groups, while higher values (>60) were imputed in Group 4. Lower NSAID index values (<30) were imputed in Group 2, but imputed values were spread out around 50 in Group 3. However, imputed values across these groups based on MI-BQR$_1$ method (data not shown), which latent class was not incorporated in, did not seem as distinctive as those from MI-BQR$_2$.

We were interested in whether longitudinal trajectory of NSAID index differs by latent class membership; mean trajectories of NSAID index over time for four groups were explored. Different longitudinal trajectories of NSAID index were observed across the groups; patients in Group 1 had minimal or no use over time, and Group 3 had low use, patients in Group 4 were on high use (Group 4) persistently over time, while patients in Group 2 started with

low level and then had decreasing trend over time. The association of the assigned class membership with other variables (e.g. disease progression) was further investigated. Progressors (42%) were defined if patients had at least one interval (2 sets of X-rays), where 2 or more mSASSS units increase within 24 months was found. The proportions of progressors were 44.1%, 34.9%, 43.2%, 37.8% in Group 1, 2, 3, 4, respectively. This finding shows that among the patients in Group 4 who were on high use of NSAID over time, lesser progressors were found, compared to those in Group 1 who had minimal or no use over time. The longitudinal trajectories among progressors were also different from those in non-progressors. Figure 3 displays mean NSAID index trajectory for A: progressors and B: non-progressors, separately by four groups of patients that were identified based on our proposed method MI-BQR$_2$. In high use group (Group 4), NSAIDs usage in progressors decreased over time, while it was not changing among non-progressors. For Group 3, NSAIDs usage was increasing faster during first 6 years among progressors, but this trend was not found for non-progressors. Progressors' NSAIDs intakes decreased much slower compared to non-progressors' in Group 2.

Using imputed NSAID index data, we finally assessed the longitudinal association between NSAID usage and mSASSS as an outcome variable, while controlling for potential confounding factors, that included BASDAI, CRP levels, TNFi usage, as well as demographic information such as sex, race, disease duration, co-morbidity, education and smoking status. Multivariable mixed effect Poisson regression models were conducted to account for the correlations of repeated measurements within a patient. In this analysis, NSAID index data were divided into three categories based on the level of intensity: no use, low (<50%) and high (≥50%), suggested by clinicians as it was believed to be a clinically relevant way. And we also found the association between NSAID index and mSASSS in log scale was not linear. Table 4 shows the adjusted rate ratios (RRs) with their 95% confidence intervals (CIs) and p-values from the multivariable analysis based on only observed data (OBSVD), and imputed NSAID index data by each of two methods: MI-MCMC and MI-BQR$_2$.

There were noticeable differences across these three methods in the estimates and corresponding p-values for NSAID use. The results from the proposed method (MI-BQR$_2$) indicated that high NSAID index (≥50%) is inversely associated with radiographic damage compared to no use (adjusted RR=0.9; 95% confidence interval (CI)=[0.82, 0.98]; p=0.017), while the other methods did not result in a significant association (p=0.759 for OBSVD; p=0.245 for MI-MCMC). This significant association was diminished when low NSAID index (<50%) was compared to no use (adjusted RR=1.03; 95% confidence interval (CI)=[0.91, 1.15]; p=0.646). The results of imputed NSAID index data from MI-BQR$_2$ were similar to those from MI-BQR$_1$ (data not shown). We believed that the difference was not big enough to capture a significant contribution of the latent class because it may be attenuated due to the original imputed NSAID index values being categorized. Hence, we conducted analyses after breaking the NSAID index down further, i.e., no use, low (<50%), moderate (50%-75%) and high (>75%), and found differences in estimates between two methods, MI-BQR$_1$ and MI-BQR$_2$; when high NSAID use (>75%) was compared to no use (0%), adjusted RR from MI-BQR$_2$ was 0.90 (p=0.034) and that from MI-BQR$_1$ was 0.92 (p=0.05). Even though the proposed imputation approach was developed under the mixed

effect model framework specifically for multi-level (or longitudinal) designs, we also further carried out imputations that is based on MI-BQR$_2$ under a fixed effect only model to show what would happen. The significant association of high NSAID usage with lower radiographic damage that was found from the proposed method MI-BQR$_2$ disappeared when fixed effect only model was used (adjusted RR=0.98, p=0.6).

5 Discussion

It has become increasingly important to study longitudinal treatment effects and evaluate an association between diseases and longitudinal patterns of pharmacological therapy. In longitudinal studies, it is possible that medication usage data are incompletely collected during study visits, which can introduce an intermittent missing data pattern. In this article we have proposed the use of a multiple imputation approach that specifically accommodates intermittent missing data under the joint latent class Bayesian quantile regression (BQR) frame work, in order to account for unobserved heterogeneity into the imputation procedure.

We used Gibbs sampling for the estimation and imputation through Bayesian quantile regression. Metropolis Hastings (M-H) can be also used for modeling our BQR model [16]. The advantage of the Gibbs sampler over metropolis has been discussed to indicate that Gibbs sampling doesn't have the convergence issue as M-H algorithm has, and unlike M-H algorithm which has proposal distribution selection, the proposal distribution of Gibbs sampling is simply taken to be the conditional distributions of the target distribution. Each density to be sampling is of low dimension and thus it is relatively easy and efficient to sample from it.

Our findings from the simulation study presented that the proposed method performs better than other methods that were used for comparisons, by having a higher relative efficiency. Since different quantiles are used for imputation rather than mean values, our approach was found not sensitive to the choice of the error structure as compared to the approach that assumes normality of data. It was more robust than completely parametric methods when dealing with heteroskedastic variance. Though the proposed imputation algorithm is relatively straightforward, a statistical package will be further developed to help users easily implement the proposed approach. We also demonstrated an application of our imputation method to real data from PSOAS by examining the longitudinal association between NSAID usage and radiographic damage for ankylosing spondylitis patients, in a situation where the NSAID index data for some patients were incompletely collected during the follow up time points.

Despite of aforementioned advantages, there are limitations of the proposed method due to the specific nature of the Bayesian ALD approach; most notably is the deliberate misspecification of the likelihood on the implications for posterior consistency and the coverage probabilities of the resulting posterior distributions [50]. We acknowledge that it is important to identify the model carefully for missing data process and interpret the analysis results cautiously. Computational burden related to MCMC algorithms for estimating regression parameters should be also carefully considered. Another drawback of the proposed method is that, as quantiles are fitted separately, the conditional quantile curves are

not smooth and the fitted regression lines may cross, which violates the basic probabilistic rule and causes problems for inference in practice [50]. The usual approach to deal with this problem is to simultaneously fit several quantiles [51]. Even though many of these methods give raise to new problems in terms of computation or data requirements [52], it would be interesting to separate quantiles well such that quantile crossing can be less likely to occur. We also acknowledge that Bayesian nonparametric (BNP) models [54], which a single model that can adapt its complexity to the data rather than comparing models that vary in complexity, can be used for the imputations to avoid the need to explicit latent classes. However, as well as improving the accuracy of the imputations for missing medication data over time, the main focus of our study was on providing the longitudinal trajectory of medication intake among the patients, especially for the longitudinal studies that involve effects of a drug treatment on disease severity or progression over time. In clinical studies, MAR is a reasonably assumptions for intermittent missing data, but the proposed method can be used to handle unobserved heterogeneity, which also helps better understanding of measurement errors in self-reported medication usage data.

There is a growing interest to develop MI methods that impute missing data across multiple variables (i.e. medications) while accounting for the correlations among them, which can be also extended by our proposed method. When we assume that the number of classes is known, the methods may also be further considered through reversible jump Markov chain Monte Carlo (RJMCMC) techniques [53].

Acknowledgments

We acknowledge the support provided by the Biostatistics/ Epidemiology/ Research Design (BERD) component of the Center for Clinical and Translational Sciences (CCTS) for this project. CCTS is mainly funded by the NIH Centers for Translational Science Award (UL1 TR000371), the National Center for Advancing Translational Sciences (NCATS). We also acknowledge the grants from the United States Department of Health and Human Services, National Institutes of Health (NIH), National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS), P01-052915-06, and the Spondylitis Association of America and the Russell/Engleman Rheumatology Research Center at UCSF. The authors also would like to recognize and thank those who have provided with their thoughtful comments and suggestions.

A: Appendix A: Conditional Distributions of Parameters

Following notation in Neelon *et al.* [22] and Luo *et al.* [55], the derivation of conditional distributions can be determined as below. [label=)]

1. Conditional density of $\boldsymbol{\delta}_k$:

$$\pi\left(\delta_k \mid \cdot\right) \propto \prod_{i=1}^{n} \left[Pr\left(C_i = k \mid \delta_k\right)\right]^{I_{C_i = k}} \pi\left(\delta_k\right)$$

$$= \prod_{i: C_i = k} \left(\frac{exp\left(w_i^T \delta_k\right)}{\sum_{h=1}^{K} exp\left(w_i^T \delta_h\right)} \right) \mathrm{MVN}\left[\delta_k; \mathbf{0}, \frac{9}{4} \boldsymbol{I}_{o \times o}\right],$$

where MVN(δ_k; ·) is an o-dimensional multivariate normal distribution evaluated at δ_k. Since this full conditional distribution does not have a closed form, δ_k as a vector was updated through a random walk Metropolis algorithm using a

multivariate-$t_3(s_g \boldsymbol{T}_k)$ proposal density centered at the previous value, $\delta_k^{(r-1)}$, where s_g and $\boldsymbol{T}_k$ are determined based on Neelon *et al.* [22].

2. Conditional density of C_i for $i = 1, \cdots, n$:

$$\pi(C_i | \cdot) = Pr(C_i = k | \cdot) = \text{Cat}(p_{ik}),$$

where

$$p_{ik} = \frac{\pi_{ik}(\delta_k)[\prod_{j=1}^{n_{ik}} f(z_{ij}|\boldsymbol{\beta}_k, \boldsymbol{\gamma}_i, e_{ij}, \sigma_k]\text{MVN}(\boldsymbol{\gamma}_i; \mathbf{0}, \boldsymbol{\Sigma}_k)}{\sum_{h=1}^{K} \pi_{ih}(\delta_h)[\prod_{j=1}^{n_{ik}} f(z_{ij}|\boldsymbol{\beta}_h, \boldsymbol{\gamma}_i, e_{ij}, \sigma_h]\text{MVN}(\boldsymbol{\gamma}_i; \mathbf{0}, \boldsymbol{\Sigma}_h)},$$

$\boldsymbol{\pi}_{ik}(\delta_k) = Pr(C_i = k \mid \delta_k)$ as given in (1). If there are no class membership covariates, i.e., $o = 1$ in 1), then update $\boldsymbol{\pi}_{ik}$ directly from a Dirichlet $(n_1 + \psi_1, \cdots, n_K + \psi_K)$ distribution, where $\psi_1, \cdots, \psi_K$ are prior hyperparameters and $n_k = \sum_{i=1}^{n} I_{(C_i = k)}$. $(\boldsymbol{\pi}_1 < \boldsymbol{\pi}_2, \cdots, < \boldsymbol{\pi}_K)$.

3. Conditional density of $\boldsymbol{\beta}_k$:

$$\pi(\boldsymbol{\beta}_k | \cdot) = \pi(\boldsymbol{\beta}_k | z_{ij}, \boldsymbol{\gamma}_i, e_{ij}, \sigma_k = \text{MVN}_p(\boldsymbol{b}_k, \boldsymbol{B}_k),$$

where $\boldsymbol{z}_k$ is a variable $\boldsymbol{z}$ for class k and

$$\boldsymbol{B}_k^{-1} = \frac{1}{\kappa_2 \sigma_k} \sum_{i=1}^{n_k} \sum_{j=1}^{n_i} \frac{\boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^T}{e_{ij}} + \boldsymbol{B}_0^{-1},$$

$$\boldsymbol{b}_k = \frac{1}{\kappa_2 \sigma_k} \sum_{i=1}^{n_k} \sum_{j=1}^{n_i} \frac{\boldsymbol{x}_{ij}(z_{ij} - \boldsymbol{v}_{ij}^T \boldsymbol{\gamma}_i - \kappa_1 e_{ij})}{e_{ij}} + \boldsymbol{B}_0^{-1} \boldsymbol{b}_0.$$

4. Conditional distribution of parameter σ_k:

$$\pi(\sigma_k | z_{ij}, \boldsymbol{\beta}_k, \boldsymbol{\gamma}_i, \phi_k^2) \propto f(z_{ij} | \boldsymbol{\beta}_k, \boldsymbol{\gamma}_i, e_{ij}, \sigma_k) \pi(\sigma_k)$$
$$\sim \text{IG}(c_1, d_1),$$

where

$$c_1 = \frac{N_k}{2} + c_0,$$

$$d_1 = d_0 + \frac{1}{2\kappa_2}\sum_{i=1}^{n_k}\sum_{j=1}^{n_i}\frac{(z_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_k - \boldsymbol{v}_{ij}^T\boldsymbol{\gamma}_i - \kappa_1 e_{ij})^2}{e_{ij}},$$

c_0, d_0 are the parameters values for the prior distribution of σ_k.

5. Conditional distribution of parameter ϕ_k^2:

$$\pi(\phi_k^2|z,\boldsymbol{\beta}_k,\boldsymbol{\gamma}_i,\sigma_k) \propto \left(\prod_{j=1}^{n_k} f(\boldsymbol{\gamma}_i|\phi_k^2)\right)\pi(\phi_k^2)$$

$$\sim \text{IG}(\nu_1,\omega_1),$$

where

$$\nu_1 = \frac{1}{2}J_q + \nu_0,$$

$$\omega_1 = \omega_0 + \frac{1}{2}\sum_{j=1}^{n_k}\frac{\boldsymbol{\gamma}_j^T\boldsymbol{\gamma}_j}{2},$$

q is the dimension of the random effects and ν_0, ω_0 are the parameters values for the prior distribution of ϕ_k^2.

6. Conditional distribution of $\boldsymbol{\gamma}_i$:

$$\pi(\boldsymbol{\gamma}_i|\cdot) \propto f(z_{ij}|C_i = k,\boldsymbol{\beta}_k,\boldsymbol{\gamma}_i,\boldsymbol{e}_{ij},\sigma_k)\text{MVN}(\boldsymbol{\gamma}_i;\mathbf{0},\boldsymbol{\Sigma}_k).$$

Conditional on $C_i = k$, $\boldsymbol{\gamma}_i$ was updated using a random walk Metropolis algorithm based on a multivariate-$t_3(s_\gamma \boldsymbol{R}_k)$ proposal density centered at the previous value, $\boldsymbol{\gamma}_i^{(r-1)}$, where the scale matrix $\boldsymbol{R}_k$ was estimated using the inverse information matrix obtained from traditional model fit and s_γ is a scaling factor used to achieve optimal acceptance rates [22].

7. Conditional distribution of e_{ij}:

$$\pi(e_{ij}|z_{ij},\boldsymbol{\beta}_k,\boldsymbol{\gamma}_i,\sigma_k) \propto f(z_{ij}|\boldsymbol{\beta}_k,\boldsymbol{\gamma}_i,e_{ij},\sigma_k)f(e_{ij}|\sigma_k)$$

$$\propto \frac{1}{\sqrt{e_{ij}}}exp\left\{\frac{-(z_{ij}-\boldsymbol{x}_{ij}^T\boldsymbol{\beta}-\boldsymbol{v}_{ij}^T\boldsymbol{\gamma}_i-\kappa_1 e_{ij})^2}{2k_2\sigma_k e_{ij}}\right\}exp\left\{-\frac{e_{ij}}{\sigma_k}\right\}$$

$$\propto \frac{1}{\sqrt{e_{ij}}}exp\left\{\frac{-1}{2}(\phi_{ij}^2 e_{ij}^{-1}+\psi_{ij}^2 e_{ij})\right\}.$$

→ kernel of a generalized inverse Gaussian (GIG) distribution, where

$$\phi_{ij}^2 = \frac{(z_{ij}-\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_k-\boldsymbol{v}_{ij}^T\boldsymbol{\gamma}_i-k_1 e_{ij})^2}{\kappa_2\sigma_k},$$

$$\psi_{ij}^2 = \frac{\kappa_1^2}{\kappa_2\sigma_k}+\frac{2}{\sigma_k}.$$

$$\pi(e_{ij}|z_{ij},\boldsymbol{\beta},\boldsymbol{\gamma}_i,\sigma_k) \sim \text{GIG}(0.5,\phi_k,\psi_k)$$

References

[1]. Miceli-Richard C, Dougados M. NSAIDs in ankylosing spondylitis. Clin Exp Rheumatol. 2002; 20(6 Suppl 28): S65–6.

[2]. Haroon N, Kim T, Inman RD. Continuance of non-steroidal anti-inflammatory drugs may reduce radiographic progression in ankylosing spondylitis patients on biological therapy. Arthritis Rheum 2011; 63(10): 1593–5.

[3]. Gensler LS, Ward MM, Reveille JD, Weisman MH, Davis JC Jr. Clinical, radiographic and functional differences between juvenile-onset and adult-onset ankylosing spondylitis: results from the PSOAS cohort. Ann Rheum Dis. 2008; 67(2):233–7. [PubMed: 17604288]

[4]. Rahbar MH, Lee M, Hessabi M, Tahanan A, Brown MA, Learch TJ, Diekman LA, Weisman MH, Reveille JD. Harmonization, data management, and statistical issues related to prospective multicenter studies in Ankylosing spondylitis (AS): Experience from The Prospective Study of Ankylosing Spondylitis (PSOAS) cohort. Contemporary Clinical Trials Communications 2018; 11(2018):127–135. [PubMed: 30094388]

[5]. Little RJA, Rubin DB. Statistical analysis with missing data. 2nd edition John Wiley & Sons; New York, 2002.

[6]. Rubin DB. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons: New York, 1987.

[7]. Mackinnon A The use and reporting of multiple imputation in medical research – a review. Journal of Internal Medicine 2010 268(6): 586–593. [PubMed: 20831627]

[8]. Rezvan RH, LEE KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. BMC Medical Research Methodology 2015; 15:30. [PubMed: 25880850]

[9]. Jørgensen AW, Lundstrøm LH, Wetterslev J, Astrup A, Gøtzsche PC. Comparison of Results from Different Imputation Techniques for Missing Data from an Anti-Obesity Drug Trial. PLOS ONE 2014; 9(11): e111964. [PubMed: 25409438]

[10]. Ayele BT, Lipkovich IA, Molenberghs G, Mallinckrodt VH. A Multiple-Imputation-Based Approach to Sensitivity Analyses and Effectiveness Assessments in Longitudinal Clinical Trials. Journal of Biopharmaceutical Statistics 2014; 24(2): 211–228. [PubMed: 24605966]

[11]. Koenker R and Park B. An interior point algorithm for nonlinear quantile regression. Journal of Econometrics 1996; 71(1–2): 265–283.

[12]. Wei Y, Ma Y, Carroll RJ. Multiple imputation in quantile regression. Biometrika 2012; 99(2): 423–438. [PubMed: 24944347]

[13]. Liu M, Daniewls MJ, Perri MG. Quantile regression in the presence of monotone missingness with sensitivity analysis. Biostatistics 2016; 17(1): 108–121. [PubMed: 26041008]

[14]. Komunjer I Quasi-Maximum Likelihood Estimation for Conditional Quantiles. Journal of Econometrics 2005; 128(1): 137–164.

[15]. Sriram K, Ramamoorthi RV, Ghosh P. Posterior Consistency of Bayesian Quantile Regression Based on the Misspecified Asymmetric Laplace Density. Bayesian Anal 2013; 8(2): 479–504.

[16]. Yu K, Moyeed RA. Bayesian quantile regression. Statistics & Probability Letters 2001; 54: 437–447.

[17]. Yuan Y, Yin G. Bayesian Quantile Regression for Longitudinal Studies with Nonignorable Missing Data. Biometrics 2012; 66(1): 4105–114.

[18]. Burgette LF, Reiter JP. Nonparametric Bayesian Multiple Imputation for Missing Data Due to Mid-Study Switching of Measurement Methods. Journal of the American Statistical Association 2012; 498: 439–449.

[19]. Nagin DS. Group-Based Modeling of Development. Harvard University Press: 2005.

[20]. Jones BL, Nagin DS. Advances in Group-based Trajectory Modeling and a SAS Procedure for Estimating Them. Sociological Methods & Research 2007; 35(4): 542–571.

[21]. Nagin DS, Odgers CL. Group-based trajectory modeling in clinical research. Annu Rev Clin Psychol. 2010; 6: 109–38. [PubMed: 20192788]

[22]. Neelon B, O'Malley AJ,Normand ST. A Bayesian Two-Part Latent Class Model for Longitudinal Medical Expenditure Data: Assessing the Impact of Mental Health and Substance Abuse Parity. Biometrics 2011; 67(1): 280–289. [PubMed: 20528856]

[23]. Gebregziabher M, DeSantis SM. Latent class based multiple imputation approach for missing categorical data. Journal of Statistical Planning and Inference 2010; 140: 3252–3262. [PubMed: 30555206]

[24]. Vidotto D, Vermunt JK, Kaptein MC. Multiple Imputation of Missing Categorical Data using Latent Class Models: State of the Art. Psychological Test and Assessment Modeling 2015; 57 (4): 542–576.

[25]. Si Y, Reiter JP. Nonparametric Bayesian Multiple Imputation for Incomplete Categorical Variables in Large-Scale Assessment Surveys. Journal of Educational and Behavioral Statistics 2013; 38(5): 499–521.

[26]. Manrique-Vallier D, Reiter JP. Bayesian multiple imputation for large-scale categorical data with structural zeros. Survey Methodology 2014; 40(1): 125–134.

[27]. Harel O, Chung H, Miglioretti D. Latent class regression: inference and estimation with two-stage multiple imputation. Biom J 2013; 55(4): 541–553. [PubMed: 23712802]

[28]. Creemers MC, Franssen MJ, van't Hof MA, Gribnau FW, van de Putte LB, van Riel PL. Assessment of outcome in ankylosing spondylitis: an extended radiographic scoring system. Ann Rheum Dis 2005; 64:127–9. [PubMed: 15051621]

[29]. Li Q, Xi R, Lin N. Bayesian regularized quantile regression. Bayesian Analysis 2010; 5: 533–556.

[30]. Ji Y, Lin N, Zhang B. Model selection in binary and Tobit quantile regression using the Gibbs sampler. Computational Statistics & Data Analysis 2012; 56: 827–839.

[31]. Andrews DR, Nallows CL. Scale mixtures of normal distributions. J. R. Statist. Soc. B 1974; 36: 99–102.

[32]. West M Outlier models and prior distributions in Bayesian linear regression. J. R. Statist. Soc. B 1984; 46: 431–439.

[33]. Dayton CM, MacReady GB. Concomitant-variable latent class models. Journal of the American Statistical Association 1988; 83(401) : 173–178.

[34]. Dayton CM, MacReady GB. Applied latent class analysis Edited by Hagenaars JA, McCurtcheon AL, Universiteit van Tilburg, The Netherlands, Allan L. McCutcheon, University of Nebraska, Lincoln. Publisher: Cambridge University Press, 2002; 213–233.

[35]. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. J. R. Statist. Soc. B 2002; 64(4):583–639.

[36]. Kozumi H, Kobayashi G. Gibbs sampling methods for Bayesian quantile regression. Journal of Statistical Computation and Simulation 2011; 81(11): 1565–1578.

[37]. Yang Y, Wang HJ, He X. Posterior Inference in Bayesian Quantile Regression with Asymmetric Laplace Likelihood. International Statisitcal Review 2016; 84(3):327–344

[38]. Gelman A Prior distribution. Encyclopedia of Envioromnentrics 2002; 3: 1634–1637.

[39]. Wang JH and Feng X. Multiple Imputation for M-regression with Censored Covariates. Journal of the American Statistical Association 2012; 107(497): 194–204.

[40]. Li KH. Imputation Using Markov Chains. Journal of Statistical Computation and Simulation 1988; 30:57–79.

[41]. Liu M, Wei L, Zhang J. Review of guidelines and literature for handling missing data in longitudinal clinical trials with a case study. Pharm Stat 2006; 5:7–18 [PubMed: 17080924]

[42]. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 2013; 8(2): 479–504.

[43]. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple Imputation by Chained Equations: What is it and how does it work?Int J Methods Psychiatr Res 2011; 20(1): 40–49 [PubMed: 21499542]

[44]. van der Linden S, Valkenburg H, Cats A. Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. Arthritis Rheum 1984; 27:361–368. [PubMed: 6231933]

[45]. Mackay K, Mack C, Brophy S, Calin A. The Bath Ankylosing Spondylitis Radiology Index (BASRI): a new, validated approach to disease assessment. Arthritis Rheum 1998; 41: 2263–70. [PubMed: 9870884]

[46]. Lee M, Rahbar MH, Brown MA, Gensler LS, Weisman MH, Diekman L, Reveille JD. Multiple Imputation method based on Weighted Quantile Regression Models for Longitudinal Censored Biomarker Data with Missing Early Visits BMC Medical Research Methodology 2018; 18(1):8. [PubMed: 29325529]

[47]. Meng X Multiple-imputation inferences with uncongenial sources of input. Statistical Science 1994; 9(4): 538–573.

[48]. Rubin DB. Multiple Imputation After 18+ years. Journal of the American Statistical Association 1996; 91(434): 473–489.

[49]. Lee M, Kong L, Weissfeld L. Multiple imputation for left-censored biomarker data based on Gibbs sampling method. Statistics in Medicine 2012; 31(17): 1838–1848. [PubMed: 22359320]

[50]. Benoit DF, Van den Poel D. bayesQR: A Bayesian approach to quantile regression. Journal of Statistical Software 2017;76.

[51]. Taddy MA, Kottas A. A Bayesian Nonparametric Approach to Inference for Quantile Regression. Journal of Business & Economic Statistics 2010;28:357–369.

[52]. Sriram K, Ramamoorthi RV, Ghosh P. Simultaneous Bayesian estimation of multiple quantiles with an extension to hierarchical models. IIM Bangalore Research Paper 2012.

[53]. Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 1995; 82(4):711–732.

[54]. Hjort N, Holmes C, Müller P, (Eds.), SW. Bayesian Nonparametrics: Principles and Practice. Cambridge University Press 2010.

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

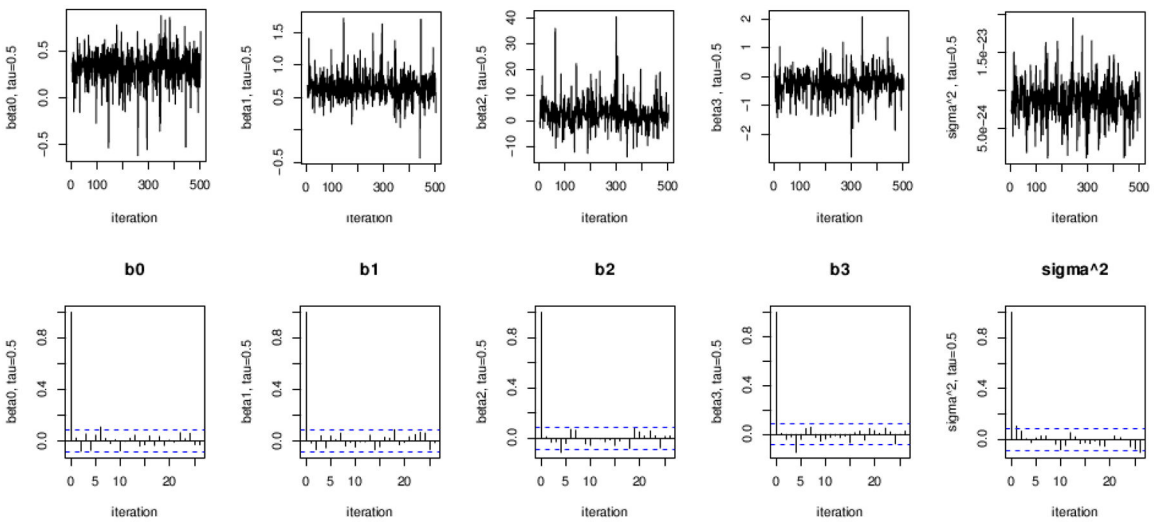

Figure 1:
Trace and autocorrelation function (ACF) plots

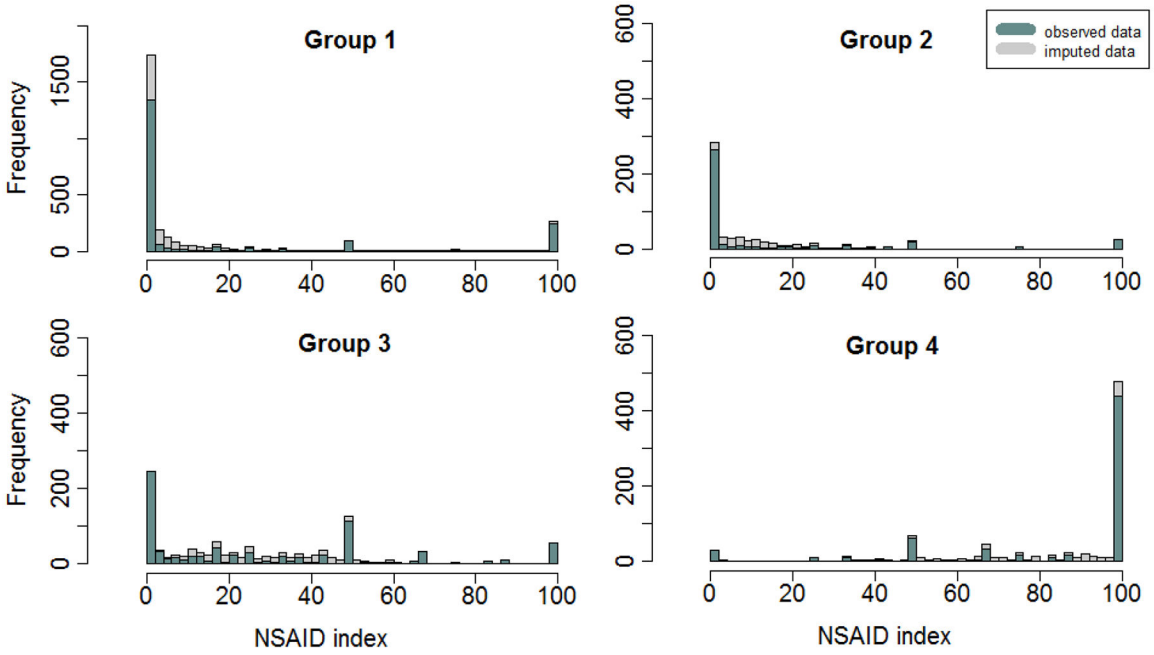

Figure 2:

Imputed NSAID index by latent class based on the proposed method MI-BQR_2 by Group Distribution of NSAID index data, distinguishing the observed data (green area) from imputed data (gray area) based on MI-BQR_2 method

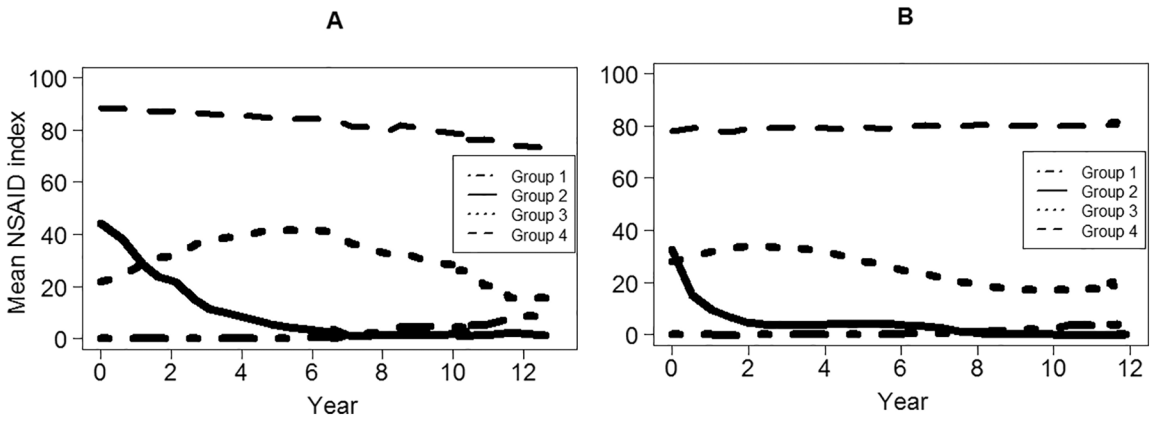

Figure 3:

Mean NSAID index trajectory over time

Mean NSAID index trajectory over time for progressor (A) and non-progressor (B): In Group 4, NSAIDs usage in progressors decreased, while it was not changing among non-progressors. For Group 3, NSAIDs usage was increasing faster during first 6 years among progressors, but this trend was not found for non-progressors. Progressors' NSAIDs intakes decreased much slower compared to non-progressors' in Group 2.

Table 1:

Simulation results (30% missing; Scenario 1- Scenario 3; n = 500)

a	a_0		a_1		a_2	
Method	**MSE**	**100xRE**	**MSE**	**100xRE**	**MSE**	**100xRE**
Scenario 1: Normal distribution						
OMNI	0.0036	–	0.0075	–	0.0009	–
OBSVD	0.0054	74.539	0.0097	83.316	0.0016	68.365
MI-MCMC	0.0052	90.475	0.0095	87.299	0.0014	86.744
MI-BQR$_1$	0.0040	91.677	0.0090	89.323	0.0011	92.626
MI-BQR$_2$	0.0032	93.554	0.0087	90.715	0.0009	98.468
Scenario 2: Asymmetric distribution						
OMNI	0.0092	–	0.0277	–	0.0036	–
OBSVD	0.0155	67.907	0.0554	81.904	0.0059	70.328
MI-MCMC	0.0115	80.065	0.0396	83.994	0.0047	77.108
MI-BQR$_1$	0.0102	90.515	0.0377	85.880	0.0041	92.907
MI-BQR$_2$	0.0098	92.735	0.0283	86.984	0.0032	95.345
Scenario 3: Heteroscedastic distribution						
OMNI	0.0096	–	0.0275	–	0.0009	–
OBSVD	0.0154	67.521	0.0597	81.271	0.0056	69.376
MI-MCMC	0.0112	81.992	0.0346	82.817	0.0051	77.733
MI-BQR$_1$	0.0109	90.588	0.0393	84.978	0.0041	93.333
MI-BQR$_2$	0.0099	92.579	0.0275	85.348	0.0038	94.349

OMNI (omniscient); **OBSVD** (using only observed data); **MI-MCMC** (MCMC-based MI method); **MI-BQR$_1$** (BQR-based MI method without latent class); **MI-BQR$_2$** (BQR-based MI method with latent class); **MSE** (mean squared error); **RE** (relative efficiency)

Table 2:

Simulation results (50% missing; Scenario 1- Scenario 3; n = 500)

a	a_0		a_1		a_2	
Method	**MSE**	**100xRE**	**MSE**	**100xRE**	**MSE**	**100xRE**
Scenario 1: Normal distribution						
OMNI	0.0036	–	0.0075	–	0.0009	–
OBSVD	0.0083	53.019	0.0138	62.263	0.0026	45.086
MI-MCMC	0.0070	73.574	0.0124	74.742	0.0020	83.294
MI-BQR$_1$	0.0044	82.640	0.0106	81.034	0.0011	91.746
MI-BQR$_2$	0.0035	86.540	0.0093	83.027	0.0019	93.372
Scenario 2: Asymmetric distribution						
OMNI	0.0092	–	0.0277	–	0.0036	–
OBSVD	0.0256	45.325	0.0691	60.427	0.0094	47.527
MI-MCMC	0.0139	74.633	0.0539	65.633	0.0060	81.375
MI-BQR$_1$	0.0115	82.139	0.0451	77.993	0.0043	89.583
MI-BQR$_2$	0.0092	84.345	0.0315	79.435	0.0034	92.536
Scenario 3: Heteroscedastic distribution						
OMNI	0.0096	–	0.0275	–	0.0039	–
OBSVD	0.0283	46.007	0.0639	60.344	0.0096	47.105
MI-MCMC	0.0137	74.851	0.0572	64.984	0.0064	81.478
MI-BQR$_1$	0.0142	81.855	0.0405	76.255	0.0041	89.716
MI-BQR$_2$	0.0114	83.011	0.0367	78.953	0.0029	91.105

OMNI (omniscient); **OBSVD** (using only observed data); **MI-MCMC** (MCMC-based MI method); **MI-BQR$_1$** (BQR-based MI method without latent class); **MI-BQR$_2$** (BQR-based MI method with latent class); **MSE** (mean squared error); **RE** (relative efficiency)

Table 3:

Simulation results (30, 50% missing; Scenario 1- Scenario 3; n = 200)

a	a_0		a_1		a_2	
Method	**MSE**	**100xRE**	**MSE**	**100xRE**	**MSE**	**100xRE**
Scenario 1: Normal distribution, 30% missing						
OMNI	0.0261	–	0.0780	–	0.0097	–
OBSVD	0.0412	58.780	0.1084	69.720	0.0153	59.198
MI-MCMC	0.0275	63.172	0.1261	80.224	0.0107	79.657
MI-BQR$_1$	0.0262	79.199	0.0785	88.897	0.0097	89.462
MI-BQR$_2$	0.0255	88.778	0.0781	90.222	0.0095	89.827
Scenario 2: Asymmetric distribution, 30% missing						
OMNI	0.0104	–	0.0200	–	0.0025	–
OBSVD	0.0451	65.970	0.0914	64.063	0.0114	69.050
MI-MCMC	0.0375	62.666	0.2132	48.873	0.0105	74.579
MI-BQR$_1$	0.0566	77.253	0.0691	83.481	0.0078	86.555
MI-BQR$_2$	0.0323	86.433	0.0687	85.742	0.0078	87.644
Scenario 3: Heteroscedastic distribution, 30% missing						
OMNI	0.0104	–	0.0208	–	0.0025	–
OBSVD	0.0443	65.083	0.0885	68.549	0.0123	57.503
MI-MCMC	0.0366	61.300	0.1997	59.563	0.0102	76.655
MI-BQR$_1$	0.0319	77.584	0.1131	83.029	0.0123	88.534
MI-BQR$_2$	0.0325	80.515	0.0675	85.880	0.0078	92.907
Scenario 1: Normal distribution, 50% missing						
OMNI	0.0260	–	0.0780	–	0.0097	–
OBSVD	0.0633	39.311	0.1394	52.920	0.0235	40.155
MI-MCMC	0.0309	63.288	0.2471	60.899	0.0130	60.187
MI-BQR$_1$	0.0273	75.043	0.0823	80.722	0.0099	88.252
MI-BQR$_2$	0.0274	85.879	0.0826	82.187	0.0099	91.720
Scenario 2: Asymmetric distribution, 50% missing						
OMNI	0.0104	–	0.0200	–	0.0025	–
OBSVD	0.0633	47.677	0.1155	51.386	0.0192	37.791
MI-MCMC	0.0477	57.188	0.5393	14.521	0.0168	59.096
MI-BQR$_1$	0.0347	70.075	0.0841	74.800	0.0087	80.609
MI-BQR$_2$	0.0344	80.302	0.0813	75.864	0.0087	81.029
Scenario 3: Heteroscedastic distribution, 50% missing						
OMNI	0.0104	–	0.0208	–	0.0025	–
OBSVD	0.0633	47.715	0.1160	51.411	0.0192	37.782
MI-MCMC	0.0475	57.830	0.5455	15.655	0.0168	60.471
MI-BQR$_1$	0.0346	70.759	0.1352	75.117	0.0129	82.645

a	a_0		a_1		a_2	
Method	**MSE**	**100xRE**	**MSE**	**100xRE**	**MSE**	**100xRE**
MI-BQR$_2$	0.0346	82.134	0.0829	77.993	0.0084	88.583

OMNI (omniscient); **OBSVD** (using only observed data); **MI-MCMC** (MCMC-based MI method); **MI-BQR$_1$** (BQR-based MI method without latent class); **MI-BQR$_2$** (BQR-based MI method with latent class); **MSE** (mean squared error); **RE** (relative efficiency)

Table 4:

Analysis results of longitudinal association between NSAIDs usage and mSASSS when NSAID index were imputed by different imputation methods

Method	**OBSVD**		**MI-MCMC**		**MI-BQR$_2$**	
Variable	**adj. RR (95% CI)**	**p-value**	**adj. RR (95% CI)**	**p-value**	**adj. RR (95% CI)**	**p-value**
NSAID index high vs. no use	**0.99** (0.91, 1.07)	0.759	**0.96** (0.89, 1.03)	0.245	**0.90** (0.82, 0.98)	0.017
NSAID index low vs. no use	**1.06** (0.99, 1.14)	0.073	**1.04** (0.98, 1.11)	0.229	**1.03** (0.91, 1.15)	0.646

OBSVD (using only observed data); **MI-MCMC** (MCMC-based MI method); **MI-BQR$_2$** (BQR-based MI method with latent class); **adj. RR**: adjusted rate ratio after controlling for sex, race, disease duration, co-morbidity, education, smoking status, C-reactive protein (CRP), Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) and medication usages of TNFi; **CI**: confidence interval.