Bayesian joint analysis for longitudinal studies with nonignorable missing data and applications to Parkinson's disease study

Jun Zhang, Sheng Luo*

Missing data are ubiquitous in longitudinal studies. Two types of missing data can happen: monotone missing data and intermittent (non-monotone) missing data, based on whether or not the patients will return to the study after the missed visit. In addition, missing data are ignorable in case of missing at random (MAR) or missing completely at random (MCAR), while missing data are not ignorable in case of missing not at random (MNAR). Statistical analysis can be biased if the missing data are not ignorable. A number of statistical approaches have been developed to handle missing data. Few studies provide inference to process both types of missing data simultaneously, especially, there is no study addressing the missingness in the data consist of dozens of ordinal responses. In this study, we propose a full latent trait model to analyze both types of missing data jointly. In addition, we provide a statistical method to test missing data mechanisms (MAR or MNAR) in our models. One contribution of this study is that the method is capable to handle the data consisting dozens of ordinal responses with existing impairment in multiple domains (e.g. motor, non-motor in Parkinson's Disease). Simulation study is conducted, and the results show our proposed model outperforms the naive model. This study is inspired by the Parkinson's Progression Markers Initiative (PPMI). We apply our approach to PPMI study and identify the specific disease domain that significantly associating with the MNAR mechanism in both intermittent pattern of missing data and monotone pattern of missing data.

Key Words: Item response, latent variables, informative dropouts, informative missing, domain-specific, Bayesian inference, multi-domain, multidimensional.

*Corresponding author: Sheng Luo is Associate Professor, Department of Biostatistics & Bioinformatics, Duke Clinical Research Institute, 2400 Pratt St. Durham, NC 27705, USA (E-mail: sheng.luo@duke.edu; Phone: 919-668-8038).

1 Introduction

Parkinson's disease (PD) is one of the common neurodegenerative disorders[1]. It is diagnosed in about 1% of individuals over the age of 60 worldwide[2]. PD is an incurable, complex and heterogeneous progressive disorder that gradually robs the individual of motor control. It is now considered as a systemic disease due to its non-motor symptoms in addition to the motor symptoms[3].

Because of no biomarker exists[4], the diagnosis method mainly depends on clinical information provided by patients, for example motor sign and symptoms (rigidity, tremor, etc.). The Unified Parkinson's Disease Rating Scale (UPDRS) is one widely used scale for clinical ratings of PD[5]. There are a lot of PD studies and clinical trials using UPDRS to follow the longitudinal course of PD[3,6,7]. In 2007, a revised UPDRS, the Movement Disorder Society-UPDRS (MDS-UPDRS) was introduced to provide more comprehensive and accurate tests than the original UPDRS[8]. The MDS-UPDRS consists of 65 items, all items are anchored with five categories, from 0 to 4, the higher the score the worse the status. The first three parts of MDS-UPDRS are commonly used in PD studies, which include 59 items. Statistical methods are required to extract useful information to define disease and its progression based on these 59 item responses.

The ordinal response in MDS-UPDRS are often summed up to obtain total score, which is treated as a continuous outcome. It is easy to implement but leads to loss of information by ignoring differences between item pattens[9]. Alternatively, multilevel item response theory (MLIRT) model was utilized to analyze the longitudinal scores. This model links the multiple items to an unobserved disease status structured as a univariate latent variable. However, the unidimensional framework limits the application of the model in analyzing PD due to the complication of disease such as the impairment and heterogeneity. PD is characterized by existing impairment across domains, for instance, non-motor symptoms often occur decade before the clinical motor signs[3]. To address these issues in the traditional models, multidimensional item response model was introduced[10,11]. Though cross-sectional impairment was addressed in this model, the longitudinal impairment information and correlations are not able to assessed in this cross-sectional multidimensional model. Recently, Wang and Luo[12] proposed a new multidimensional latent trait linear mixed model (MLTLMM) to address the disease impairment in longitudinal study. This new multidimensional latent trait model allows multiple latent variables and within-item multidimensionality. By adopting latent disease score to reduce the number of observed outcomes, MLTLMM is more computational scalable than multivariate marginal and random effects models.

In PD studies, participants are monitored longitudinally with respected to the aforementioned dozens ordinal outcomes plus other outcomes. During the follow-up, outcomes to be collected can be missing due to subjects' non-response, missed visits, dropout and etc. Rubin defined three missing data mechanisms[13]. If the missingness is independent of the observed and unobserved data, this missing data mechanism is missing completely at random (MCAR). When missingness is not dependent on unobserved data, it is missing at random (MAR). The missing data due to these two missing data mechanisms are treated as 'ignorable' missingness, which do not cause bias in statistical inference for likelihood-based estimation. In this study, we use MAR to denote the ignorable missingness, as MCAR is rare and can be handled in same way as MAR in analysis. However, when missingness is associated with the unobserved underlying response process, this missingness is missing not at random (MNAR). For example, patients' dropouts are due to worsening of disease or death. MNAR mechanisms are 'nonignorable'. Under the MNAR assumption, the missing data mechanism needs to be modeled simultaneously with the outcome variables to avoid biased parameter estimates[14].

In addition to these missing data mechanisms, two types of missing data are commonly observed. The first type is the 'intermittent missing data' or non-monotone missing data, for example, an individual may miss some visits before the last visit. While the other type is 'monotone missing data', denoting the data with the pattern that an individual leaves the study and never returns, or the observations are completely disrupted by some events (e.g. dropout or initiation of symptomatic treatment). Generally, the aforementioned two types missing data could have varied missing data mechanisms, which are difficult to justify whether the missing data mechanisms are MAR or MNAR. The possibility of missing data being MNAR can hardly be ruled out. Estimating parameters with nonignorable missing data is complex. Modeling longitudinal observations with nonignorable missing data has drawn much attention recently[15–17]. Assorted models are proposed, these models can be classified into three types: selection model, pattern-mixture model and shared-parameter models[16]. The selection approach combines the hypothetical complete data together with the missing data process based on likelihood. The pattern-mixture approach models the distribution of the data conditional on the missing data pattern. While the shared-parameter approach incorporates the dependence between measurements and missingness processes by means of random effects, it can be extended to latent variable models. Molenberghs *et al.* discussed a selection model for longitudinal ordinal data with nonrandom dropout[18]. Ekholm and Skinner proposed a pattern-mixture model for a longitudinal binary incomplete data set[19]. The full likelihood approach has been used to specify the joint likelihood of outcomes and missing indicators when handling nonmonotone pattern of missing data[20]. For example, the random-coefficient-based selection models were adopted to link dropout time to the longitudinal outcomes through individual random effects[21–23]. Alternatively, pattern-mixture model was used to analyze the joint distribution of repeated measures and monotone missing by stratifying the sample by time of dropout and then models the distribution of the repeated measures within each stratum[24]. Besides, pseudo likelihood was proposed to provide statistical solutions[25]. Elashoff *et al.*[26] developed the latent random effects model to incorporate effects from nonignorable monotone missing data. Most statistical models focus on one type of missing data pattern (either monotone or non-monotone missing). Those models are based on one or two outcomes and use the latent traits as predictors for monotone missing and other missing observations. Wu *et al.*[27] proposed a nonlinear mixed-effects model for both monotone and non-monotone patterns of missing data. However, when dealing with multivariate longitudinal data, especially when the study includes dozens of ordinal outcomes such as MDS-UPDRS in PD studies, these approaches are still not applicable due to large number of random effects and potential impairment across the multivariate covariates.

In this article, we present a generalized approach to analyze the longitudinal data with the presence of two types of missing data patterns with both missing data mechanisms. We extended multidimensional latent trait methods to model and test missing data mechanisms based on the responses from dozens of ordinal outcome. We jointly analyze the data without excluding MNAR assumption, and assess missing data mechanisms under the impact from heterogeneous disease development in multiple domains. This is the first study addressing the multiple ordinal responses (59 ordinal responses) with nonignorable missing data, carrying impairment information from multiple domains. The remainder of this article proceeds as follows. In section 2, we describe motivating study and the data. Section 3 discusses the proposed model, and Bayesian inference. Section 4 presents studies to assess the performance of the proposed models. In section 5, we apply our method to the motivating studies. Section 6 provides concluding remarks and discussion.

2 Motivating clinical studies

This methodological development is motivated by Parkinson's Progression Markers Initiative (PPMI) study. PPMI is an ongoing longitudinal observational study that aims to identify one or more markers of progression for Parkinson's disease (PD). All participants were grouped into several cohorts. At baseline, patients were not expected to require PD medications within at least 6 months. In PPMI study, MDS-UPDRS scale is used to assess the disease status and progression. MDS-UPDRS which has been proposed in 2007, is a new diagnosis tool for PD[8]. MDS-UPDRS scale consists of 65 items, each is anchored with five categories from 0 (lack of symptom) to 4 (severe symptom), categorized as measure for motor and non-motor symptoms. In clinical trial, the first three parts of MDS-UPDRS (59 items) are commonly used in study. According to the Movement Disorder Society (MDS), the 13 items in Part I are intended to measure the non-motor aspects of experience of daily living (nM-

EDL), the 13 items in Part II are intended to measure motor aspects of experiences of daily living (M-EDL), while the 18 grouped items (total 33 items, several with right, left or other body parts' sub-items) in Part III focus on motor examination. Items of the different subscales are assumed to be manifestation of different disease domains (motor, cognitive and behavior). To utilize domain-specific information carried in 59 items, we use item response latent trait model to capture the heterogeneity of disease[28]. MDS-UPDRS provides a relatively good measure to follow and define PD progression. Fitting MDS-UPDRS longitudinal observations into multidimensional latent trait model, we are able to characterize the natural disease progression of PD patients in PPMI study. However, each item of MDS-UPDRS carries pieces of disease information, in order to accurately define disease status, we need complete outcomes from each part. When a participant fails to answer one or more items out of total 59 items, those partial or complete non-responses will constitute missing data. The partial missed responses can prevent us to build the complete picture of PD status and progression, or our statistical approach depend on the full responses of these 59 items. Although PD studies are designed to collect complete data on all participants, missing data commonly happen and must be properly addressed in the analysis.

Besides, during the long course of follow up, the repeated observations are subject to the risks of endpoints, the follow-up of patients might be terminated long before the end of trial for different reasons, which can be treated as aforementioned monotone pattern of missing data. First, the longitudinal observations can be stopped because of dependent censoring (e.g., dropout, death), in addition, the symptomatic therapy (ST) can cause the repeated observations being interrupted (the following observations after ST will not reflect the natural PD progression). In PD studies, ST is generally being treated as endpoint[29]. These events are likely informative for disease progression and status. When modeling these endpoints, we are making statistical inference on whether these events associated with disease status or not. Generally, we are not able to exclude the possibilities that these events as disease-related or the monotone pattern of missing data belongs to MNAR.

Figure 1 uses sum score of MDS-UPDRS part III to illustrate the two types of missing data and how the longitudinal observations being interrupted. Indeed, we do not need an explicit model for the probabilities of missingness if missing data are MAR. However, we can not just conduct analysis based on MAR or MCAR assumption, and we need a statistical framework to test these hypothesis, and model the missing data simultaneously with outcome variables to avoid biased parameter estimates[14].

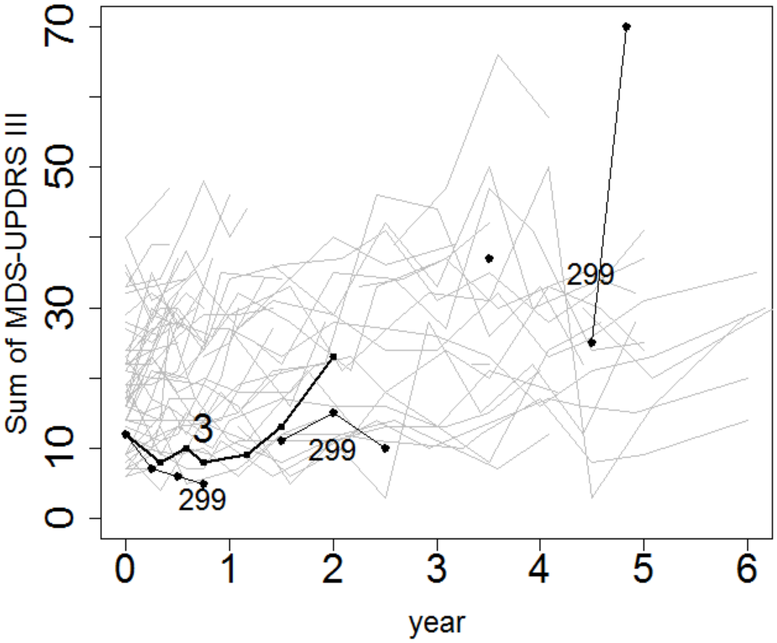

Figure 1: 50 randomly selected patients from PPMI study. Patient 299 had three missed visits before the last visit (intermittent pattern of missing data), patient 3 dropped out at end of year two (monotone pattern of missing data).

3 Model and estimation

3.1 Latent trait model

Let $y_{ik}(t)$ be the observed outcome k from ith subject at time t, where $i = 1, \ldots, N$, $k = 1, \ldots, K$, and $t = t_{i1}, \ldots, t_{iJ_i}$. All outcomes are coded so that larger values are worse clinical conditions.

To start building the MLTLMM modeling framework, we assume that there are P (with $P < K$) latent variables (LVs) representing the underlying disease severity scores and denote them as $\boldsymbol{\theta}_i(t) = (\theta_i^{(1)}(t), \ldots, \theta_i^{(p)}(t), \ldots, \theta_i^{(P)}(t))'$ for subject i at time t, where the superscript $(p = 1, \ldots, P)$ denotes the pth latent variable. From a clinical perspective, each latent variable denotes the severity of a PD domain (e.g., non-motor and motor). We introduce the linear model for continuous outcomes, logistic model for binary outcomes, and ordinal logistic model for ordinal responses.

$$y_{ik}(t) = a_k + \boldsymbol{b_k'\theta_i(t)} + \varepsilon_{ik}(t), \tag{1}$$

$$\text{logit}\big\{p(y_{ik}(t) = 1|\theta_i(t))\big\} = a_k + \boldsymbol{b_k'\theta_i(t)}, \tag{2}$$

$$\text{logit}\big\{p(y_{ik}(t) \leq l|\theta_i(t))\big\} = a_{kl} - \boldsymbol{b_k'\theta_i(t)}, \tag{3}$$

where a_k and $\boldsymbol{b_k}$ are the outcome-specific parameters, while the random errors $\varepsilon_{ik} \sim N(0, \sigma_{\epsilon_k})$, are independent and identically distributed. Note that for continuous outcome, $a_k = E[y_{ik}(t)|\boldsymbol{\theta}_i(t) = \mathbf{0}]$ is the mean of the kth outcome if the disease severity scores are 0. The parameter $\boldsymbol{b_k}$ also plays the role of bringing up disease severity score to the scale of the kth outcome. The negative sign for $\boldsymbol{b_k}$ in the ordinal outcome model is to ensure that worse disease severity (higher $\theta_i(t)$) is associated with more severe outcomes (higher $y_{ik}(t)$). The probability of being in a particular category is $p(y_{ik}(t) = l) = p(y_{ik}(t) \leq l|\boldsymbol{\theta_i(t)}) - p(y_{ik}(t) \leq l - 1|\boldsymbol{\theta_i(t)})$, where $l = 1, 2, \ldots, n_k - 1$ is the lth level of the kth random variable, which is ordinal with n_k levels. Interpretation of parameters is similar with continuous outcomes, except that modeling is on the log-odds, not the native scale of the data. Because the ordinal model is over-parameterized, additional constraints are required to make model identifiable. Using this model, we are able to explicitly combine information from all outcomes, specifically those dozens of ordinal outcomes. This is one of the simplest ways to conceptualize the disease severity scores that allows to define the disease status and progression when there is no gold standard.

$$\theta_i^{(p)}(t) = \boldsymbol{X}_i^{(p)}(t)\boldsymbol{\beta}^{(p)} + \boldsymbol{Z}_i^{(p)}(t)\boldsymbol{u}_i^{(p)} + e_i^{(p)}(t), \tag{4}$$

where $\boldsymbol{X}_i^{(p)}(t)$ and $\boldsymbol{Z}_i^{(p)}(t)$ are the covariates corresponding to fixed and random effects respectively, latent variable $\theta_i^{(p)}(t)$ denotes ith subject's unobserved disease severity in the pth domain at time t. The latent variables are continuous, higher value indicating worse severity of disease. The vector $\boldsymbol{u}_i = (\boldsymbol{u}_i^{(1)'}, \dots, \boldsymbol{u}_i^{(P)'})'$ contains the random effects for the ith subject, it follows a multidimensional normal distribution, $\boldsymbol{N(0, \Sigma)}$, where $\boldsymbol{\Sigma}$ is the covariance matrix with dimension equal to the number of random effects incorporated. There are several ways to model random effects. For example, when we incorporate fully correlated random intercepts and random slopes in framework (4), this covariance matrix will have the dimension of $2p \times 2p$. The residual term $e_i^{(p)}(t)$ is assumed to be mutually independent, and $e_i^{(p)}(t) \sim N(0, \sigma_e^{(p)})$.

For notational convenience, we let $\boldsymbol{a} = (\boldsymbol{a}_1', \dots, \boldsymbol{a}_k', \dots, \boldsymbol{a}_K')'$, and $\boldsymbol{a}_k = (a_{k,1}, \dots, a_{k,n_k-1})'$ for the kth ordinal outcome with n_k categories. We let $\boldsymbol{b} = \boldsymbol{(b_1, \dots, b_K)}'$, a K by P matrix, where $\boldsymbol{b_k} = (b_k^{(1)}, \dots, b_k^{(p)})'$. Because the model is over-parameterized, additional constraints are required to make it identifiable. The indeterminacy between the latent variable loadings $\boldsymbol{b}_k$ and the scales of the latent variables $\boldsymbol{\theta}_i(t)$ can be fixed by either setting one element in each column of $\boldsymbol{b}$ to be 1, or letting $\sigma_e^{(p)} = 1$ for $p = 1, \dots, P$ with at least one of the loadings constrained to be positive for each factor[30]. Finally, to identify parameters $\boldsymbol{a}$ and intercepts in regression coefficients, we set the constraints on one selected item in each domain, we let $a_{p,1} = 0$ (or other constant) for $p = 1, \dots, P$ ordinal outcomes and the order constraint $a_{k,1} < \dots < a_{k,l} < \dots < a_{k,n_{k-1}}$ must be satisfied. Besides, we set identifiability constraints on p orthogonal vectors $\boldsymbol{b}$, for example, when $P = 3$ and the constraints are put on the first three items, we let $b_1^{(1)} = b_2^{(2)} = b_3^{(3)} = 1$, all other elements are 0. In real data analysis, in order to achieve the better domain calibration and locate the three presumptive optimal bases, we have to carefully select the item to put constraints for each domain.

3.2 Model for monotone missing data

For the monotone missing data, we use proportional hazard model to incorporate this missing data pattern. Here we use y_{ikj} to denote the jth scheduled outcome k from subject i. When y_{ikj} belongs to monotone missing data, y_{ikq} is missing for all $q \geq j$, or it is the first missing observation for following consecutive missed visits. We use δ_i to code this missing data pattern, let $C_i = (T_i, \delta_i)$ be the endpoint observation for ith subject, where $\delta_i = 0$ indicating the latest visit and the following visits are still on going, $\delta_i = 1$ for event (either dropout or ST) and no following observation. To quantify the effect of $\theta_i(t_j)$ on the risks for events, we build the hazard model as,

$$\begin{aligned} \lambda_i(t) &= \lim_{h \to 0} \frac{P[t \leq T_i < t+h | T_i \geq t, \theta_i, \nu \boldsymbol{X_i(t)}, \gamma]}{h}, \\ &= \lambda_0(t) exp\{\boldsymbol{\gamma' W_i} + \boldsymbol{\nu' \theta_i(t)}\}, \end{aligned} \tag{5}$$

where λ_i is the hazard, and $\lambda_0(t)$ is the baseline hazard, $\boldsymbol{W_i}$ is a vector of fixed effect covariates, $\boldsymbol{\gamma}$ is the vector of regression coefficients. The regression coefficient $\boldsymbol{\nu'} = (\nu^{(1)}, \ldots, \nu^{(P)})$ links event times and the domain-specific disease status. In this model, $\lambda_i(t)$ is the instantaneous failure rate at time t given the covariates $\boldsymbol{W_i}$ and latent traits $\boldsymbol{\theta_i}$. More precisely, if $\boldsymbol{\nu = 0}$, this monotone missingness is not disease related, categorized as MAR. In case of MAR, this cox model is redundant, and not parsimonious in modeling.

3.3 Model for intermittent missing data

To incorporate intermittent missingness, we use a mixed effect logistic regression model to model the conditional probability of r_{ij}. Intermittent missing observations for y_{ikj} are those y_{ikj} may be missing while y_{ikq} is observed for some $q > j$. We use intermittent missing indicator $r_{ij} = 1$ to code this missing if he/she returns to study later. We now obtain the augmented data (y_{ij}, r_{ij}), where $r_{ij} = 1$ or 0, denoting whether ith subject's jth visit is missing or fully recorded respectively. For example, the outcome y_{ikj} ($j < J_i$, J_i is the last visit) is missing for $k = 1, \ldots, K$, when $r_{ij} = 1$, in this setting

$r_{iJ_i} = 0$ or y_{ikJ_i} (the last) must be observed.

$$logit(P(r_{ij} = 1|\theta_{ij})) = \boldsymbol{\alpha' X_i} + \boldsymbol{\eta' \theta_i(t_j)}, \tag{6}$$

where $\boldsymbol{X_i}$ is the vector of covariates and $\boldsymbol{\alpha}$ is the corresponding vector of regression coefficients. The parameter $\boldsymbol{\eta'} = (\eta^{(1)}, \dots, \eta^{(P)})$ governs the association between the intermittent missing data process and the domain-specific disease severity process modeled by latent variable. Both $y_i(t_j)$ and r_{ij} are censored by censoring time T_i. Moreover, the parameter $\boldsymbol{\eta}$ plays the role of sensitivity parameters to test MAR assumption for intermittent missing. When $\boldsymbol{\eta} = 0$ the missing data mechanism is MAR, otherwise it is MNAR, this is because the missingness (modeled missed visit probability) only depends on observed data. In case of MAR, this logistical function is redundant.

By combining equations 1 to 6, the joint models incorporate two missingness patterns into frameworks without excluding nonignorable mechanism assumption. We provide statistical test for MAR. If test results do not reject MAR for either (or both) missing data mechanisms, equation 5 or equation 6 (or both) will be independent of unobserved disease status, in this scenario the missing data are ignorable, equation 5 or equation 6 (or both) is redundant in models (not parsimonious).

3.4 Likelihood

For intermittent missing, the conditional likelihood function for the ith subject given parameters, covariates and random effects $\boldsymbol{X, U}$ is $L_i = f(\boldsymbol{y_i}|\boldsymbol{X, U, \Theta})f(\boldsymbol{r_i}|\boldsymbol{X, U, \Theta})$. The validity of likelihood is based on the assumption of independence of $\boldsymbol{y_i}$ and $\boldsymbol{r_i}$ given the latent variable (or random effect). The combined likelihood for ith subject by incorporating both monotone missing and non-monotone missing is,

$$L_i \ = f(\boldsymbol{y_i}|\boldsymbol{X, U, \Theta})f(\boldsymbol{r_i}|\boldsymbol{X, U, \Theta})f(T_i, \delta_i|\boldsymbol{X, U, \Theta})f(\boldsymbol{U}). \tag{7}$$

3.5 Bayesian inference

To make inference on the parameter vector $\boldsymbol{\Theta}$, we use Bayesian methods based on Markov chain Monte Carlo (MCMC) posterior simulations. We use vague priors on all elements in $\boldsymbol{\Theta}$, except for the aforementioned constrained parameters, i.e., $a_{1,1} = 0$ (or other constant) , and $b_1^{(p)} = 1$, for the selected item to ensure the item response model identifiable. Specifically, the prior distributions of unconstrained ordinal model's parameters a_k of the continuous outcomes is $a_k \sim N(0, 100)$. To obtain the prior distributions for the threshold parameters of ordinal outcome k, we let $a_{1,1} \sim N(0, 100)$, and $a_{k,l} = a_{k,l-1} + \Delta_l$ for $l = 2, \ldots, n_k - 1$, with $\Delta_l \sim N(0, 100)I(> 0)$, i.e., normal distribution left truncated at 0. Prior distributions for unconstrained elements in $\boldsymbol{b}$ and $\boldsymbol{\beta}$ are from $N(0, 50)$. We use the Cholesky factorization to estimate the correlation coefficients, the random effects covariance matrix is expressed as $\boldsymbol{\Sigma = \sigma'_u \Sigma_U \sigma_u}$, where Σ_U is the correlation matrix. All variances are from Inverse-Gamma$(0.01, 0.01)$. We have investigated other selections of vague prior distributions with various hyper-parameters and obtained very similar results.

The posterior samples are obtained from the full conditional of each unknown parameter using Hamiltonian Monte Carlo (HMC)[31] and No-U-Turn Sampler (NUTS)[32]. Both HMC and NUTS samplers are implemented in `Stan` (version 2.17.0)[33], which is a probabilistic programming language implementing statistical inference. For large datasets, `Stan` may be more efficient than `BUGS` language[34] in achieving faster convergence and requiring smaller number of samples[32]. To monitor Markov chain convergence, we use the trace plots and view the absence of apparent trends in the plot as evidence of convergence. In addition, we use the Gelman-Rubin diagnostic to ensure the scale reduction $\hat{R}$ of all parameters are smaller than 1.1 as well as a suite of convergence diagnosis criteria to ensure convergence[35].

4 Simulation

We conduct simulation studies to evaluate the proposed method and compare the method with the naive method (ignoring intermittent missing). We generate two continuous outcomes and a series of ordinal responses with 10 items, each item has five levels. To mimic the characteristics of PD study data, we let the outcome responses to be predicted by the latent traits from two domains. These twelve outcomes are longitudinal outcomes. Each response is predicted by two latent variables (cross loading). Here, we use multidimensional latent trait model[12] to model longitudinal observations. Now, the updated models are,

$$\theta_i^{(p)}(t) = \beta_0^{(p)} + \beta_1^{(p)} X_i + \beta_2^{(p)} t + U_i^{(p)} + e_i^{(p)}(t), \tag{8}$$

$$\lambda_i(t)|\boldsymbol{\theta_i(t)} = \lambda_0(t) exp(\gamma V_i + \boldsymbol{\nu'\theta_i(t)}, \tag{9}$$

$$logit(P(r_{ij} = 1|\boldsymbol{\theta_i(t)}) = w + \boldsymbol{\eta'\theta_i(t_j)}, \tag{10}$$

where $p = 1, 2$ denoting two disease domains, the vector $\boldsymbol{U_i'} = (u_i^{(1)}, u_i^{(2)}) \sim N(0, \boldsymbol{\Sigma})$. We simulate 1000 subjects. Each subject could have 16 sequential longitudinal observations in maximum. For monotone missing, we set the parameters for the fixed covariates as $\gamma = 1$, baseline hazards as $\lambda_0 = 0.006$, association parameter $\boldsymbol{\nu'} = c(0.4, 0.2)$. The censoring time is generated from exponential distribution with mean 50 in additional to the administrative censoring time 10. About 29% of 1000 subjects have dependent censoring. The latent variables are simulated with $(\beta_0^{(1)}, \beta_1^{(1)}, \beta_1^{(2)}, \beta_0^{(2)}) = (-0.2, 0.2, 0.5, -0.5)$, the time effects $(\beta_2^{(1)}, \beta_2^{(2)}) = (0.4, 0.7)$. The random effects are from $N_2(0, \boldsymbol{\Sigma})$, where diagonal elements of $\boldsymbol{\Sigma}$ (variance part) are $(1, 1.69)$, while the off-diagonal element (covariance) is 0.52, or equivalent to $(\sigma^{(1)}, \sigma^{(2)}, \rho) = (1, 1.69, 0.4)$. We simulate random errors from independent normal distribution $N(0, \epsilon^{(p)})$, while $(\epsilon^{(1)}, \epsilon^{(2)}) = (1, 0.64)$. For intermittent missing, we use $logit[P(r_{ij} = 1)] = W + \boldsymbol{\eta\theta_i(t_j)}$ to generate missing indicators. We let $(W, \eta^{(1)}, \eta^{(2)}) = (-4, 0.5, 0.7)$ to generate around 20% intermittent missing data (based on the summaries of 240 datasets). For comparison purpose, we generate another set of data with around

30% intermittent missing by letting $(W, \eta^{(1)}, \eta^{(2)}) = (-3, 0.5, 0.7)$. Taking into account the monotone missing, there are total about 30% missed data in the first setting, and about 40% in the second setting. The other parameter settings for continuous outcomes and ordinal outcomes are presented in Web Table 1. Total 240 datasets are generated.

In addition, we run naive model, treating missing as ignorable and only use the observed observations (intermittent missing data are ignored) for comparison purpose. The simulation analysis is conducted using a Bayesian approach via MCMC. Two chains are used in each setting, each has 4000 iterations with 3000 burn-in. For each estimated parameter, we compute percent bias as follows, for parameter β_j, percent bias= $100(\hat{\beta}_j - \beta_j)/\beta_j$. The biases is the average over all simulations. The simulation results are presented in the Table 1, more results are in Web Table 1.

Table 1: Simulation results with different intermittent missing proportions.

	Naive (setting 1)			**Joint (setting 1)**			**Naive (setting 2)**			**Joint. (setting 2)**		
	BIAS %	SD	CP	BIAS %	SD	CP	BIAS %	SD	CP	BIAS %	SD	CP
$\beta_0^{(1)} = -0.2$	13.000	0.045	0.900	0.000	0.044	0.949	31.000	0.044	0.691	0.000	0.046	0.930
$\beta_0^{(2)} = 0.2$	−9.000	0.051	0.927	2.500	0.050	0.927	−24.500	0.049	0.845	2.000	0.052	0.940
$\beta_1^{(1)} = 0.5$	0.200	0.034	0.959	0.800	0.037	0.967	0.400	0.037	0.973	−0.200	0.038	0.973
$\beta_1^{(2)} = -0.5$	−2.000	0.043	0.950	−0.600	0.044	0.953	−2.600	0.044	0.955	−0.600	0.045	0.958
$\beta_2^{(1)} = 0.4$	−4.250	0.008	0.568	−0.250	0.009	0.953	−3.500	0.009	0.677	−0.250	0.009	0.963
$\beta_2^{(2)} = 0.7$	−2.714	0.011	0.677	−0.143	0.013	0.963	−3.000	0.013	0.668	−0.143	0.014	0.963
$\rho = 0.4$	−7.000	0.034	0.873	−0.250	0.033	0.940	−9.000	0.036	0.823	−0.500	0.034	0.953
$\sigma_1^{(1)} = 1$	−5.600	0.065	0.827	−1.100	0.066	0.935	−6.500	0.065	0.782	−0.800	0.069	0.953
$\sigma_2^{(2)} = 1.69$	−4.970	0.100	0.818	0.059	0.100	0.953	−7.278	0.097	0.723	−0.118	0.104	0.940
$\epsilon^{(1)}= 1$	−2.100	0.058	0.886	−0.600	0.053	0.935	−2.400	0.055	0.905	−0.400	0.057	0.930
$\epsilon^{(2)}= 0.64$	2.500	0.033	0.927	0.000	0.034	0.935	−3.750	0.034	0.882	0.000	0.036	0.963
$\gamma = 1$	−0.800	0.171	0.959	−0.400	0.185	0.963	−0.600	0.185	0.973	−0.200	0.186	0.986
$\nu^{(1)} = 0.4$	6.250	0.061	0.900	4.000	0.056	0.926	6.500	0.057	0.918	4.000	0.057	0.926
$\nu^{(2)} = 0.2$	6.000	0.038	0.923	−0.500	0.035	0.958	6.500	0.036	0.932	0.000	0.036	0.935
$\lambda_0= 0.006$	0.000	0.001	0.945	0.000	0.001	0.949	0.000	0.001	0.935	0.000	0.001	0.935

Simulation results show that parameter estimations from joint model outperform the naive model in both setting. Generally, joint model had small bias (except one, but have better coverage probability), and coverage rate. Indeed, the simulation show that the proposed joint model can accurately estimate the covariate coefficients with the presence of both monotone missing and intermittent missing data.

5 Application to PPMI study

In this section, we use our proposed model to handle PPMI data which carry both monotone and nonmonotone missing data. The dataset used in this study were downloaded on Sept. 28, 2017. In PPMI study, all subjects were grouped into several cohorts, Parkinson Disease (PD), Scans Without Evidence of Dopaminergic Degeneration (SWEDD) and Healthy Control (HC) etc. We use the PD cohort, which includes 423 subjects. Excluding those having only one visit, there are 415 subjects in our study. Among these subjects, total 40 dropped out early for different reasons, and 197 individuals underwent ST. There are 3151 observations with complete recorded responses, and 151 visits with partial or complete missed records. We use the events of dropout and ST as the start of monotone missing data, and ignore the observation after ST (discussed in Section 2). There are total 237 individuals underwent events (dropouts or having ST) which are treated as having monotone missing data, and 158 missed visits which are recorded as the intermittent pattern of missing data. All the item responses in MDS-UPDRS part I, II and III are used as outcome responses. Based on guideline of MDS-UPDRS, the 13 items in part I are targeting for nM-EDL, part II's 13 items are for M-EDL, while part III's 27 items are for motor examination. The structured questionnaire design (part I, II, and III) defines the information manifested by those ordinal responses in each part to the corresponding domains. To fit to data structure, and incorporate the impairment of disease status across domains, we adept our models based on the assumption that item responses in three parts manifest the unobserved status of corresponding do- mains. We add dependency across domains by incorporating correlated random effects. Specifically, the models for the two missingness patterns are $\text{logit}\{p(r_{ij} = 1|\boldsymbol{\theta_i(t_j)})\} = W_k + \boldsymbol{\eta_k'\theta_i(t_j)}$, and $\lambda_i(t) = \lambda_0(t)exp(\gamma V_i + \boldsymbol{\nu'\theta_i(t)})$, where vector $\boldsymbol{\eta_k}' = (\eta_1, \eta_2, \eta_3)$, and vector $\boldsymbol{\nu'} = (\nu_1, \nu_2, \nu_3)$, corresponding to the coefficient to the domain-specific disease status in specific domain. This domain-specific setting enable us to incorporate the impairment across domains and regress the effects of heterogeneity of disease development among domains.

Table 2: Parameter estimates for PPMI's three domains.

	nM-EDL				M-EDL				Motor Examination			
	MEAN	SD	95 % CI		MEAN	SD	95% CI		MEAN	SD	95% CI	
Disease Status (Latent Variable)												
Int.	0.011	0.079	−0.149	0.152	−0.045	0.089	−0.213	0.124	0.046	0.189	−0.333	0.392
Age (yr)	0.093	0.058	−0.024	0.200	0.113	0.070	−0.021	0.245	0.254	0.148	−0.031	0.531
Time (yr)	0.289	0.023	0.246	0.333	0.416	0.033	0.355	0.479	0.656	0.044	0.571	0.745
Random effects												
ρ	−0.179	0.077	-0.324	−0.027	−0.120	0.074	−0.259	0.028	−0.387	0.067	−0.516	-0.258
σ_i *	1.160	0.065	1.037	1.286	1.484	0.083	1.327	1.655	3.506	0.177	3.172	3.860
σ_s **	0.246	0.021	0.206	0.287	0.360	0.031	0.302	0.422	0.576	0.042	0.500	0.662
σ_e	0.264	0.025	0.215	0.309	0.532	0.030	0.477	0.596	1.280	0.051	1.179	1.393
Intermittent pattern of missing data												
η	0.237	0.107	0.030	0.452	-0.056	0.093	-0.234	0.125	-0.013	0.030	-0.074	0.046
Monotone pattern of missing data												
ν	0.105	0.086	-0.061	0.273	0.183	0.072	0.044	0.326	0.009	0.025	-0.038	0.058

*: random intercept

**: random slope

Table 2 shows the domain-specific parameter estimates and the 95% confidence intervals. We find not significant age effects in all domains. Several studies addressed the role of age in PD severity with mixed results[36]. In this study, our analysis is based on a small cohort of early stage PD patients, and the findings suggest age is not an important factor in PD progression. The time effects are significant in all three domains. Measured by the structured disease statues, the disease is getting worse at average rate of 0.289 units per year (CI: [0.246, 0.333]) in nM-EDL domain, the disease progresses to worse at average rate of 0.416 units per year (CI: [0.355, 0.479]) in M-EDL domain, while in Motor Examination domain, the disease is getting worse at average rate of 0.656 units per year (CI: [0.571, 0.745]). The intermittent missingness pattern is significant associated with disease status and development in nM-EDL with log odd increased 0.237 (CI=[0.03, 0.452]) for ever unit worsening of disease in nM-EDL while controlling other two disease status unchanged. The occurrence of monotone missingness is significantly associated with the disease status and progression in M-EDL, the log hazard ratio increases 0.183 units (CI=[0.044, 0.326]) for every unit worsening of disease in M-EDL while controlling other two disease status unchanged. Overall, we can not exclude the MNAR for both types of missing data. This new finding signifies the unique advantage of our model.

Web Table 2 shows all significant random intercepts and random slopes across domains. The correlation matrix (Web Table 3) provides additional information of inter-dependency of diseases and correlated (in high dimension) subject-specific characteristics between and within domains. Every individual has 3 pairs of random effect terms sampled from a 6×6 covariance matrix, denoting the random effects in 3 dimensions or domains, each domain has its domain-specific random intercept and random slope. For each random effects vector, the 1st and 2nd elements are for nM-EDL domain, the 3rd and 4th are for domain in M-EDL, while the last two are for motor examination domain. The significant within-domain correlations (between random intercept and random slope) resides in nM-EDL ($\rho_{01} = -0.179,$ 95% CI: $[-0.324, -0.027]$) and motor examination ($\rho_{01} = -0.387,$ 95% CI: $[-0.516, -0.258]$), explained as: if a participant's nM-EDL disease status is worse at the start of the trial, his or her nM-EDL's disease progression is slow during follow-up. Same interpretation is for disease in Motor Examination domain. The intercepts between-domain correlations are all positive and significant (random intercept vs random intercept), these associations reveal that in the study if any initial disease status is worse, the other two are also worse, for example, the initial disease status in nM-EDL is worse, the disease severities in M-EDL and motor examination are also worse. Same finding is discovered in the correlations of within-domain random slope (random slope vs random slope), if disease progresses fast in any domain, it also develops fast in other domains, for example, disease develops fast in M-EDL, it also deteriorates fast in nM-EDL and motor examination. Overall, in this study, we incorporate the cross domain interaction, Our high dimensional random effect matrix (6X6) reveals the internal, hidden, between and within domain disease correlation.

6 Discussion

This article provides a joint model approach to analyze longitudinal outcomes in the presence of nonignorable missing data. We use a logistic model to describe the intermittent missingness pattern, while modeling the monotone missingness pattern using Cox model. Our model provides a statistical test for missing data mechanism, we simplify this process to covariate effect hypothesis test. Moreover, in analyzing multiple mixed types data, our model incorporates domain-specific variability for disease. To incorporate the impairment among domains, we refine our model to obtain the heterogeneity of disease from different domains. We quantify the impact on missing data mechanisms from domain-specific disease status, and provide a direct interpretation for the impact. This is the strength of our proposed model. In addition, our model can be extended to incorporate the high order impact, such as those from time-dependent disease progression. This approach can also be extended to competing risk model for multiple cause of endpoints.

We apply our approach to PPMI study, our model discloses that the intermittent missingness pattern is significantly associated with the disease development in nM-EDL, while the monotone missing pattern are mainly affected by the disease status in M-EDL. Specifically, both missingness patterns can not exclude MNAR, however, these two missingness patterns are not related to the disease progression in motor examination. One possible reason may be from the design of study, and cohort used in this study, as the goal of PPMI is to identify, test and verify markers of progression for early-stage Parkinson's disease, at which stage, motor impairment is less severe compared with non-motor impairment.

Our model has some limitations. Our joint analysis is based on latent traits to provide inference for nonignorable missingness with fully correlated random effects cross domain. We do not provide a method to address the partial missing data. We do not include continuous responses in this study, the reason is that the continuous responses are more frequently missed than MDS-UPDRS in PPMI datasets. To incorporate these continuous variables into real analysis, we need additional logistic models (to incorporate missing heterogeneity, such as missing continuous outcomes while ordinal outcomes available) which makes modeling complicated. To assess the underling association when patients intentionally avoid certain medical tests or answering some questions, we need build a more general approach to handle the partial missing data, and test the missing data mechanism. In this study we provide a statistical test to test missing data mechanism, there are other models with different frameworks, we do not provide comparison of test results to other applicable models, such as selection model and mixture model. In general, local sensitivity analysis can be use to evaluate the robustness of inference of departure from assumption[37–39].

References

[1] Pringsheim T, Jette N, Frolkis A, Steeves TD. The prevalence of Parkinson's disease: A systematic review and meta-analysis. Movement Disorders. 2014;29(13):1583–1590.

[2] Bellou V, Belbasis L, Tzoulaki I, Evangelou E, Ioannidis JP. Environmental risk factors and Parkinson's disease: an umbrella review of meta-analyses. Parkinsonism & Related Disorders. 2016;23:1–9.

[3] Chaudhuri KR, Healy DG, Schapira AH. Non-motor symptoms of Parkinson's disease: diagnosis and management. The Lancet Neurology. 2006;5(3):235–245.

[4] Michell A, Lewis S, Foltynie T, Barker R. Biomarkers and Parkinson's disease. Brain. 2004;127(8):1693–1705.

[5] Ramaker C, Marinus J, Stiggelbout AM, Van Hilten BJ. Systematic evaluation of rating scales for impairment and disability in Parkinson's disease. Movement Disorders. 2002;17(5):867–876.

[6] Tison F, Yekhlef F, Chrysostome V, Balestre E, Quinn NP, Poewe W, et al. Parkinsonism in multiple system atrophy: Natural history, severity (UPDRS-III), and disability assessment compared with Parkinson's disease. Movement Disorders. 2002;17(4):701–709.

[7] Taylor Tavares AL, Jefferis GS, Koop M, Hill BC, Hastie T, Heit G, et al. Quantitative measurements of alternating finger tapping in Parkinson's disease correlate with UPDRS motor disability

and reveal the improvement in fine motor control from medication and deep brain stimulation. Movement Disorders. 2005;20(10):1286–1298.

[8] Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. Movement Disorders. 2008;23(15):2129–2170.

[9] Gorter R, Fox JP, Twisk JW. Why item response theory should be used for longitudinal questionnaire data analysis in medical research. BMC Medical Research Methodology. 2015;15(1):15–55.

[10] Gnaldi M. A multidimensional IRT approach for dimensionality assessment of standardised students' tests in mathematics. Quality & Quantity. 2017;51(3):1167–1182.

[11] Osteen P. An introduction to using multidimensional item response theory to assess latent factor structures. Journal of the Society for Social Work and Research. 2010;1(2):66–82.

[12] Wang J, Luo S. Multidimensional latent trait linear mixed model: an application in clinical studies with multivariate longitudinal outcomes. Statistics in Medicine. 2017;36(20):3244–3256.

[13] Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–592.

[14] Diggle P. Analysis of Longitudinal Data. Oxford University Press; 2002.

[15] Little RJ. Pattern-mixture models for multivariate incomplete data. Journal of the American Statistical Association. 1993 3;88(421):125–134.

[16] Little RJ. Modeling the drop-out mechanism in repeated-measures studies. Journal of the American Statistical Association. 1995;90(431):1112–1121.

[17] Yuan Y, Yin G. Bayesian quantile regression for longitudinal studies with nonignorable missing data. Biometrics. 2010;66(1):105–114.

[18] Molenberghs G, Kenward MG, Lesaffre E. The analysis of longitudinal ordinal data with non-random drop-out. Biometrika. 1997;84(1):33–44.

[19] Ekholm A, Skinner C. The Muscatine children's obesity data reanalysed using pattern mixture models. Journal of the Royal Statistical Society: Series C (Applied Statistics). 1998;47(2):251–263.

[20] Ibrahim JG, Chen MH, Lipsitz SR. Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. Biometrika. 2001;88(2):551–564.

[21] Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. Biometrics. 1988;44(1):175–188.

[22] De Gruttola V, Tu XM. Modelling progression of CD4-lymphocyte count and its relationship to survival time. Biometrics. 1994;50(4):1003–1014.

[23] Ribaudo HJ, Thompson SG, Allen-Mersh TG. A joint analysis of quality of life and survival using a random effect selection model. Statistics in Medicine. 2000;19(23):3237–3250.

[24] Hogan JW, Laird NM. Model-based approaches to analysing incomplete longitudinal and failure time data. Statistics in Medicine. 1997;16(3):259–272.

[25] Troxel AB, Lipsitz SR, Harrington DP. Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data. Biometrika. 1998;85(3):661–672.

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

[27] Wu L, Hu XJ, Wu H. Joint inference for nonlinear mixed-effects models and time to event at the presence of missing data. Biostatistics. 2007;9(2):308–320.

[28] Baker FB. The Basics of Item Response Theory. ERIC; 2001.

[29] Simuni T, Long JD, Caspell-Garcia C, Coffey CS, Lasch S, Tanner CM, et al. Predictors of time to initiation of symptomatic therapy in early Parkinson's disease. Annals of Clinical and Translational Neurology. 2016;3(7):482–494.

[30] Dunson D. Dynamic latent trait models for multidimensional longitudinal data. Journal of the American Statistical Association. 2003;98(463):555–563.

[31] Duane S, Kennedy AD, Pendleton BJ, Roweth D. Hybrid monte carlo. Physics Letters B. 1987;195(2):216–222.

[32] Hoffman MD, Gelman A. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research. 2014;15(1):1593–1623.

[33] Stan Development Team. Stan Modeling Language Users Guide and Reference Manual, Version 2.17.0; 201. Available from: `http://mc-stan.org/`.

[34] Lunn D, Thomas A, Best N, Spiegelhalter D. WinBUGS - A Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing. 2000;10(4):325–337.

[35] Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis. Chapman & Hall/CRC Boca Raton, FL, USA; 2014.

[36] Levy G. The relationship of Parkinson disease with aging. Archives of Neurology. 2007;64(9):1242–1246.

[37] Verbeke G, Molenberghs G, Thijs H, Lesaffre E, Kenward MG. Sensitivity analysis for nonrandom dropout: a local influence approach. Biometrics. 2001;57(1):7–14.

[38] Ma G, Troxel AB, Heitjan DF. An index of local sensitivity to nonignorable drop-out in longitudinal modelling. Statistics in Medicine. 2005;24(14):2129–2150.

[39] Moreno-Betancur M, Chavance M. Sensitivity analysis of incomplete longitudinal data departing from the missing at random assumption: Methodology and application in a clinical trial with drop-outs. Statistical Methods in Medical Research. 2016;25(4):1471–1489.