Statistica Sinica **27** (2017), 187-205
doi:http://dx.doi.org/10.5705/ss.202015.0120

PERSONALIZE TREATMENT FOR LONGITUDINAL DATA USING UNSPECIFIED RANDOM-EFFECTS MODEL

Hyunkeun Cho, Peng Wang and Annie Qu

Western Michigan University, University of Cincinnati and University of Illinois at Urbana-Champaign

Abstract: We develop new modeling for personalized treatment for longitudinal studies involving high heterogeneity of treatment effects. Incorporating subject-specific information into the treatment assignment is crucial since different individuals can react to the same treatment very differently. We estimate unobserved subject-specific treatment effects through conditional random-effects modeling, and apply the random forest algorithm to allocate effective treatments for individuals. The advantage of our approach is that random-effects estimation does not rely on the normality assumption. In theory, we show that the proposed random-effect estimator is consistent and more efficient than the random-effect estimator that ignores correlation information from longitudinal data. Simulation studies and a data example from an HIV clinical trial also confirm that the proposed method can efficiently identify the best treatments for individual patients.

Key words and phrases: Generalized linear mixed model, penalized quasi-likelihood, personalized treatment, quadratic inference functions, random forest.

1. Introduction

In recent years there has been an increasing demand for effective personalized treatments for individuals since their responses to the same treatment can be very different. For example, in a nutrition study conducted at the University of Illinois at Urbana-Champaign for patients with stage-five chronic kidney disease, personal information such as nutrition levels from blood tests are collected repeatedly to assess treatment effects at different protein intake levels. Since the patients' responses to the assigned treatment have high variability, this imposes a challenge for the accurate estimation of treatment effects without incorporating heterogeneity variability among individuals. Traditional approaches applying one treatment for all patients are no longer seen as effective in achieving the best treatment outcomes for different individuals. In this paper, we propose an efficient treatment assignment strategy to categorize patients who might benefit from certain treatments more than others.

Recent developments for personalized treatment include Bonetti and Gelber (2000, 2004) approaches on discovering patterns of treatment effects for overlapping subgroups of patients based on moving average techniques; Song and Pepe's (2004) subgroup identifications to select ideal treatments for patients assuming a monotone relationship between the covariate and the binary response; and Qian and Murphy's (2011) optimal treatment scheme derived from conditional mean estimation using the penalized least square. In addition, scoring systems under parametric or semiparametric regression models have been proposed to specify a desired level of treatment difference (Marlowe et al. (2007); Cai et al. (2011); Zhao et al. (2013)). These approaches require that the optimal treatment for subjects at each risk level be specified beforehand.

Nonparametric classification methods are also popular for categorizing subjects into different groups through a multiple testing approach (Su et al. (2009); Lipkovich et al. (2011)). Foster, Taylor and Ruberg (2011) introduce the concept of "virtual twins" to assess treatment differences based on the prediction of the interaction effects for treatment and covariate. Zhao et al. (2012) utilized a weighted-outcome support vector machine to maximize the conditional mean of the responses. However, these methods are mainly applicable for independent observations, and do not consider heterogeneity variation among subjects. In practice, there are often unobserved latent factors which contribute to individual treatment effects, in addition to the observed covariate information.

To address subject-specific variation among patients, Diaz et al. (2007), Diaz, Yeh, and Leon (2012) applied linear mixed models to estimate the frequency of disease occurrence corresponding to different drug dosages. These were the first attempts to provide random-effects model estimation and interpretation in the personalized medicine literature. However, their approach does not consider important latent factors induced from individual information such as clinical, genetic, environmental, and demographic variables; these are essential to formulate personalized treatment effectively. In addition, their model is mainly applicable for the normal random-effects distribution.

In many clinical trials, a treatment which is beneficial for some patients might be ineffective or have an adverse effect on others and the random effects for modeling heterogeneity of the treatment effects need not be normal. Standard mixed-effects models assuming normality of random effects (Laird and Ware (1982); Breslow and Clayton (1993); McCulloch (1997); Jiang and Zhang (2001); Vonesh et al. (2002); McCulloch, Searle and Neuhaus (2008)) may not be effective at capturing subject-specific treatment effects.

In this paper, we provide a new personalized treatment assignment strategy that is applicable for continuous, discrete, and categorical longitudinal responses. We estimate the random effect through generalized linear mixed modeling, and

utilize supervised learning algorithms by treating the random-effect estimates as responses. The proposed strategy provides a personalized treatment rule that specifies a treatment according to an individual's characteristics. This is achieved by taking both observed covariates and unobserved random effects associated with individuals into consideration. This leads to a more effective treatment assignment to new patients. Among several supervised learning algorithms, we adopt the random forest algorithm due to its accurate prediction and interpretation for tree-based models, see Breiman (2001, 2004), Biau, Devroye and Lugosi (2008), Biau (2012), and Denil, Matheson and de Freitas (2014) for more details. In addition, the random forest algorithm is able to identify important factors that influence the outcome of the treatment. Simulations also show that the random forest algorithm provides a better decision rule than the other supervised learning algorithms in the sense of reducing prediction errors.

Estimating random effects accurately is a key step for optimal treatment assignments for patients. We propose estimating random effects without imposing any distributional assumption on them. This is essential to obtaining accurate subject-specific treatment effects in the mixed-effects model. An advantage of the proposed estimation procedure is that it does not require specification of the likelihood function, yet still accommodates serial correlation over time in estimating both fixed and random effects. In theory, we show that the random-effects estimator of the proposed approach is consistent and more efficient than an estimator which ignores correlation information from longitudinal data.

The paper is organized as follows. Section 2 describes the existing estimation procedures for the generalized linear mixed model. Section 3 introduces a personalized treatment assignment strategy and provides asymptotic properties of the estimators and the implementation of the proposed method. Sections 4 and 5 provide simulation studies and data analysis for a human immunodeficiency virus (HIV) clinical trial study. The final section gives concluding remarks. Proofs and necessary conditions are provided in the Appendix.

2. Notation and Framework

For longitudinal data, let y_{it} be a response variable and $\mathbf{x}_{it}$ be an $1 \times p$ vector of covariates, measured at time $t = t_1, \ldots, t_{n_i}$ for subjects $i = 1, \ldots, N$. To simplify the notation, we set $n_i = n$ for all subjects; the implementation for unbalanced data will be introduced in Section 3. We assume that the model satisfies $\mu_{it} = E\{y_{it}\} = \mu\{\mathbf{x}_{it}\boldsymbol{\beta}\}$, where $\mu(\cdot)$ is a known inverse link function and $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector; this is required to be correctly specified for the marginal model approach. It is suitable when the inference of the population average is of main interest. However, if there is strong indication of individual variations, it is

more sensible to apply a random-effects model. For the generalized linear mixed model, the conditional mean of the response given the random effects $\mathbf{b}_i$ is

$$E\{y_{it}|\mathbf{b}_i\} = \mu\{\mathbf{z}_{it}\mathbf{b}_i + \mathbf{x}_{it}\boldsymbol{\beta}\} = \mu_{it}, \quad i = 1, \ldots, N, \tag{2.1}$$

where $\mathbf{b}_i$ is a $q \times 1$ vector of random effects corresponding to covariates $\mathbf{z}_{it}$ for the ith subject at time t.

If the conditional likelihood of y_{it} given $\mathbf{b}_i$ is unknown, we can solve the quasi-likelihood equation (Wedderburn (1974)) to obtain fixed and random effects estimators. Specifically, the estimating equations corresponding to the fixed-effects parameters $\boldsymbol{\beta}$ and random-effects $\mathbf{b}_i$ are

$$\sum_{i=1}^{N} \dot{\boldsymbol{\mu}}'_{i,\beta}\mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0, \tag{2.2}$$

$$\dot{\boldsymbol{\mu}}'_{i,\mathbf{b}_i}\mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0, \ i = 1, \ldots, N, \tag{2.3}$$

where $\dot{\boldsymbol{\mu}}_{i,\beta} = \frac{\partial}{\partial\beta}\boldsymbol{\mu}_i(\boldsymbol{\beta}|\mathbf{b}_i)$, $\dot{\boldsymbol{\mu}}_{i,\mathbf{b}_i} = \frac{\partial}{\partial\mathbf{b}_i}\boldsymbol{\mu}_i(\boldsymbol{\beta}|\mathbf{b}_i)$, $\mathbf{V}_i = \mathrm{Var}\,(\mathbf{y}_i|\mathbf{b}_i)$, $\mathbf{y}_i = (y_{it_1}, \ldots, y_{it_n})'$, and $\boldsymbol{\mu}_i = (\mu_{it_1}, \ldots, \mu_{it_n})'$. If the dimension of random effects $\mathbf{b} = (\mathbf{b}'_1, \ldots, \mathbf{b}'_N)'$ in (2.3) increases as the sample size increases, then the estimation of random effects could be non-convergent and inconsistent without additional constraints or distributional assumptions.

Breslow and Clayton (1993) proposed penalized quasi-likelihood; this requires the normality of random effects to ensure the consistency of random-effects estimation. Neuhaus, Hauck and Kalbfleisch (1992) and Wang, Tsai and Qu (2012) indicate that, when the distribution of the random effects is not normal, the estimators of the fixed effects could be biased. For personalized treatment identification, the normality of random effects is too restrictive for effectively distinguishing individual treatments. Jiang (1999) estimates random effects without any distributional assumption, but assumes that the observations are independent conditional on the random effects. It is important to incorporate serial correlations over time for longitudinal studies. However, the true covariance matrix $\mathbf{V}_i^{-1} = \mathbf{A}_i^{-1/2}\mathbf{R}^{-1}\mathbf{A}_i^{-1/2}$ in (2.2) and (2.3) is often unknown, where $\mathbf{R}$ is the true correlation matrix and $\mathbf{A}_i = \mathrm{diag}\{\mathrm{var}(y_{i1}|\mathbf{b}_i), \ldots, \mathrm{var}(y_{in}|\mathbf{b}_i)\}$. Qu, Lindsay and Li (2000) approximate $\mathbf{R}^{-1}$ by $\sum_{j=1}^{m} a_j\mathbf{M}_j$, where $\mathbf{M}_1, \ldots, \mathbf{M}_m$ are known basis matrices, and the a_j's are unknown constants. More details on choosing the basis matrices of $\mathbf{M}_j$ $(j = 1, \ldots, m)$ can be found in Zhou and Qu (2012).

Wang, Tsai and Qu (2012) define the fixed-effects extended score vector to represent the quasi-likelihood equations in (2.2), conditional on the random effects $\mathbf{b}$, as

$$\mathbf{G}_N^f = \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i^f(\boldsymbol{\beta}) = \frac{1}{N}\begin{pmatrix} \sum_{i=1}^{N}\dot{\boldsymbol{\mu}}'_{i,\beta}\mathbf{A}_i^{-1/2}\mathbf{M}_1\mathbf{A}_i^{-1/2}(\mathbf{y}_i - \boldsymbol{\mu}_i) \\ \vdots \\ \sum_{i=1}^{N}\dot{\boldsymbol{\mu}}'_{i,\beta}\mathbf{A}_i^{-1/2}\mathbf{M}_m\mathbf{A}_i^{-1/2}(\mathbf{y}_i - \boldsymbol{\mu}_i) \end{pmatrix}, \tag{2.4}$$

and the random-effects extended score vector conditional on the fixed effects $\boldsymbol{\beta}$ as $\mathbf{G}^r = \{(\mathbf{g}_1^I)', \ldots, (\mathbf{g}_N^I)', \lambda \mathbf{b}', \lambda(\mathbf{P_J b})'\}'$. Here $\mathbf{g}_i^I = \dot{\boldsymbol{\mu}}_{i,\mathbf{b}_i}' \mathbf{A}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i)$, the tuning parameter λ is chosen to be $\log(N)$, $\mathbf{P_J} = \mathbf{J}(\mathbf{J}'\mathbf{J})^{-1}\mathbf{J}'$ is a known projection matrix on the null space of $(\mathbf{I} - \mathbf{P_X})\mathbf{Z}$, $\mathbf{J}$ is a matrix whose columns constitute bases for the null space of $(\mathbf{I} - \mathbf{P_X})\mathbf{Z}$, and $\mathbf{P_X}$ is defined similarly as $\mathbf{P_J}$ with covariates $\mathbf{X}$ and $\mathbf{Z}$ associated with fixed effects and random effects, respectively. Two quadratic inference functions are minimized iteratively to obtain fixed and random effects estimators:

$$L^f(\boldsymbol{\beta}) = (\mathbf{G}_N^f)'(\mathbf{W}_N^f)^{-1}(\mathbf{G}_N^f), \tag{2.5}$$

$$L^r(\mathbf{b}) = (\mathbf{G}^r)'(\mathbf{G}^r), \tag{2.6}$$

where $\mathbf{W}_N^f = (1/N)\sum_{i=1}^N (\mathbf{g}_i^f)(\mathbf{g}_i^f)'$. For estimation of the parameter $\boldsymbol{\beta}$, $\mathbf{G}_N^f$ does not involve any nuisance parameter to be estimated yet still accommodates correlations within subjects holding $E\{\mathbf{g}_i^f(\boldsymbol{\beta})\} = 0$ under the true parameter. Still, Wang, Tsai and Qu (2012) do not use the correlation information from longitudinal observations in formulating $\mathbf{g}_i^I$.

3. Methodology

In this section, we introduce a personalized treatment strategy to assign the best treatment for individuals that uses a generalized linear mixed model and the nonparametric regression approach to longitudinal analysis.

3.1. Two-step procedure for personalized treatment

Random-effects modeling is useful for providing an interpretation of heterogeneous variation among subjects, and for estimating the personalized treatment effects in personalized medicine. We divide our covariates $\mathbf{x}_{it}$ as $\mathbf{x}_{it} = (\mathbf{x}_{it}^0, \mathbf{x}_{it}^1)$, where $\mathbf{x}_{it}^0$ are the covariates which are the same for every subject, and $\mathbf{x}_{it}^1$ are the baseline covariates providing the individual characteristics that might be relevant to treatment assignment. Suppose there are k treatments. We consider the generalized linear mixed model:

$$E\{y_{it}|\mathbf{b}_i, b_{0i}\} = \mu\{\mathbf{z}_i(\boldsymbol{\beta} + \mathbf{b}_i) + \mathbf{x}_{it}^0\boldsymbol{\beta}_0 + b_{0i}\} = \mu_{it}, \quad i = 1, \ldots, N, \tag{3.1}$$

where $\mathbf{z}_i$ is the treatment vector with $(k-1)$ binary variables, $\boldsymbol{\beta}$ is the vector of the average treatment effects, $\mathbf{b}_i$ is the vector of subject-specific treatment effects corresponding to $\mathbf{z}_i$ for the ith subject, b_{0i} is a random intercept, and $\boldsymbol{\beta}_0$ are fixed effects corresponding to covariates $\mathbf{x}_{it}^0$. For example, given three treatments (A, B, C), the vector $\mathbf{z}_i = (z_{1i}, z_{2i})$ represents the treatment the ith subject receives, where $\mathbf{z}_i = (1,0)$, $(0,1)$, and $(0,0)$ corresponds to treatments A, B, and C, respectively.

We propose a two-step procedure to identify the best treatment for the ith individual by comparing the random effects corresponding to each treatment. In the first step, we estimate the random effects by solving (3.1), where the normality assumption for the random effects $\mathbf{b}_i$ is not required. This step provides the estimate of the treatment effect only if patients receive the corresponding treatment. In the second step, we grow a random forest using the estimated random effects as the response variable and patient information as the covariates. This allows us to compare all treatment effects even if some patients do not receive certain treatments. For example, if a patient i receives treatment A, we first estimate the effect of treatment A by $b_{1i} + b_{0i}$ for the ith patient; in the second step, we grow a random forest using the estimate of $b_{1i} + b_{0i}$ as the response variable, and the patient information as the covariates. This strategy allows patients not receiving treatment A to have their effects predicted for treatment A through the random forest.

The crucial step of solving the personalized treatment problem relies on accurate estimation for the subject-specific effects in the mixed-effects model (3.1). We estimate the random effect $\mathbf{b}_i$ without assuming the normality condition for the random effects, as in Wang, Tsai and Qu (2012). However, in contrast to their approach that ignores correlation information for random-effects estimation, we utilize the correlation of responses for random-effects estimation. We show in Sections 3.2 and 4 that incorporating correlation information for the random-effects estimation yields more accurate and efficient subject-specific effect estimation in both theory and practice.

We formulate the estimating equations corresponding to the random-effects parameters in (2.3) as

$$\mathbf{g}_i^C = \dot{\boldsymbol{\mu}}'_{i,\mathbf{b}_i} \mathbf{A}_i^{-1/2} \mathbf{C}^{-1} \mathbf{A}_i^{-1/2} (\mathbf{y}_i - \boldsymbol{\mu}_i), \ i = 1, \ldots, N, \tag{3.2}$$

where $\mathbf{C}$ is the correlation matrix estimator based on the method of moments, $1/N \sum_{i=1}^{N} (\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)'$. We construct extended scores with constraints of the mean and variance for the random effects $\mathbf{b}$ as

$$\mathbf{G}^c(\mathbf{b}) = \{(\mathbf{g}_1^C)', \ldots, (\mathbf{g}_N^C)', \lambda_1 \mathbf{b}', \lambda_2 (\mathbf{P_J b})'\}'. \tag{3.3}$$

We utilize tuning parameters λ_1 and λ_2 in (3.3), and provide a data-dependent cross-validation approach to select λ_1 and λ_2 in Section 3.4. The fixed-effects and random-effects parameters are obtained by iteratively minimizing $L^f(\boldsymbol{\beta})$ in (2.5) and the distance function

$$L^c(\mathbf{b}) = (\mathbf{G}^c)'(\mathbf{G}^c). \tag{3.4}$$

We apply nonparametric regression methods to develop a personalized treatment rule. This is achieved by taking the observed data and the estimated random effects associated with individuals into consideration. Specifically, we treat

the estimator of the random effects $\hat{\mathbf{b}}_i$ in (3.1) as the response, $\mathbf{x}_{it}^1$ as the covariates, and apply supervised learning algorithms to get better predictions for individual treatment effects. Among them, we adopt the random forest algorithm due to its accurate prediction and interpretation for tree-based models without parametric assumptions. It also provides ranks for the covariates' importance which identifies important factors influencing the treatment outcome. The random forest algorithm is an ensemble of a number of regression trees to formulate a prediction rule. Specifically, we grow a regression tree from each bootstrap sample, and the final prediction is obtained through averaging over all the estimators from the regression trees. In building regression trees, we consider a random sample of the covariates for splitting a tree, and choose the size of random samples as the square root of the number of covariates. Here we do not need to use all the covariates for the random samples, as we can reduce the correlations among the regression trees and therefore reduce the variation of the final prediction.

3.2. Asymptotic properties

In this section, we investigate the asymptotic properties of the proposed random-effects estimators in the generalized linear mixed model (2.1), and show that it is important to incorporate correlation information for random-effects estimation. Let $\mathbf{b}_0 = (\mathbf{b}_{01}', \ldots, \mathbf{b}_{0N}')'$ be the true realization of the random effects and let $\hat{\mathbf{b}} = (\hat{\mathbf{b}}_1', \ldots, \hat{\mathbf{b}}_N')'$ be the corresponding random-effects estimators, where $\mathbf{b}_{0i}$ and $\hat{\mathbf{b}}_i$ are $q \times 1$ vectors of random effects for the ith subject.

Theorem 1. *Under the regularity conditions provided in the Appendix,* $||\hat{\mathbf{b}}_i - \mathbf{b}_{0i}|| = O_p(n^{-1/2})$ *where* $||\cdot||$ *is the Euclidean norm.*

In general, when the repeated observations are independent conditional on the random effects, it is relatively straightforward to obtain $\sqrt{n}$-consistency for the random-effects estimation. Here it is challenging due to an additional serial correlation from repeated measurements conditional on the random effects; standard techniques for random-effects estimation for the independent case are not applicable for correlated data. To obtain the $\sqrt{n}$-consistency of random-effects estimation, we impose an L_2-mixingale condition on the serial correlations for the repeated measurements; the details of this are in the Appendix.

In contrast to the existing approaches assuming independent working structure for estimating the random effects (Breslow and Clayton (1993); Lee and Nelder (1996); Jiang (1999); and Wang, Tsai and Qu (2012)), we incorporate the serial correlation information for random-effects estimation. In the following result, we show that our proposed estimator $\hat{\mathbf{b}}$ improves on the random-effects estimator $\hat{\mathbf{b}}^I$ of Wang, Tsai and Qu (2012).

Theorem 2. *Under the regularity conditions provided in the Appendix,* $\mathrm{Var}\,(\mathbf{a}'\hat{\mathbf{b}}_i) \leq \mathrm{Var}\,(\mathbf{a}'\hat{\mathbf{b}}_i^I)$ *for any constant vector* $\mathbf{a}$.

If the estimated correlation matrix $\mathbf{C}$ in (3.2) is consistent with the true correlation matrix, the estimator $\hat{\mathbf{b}}$ is also optimal among all estimators solved by a linear combination of estimating equations $\mathbf{G}_i(\mathbf{b}_i) = \big\{\dot{\boldsymbol{\mu}}'_{i,\mathbf{b}_i}\mathbf{A}_i^{-1/2}\mathbf{M}_1\mathbf{A}_i^{-1/2}(\mathbf{y}_i - \boldsymbol{\mu}_i), \ldots, \dot{\boldsymbol{\mu}}'_{i,\mathbf{b}_i}\mathbf{A}_i^{-1/2}\mathbf{M}_m\mathbf{A}_i^{-1/2}(\mathbf{y}_i - \boldsymbol{\mu}_i)\big\}'$ for $i = 1, \ldots, N$. The proofs of Theorems 1 and 2 are provided in the Appendix.

3.3. Implementation and algorithm

In this section, we provide an algorithm to estimate fixed-effects and random-effects parameters, and we formulate a prediction rule for the personalized treatment. We demonstrate a cross-validation approach to select tuning parameters, and an implementation strategy for handling unbalanced longitudinal data.

We develop the personalized treatment assignment strategy using a two-step procedure.

Step 1: Obtain the random-effect estimator $\hat{\mathbf{b}}$ using the procedure described in *Step* 1.1−1.6.

Step 1.1: Set the initial value of random effects as $\hat{\mathbf{b}} = 0$.

Step 1.2: Obtain the initial estimate of $\boldsymbol{\beta}$ by minimizing (2.5).

Step 1.3: Update the correlation matrix $\mathbf{C}$ using the current estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{b}}$.

Step 1.4: Replace $\boldsymbol{\beta}$ with the current $\hat{\boldsymbol{\beta}}$ and update $\hat{\mathbf{b}}$ by minimizing (3.4).

Step 1.5: Replace $\mathbf{b}$ with $\hat{\mathbf{b}}$ and update $\hat{\boldsymbol{\beta}}$ by minimizing (2.5).

Step 1.6: Repeat *Steps* 1.3−1.5 until the convergence criterion is reached.

Step 2: Perform the random forest algorithm using the R package randomForest (Breiman (2001)) by treating $\hat{\mathbf{b}}$ as the response and covariates $\mathbf{x}_{it}^1$ as the predictors.

In the first step, for fixed λ_1 and λ_2, the random effects are estimated by minimizing two objective functions (2.5) and (3.4) iteratively. In the second step, we treat the random-effects estimator $\hat{\mathbf{b}}$ as the response variable and apply the random forest algorithm. The R coding of the above procedure is provided in the online supplementary material.

The choice of tuning parameters is essential to accurate estimation of the fixed and random effects. Wang, Tsai and Qu (2012) choose both λ_1 and λ_2 to be $\log(N)$. Random-effects estimation is less sensitive to the choice of λ_2, since it ensures the identifiability of random effects, and we fix λ_2 to be $\log(N)$.

The choice of λ_1 is more critical as random-effects estimation depends on the magnitude of its variance. Here we adopt a cross-validation approach for longitudinal data to tune the parameter λ_1. We fit the model with a parameter λ_1 to all observations except the kth repeated measurement, and obtain $\hat{\boldsymbol{\beta}}^{-k}(\lambda_1)$ and $\hat{\mathbf{b}}^{-k}(\lambda_1)$ following Step 1 of the algorithm. We compute the prediction error for the kth measurement based on this estimator as

$$PE_k(\lambda_1) = \sum_{i=1}^{N} \Big[y_{ik} - \mu\big\{ \mathbf{z}'_{ik}\hat{\mathbf{b}}^{-k}(\lambda_1) + \mathbf{x}'_{ik}\hat{\boldsymbol{\beta}}^{-k}(\lambda_1) \big\} \Big]^2, k = 1, \ldots, K,$$

where $K = \min(n_1, \ldots, n_N)$. The cross-validation error is specified as $CV(\lambda_1) = (1/K)\sum_{k=1}^{K} PE_k(\lambda_1)$. We choose the value of λ_1 to minimize $CV(\lambda_1)$.

We follow Wang, Tsai and Qu's (2012) strategy to implement our method for unbalanced data. Let t_n be the cluster size of subjects for fully observed data without any missing observation. Let $\mathbf{T}_i$ be a $t_n \times t_{n_i}$ transformation matrix for the ith cluster, where the $\mathbf{T}_i$'s are generated by deleting the columns of the $t_n \times t_n$ identity matrix corresponding to missing observations. We then transform the response variable, the mean and marginal variance of the responses for the ith subject, using $\boldsymbol{\mu}_i^* = \mathbf{T}_i\boldsymbol{\mu}_i$, $\dot{\boldsymbol{\mu}}_{i,\beta}^* = \mathbf{T}_i\dot{\boldsymbol{\mu}}_{i,\beta}$, $\mathbf{y}_i^* = \mathbf{T}_i\mathbf{y}_i$, and $\mathbf{A}_i^* = \mathbf{T}_i\mathbf{A}_i\mathbf{T}_i'$. We replace $\mathbf{g}_i^f$ in (2.4) by $\mathbf{g}_i^{f*}$ through substituting $\mathbf{y}_i$, $\boldsymbol{\mu}_i$, $\dot{\boldsymbol{\mu}}_{i,\beta}$, and $\mathbf{A}_i$ by $\mathbf{y}_i^*$, $\boldsymbol{\mu}_i^*$, $\dot{\boldsymbol{\mu}}_{i,\beta}^*$, and $\mathbf{A}_i^*$ to estimate the fixed-effects parameters. In similar fashion, we formulate estimating equations $\mathbf{g}_i^{C*}$ to estimate random-effects parameters for the unbalanced data.

4. Simulation Studies

We conducted simulation studies to evaluate the performance of the proposed method in identifying the optimal treatment for individuals. The conditional correlated responses were generated using the linear mixed model for longitudinal data given by

$$y_{it} = z_i(\beta_1 + b_{1i}) + \beta_0 + b_{0i} + e_{it}, \quad \text{for } i = 1, \ldots, N \text{ and } t = 1, \ldots, n_i,$$

where treatments z_i were generated from a Bernoulli, ($P(z_i = 1) = 0.5$), determining treatment, 1, or control, 0. Here $\mathbf{e}_i = (e_{i1}, \ldots, e_{in_i})' \sim N(0, \mathbf{R})$ with $\mathbf{R}$ being an AR(1) correlation matrix with the correlation coefficient $\rho = 0.8$. We set $\beta_0 = \beta_1 = 0$, indicating no overall treatment effect for the population. Baseline covariates $x_{i,j}$, $j = 1, \ldots, 6$, were generated independently from a uniform (-1,1). We generated random effects as $b_{0i} = x_{i,1}$ and $b_{1i} = x_{i,2} - x_{i,1}$. The expected value of the ith subject was $x_{i,2}$ if the treatment was received (with $z_i = 1$), and $x_{i,1}$ otherwise. If the ith random slope b_{1i} is positive, then the treatment is beneficial for the ith subject. The sample size was N = 100 or 300 in our simulations.

Table 1. Performance of the penalized quasi-likelihood (PQL), Wang, Tsai, and Qu's approach (WQIF), and the proposed approach (PQIF). Random forest (RF), support vector machines (SVM), and decision trees (TREE) were implemented for prediction.

N	$\max(n_i)$	Method	MSE($\hat{\mathbf{b}}$)	PE($\tilde{\mathbf{b}}$)			CE		
				RF	TREE	SVM	RF	TREE	SVM
		PQIF	0.231	0.118	0.132	0.134	0.19	0.22	0.21
	5	WQIF	0.484	0.192	0.207	0.170	0.25	0.28	0.24
		PQL	0.615	0.171	0.208	0.218	0.23	0.27	0.25
100									
		PQIF	0.216	0.107	0.112	0.115	0.17	0.21	0.20
	10	WQIF	0.419	0.145	0.174	0.147	0.21	0.22	0.21
		PQL	0.554	0.141	0.176	0.161	0.20	0.23	0.22
		PQIF	0.226	0.090	0.098	0.108	0.14	0.20	0.16
	5	WQIF	0.434	0.148	0.151	0.150	0.18	0.22	0.19
		PQL	0.635	0.113	0.140	0.231	0.16	0.21	0.18
300									
		PQIF	0.190	0.070	0.074	0.083	0.13	0.18	0.15
	10	WQIF	0.382	0.091	0.112	0.135	0.15	0.20	0.16
		PQL	0.512	0.080	0.115	0.167	0.16	0.20	0.17

The repeated measurements were monotone missing, and the cluster sizes were unequal, with 70% of subjects having $n_i = 10$ and 30% of having $n_i = 8$. We also investigated performance with relatively small cluster sizes $n_i = 5$ for 70% and $n_i = 4$ for 30% of subjects.

We compared the proposed mixed-effects approach utilizing the correlation structure for the fixed-effect estimation to the penalized quasi-likelihood method and to Wang, Tsai and Qu's (2012) mixed-effects approach. We generated 200 simulated data sets and report the mean square error for $\hat{\mathbf{b}}$, $\text{MSE}(\hat{\mathbf{b}}) = \sum_{k=1}^{200}\sum_{i=1}^{N}||\hat{\mathbf{b}}_i^{(k)} - \mathbf{b}_i||^2/200N$, where $\hat{\mathbf{b}}_i^{(k)}$ is the estimator of $\mathbf{b}_i = (b_{0i}, b_{1i})$ from the kth simulation, $\mathbf{b}_i$ is the true parameter, and $||\cdot||$ denotes the Euclidean norm. Table 1 shows that the proposed mixed-effects approach performs the best in terms of the mean squared errors of $\hat{\mathbf{b}}$ for all cases, and substantially so.

We treated the random-effects estimators as the responses, and investigated how the six baseline covariates (two of them relevant) were associated with the response based on three supervised learning algorithms: the random forest, the support vector machine, and the decision tree. To evaluate their performance, we built a predictive model for random effects for the training data (randomly selecting 80% of the subjects), and checked the validity of the model through the prediction error for the testing data (the other 20% of the subjects). We took

$\mathrm{PE}(\tilde{\mathbf{b}}) = \sum_{k=1}^{200} \sum_{i=1}^{N/5} ||\tilde{\mathbf{b}}_i^{(k)} - \mathbf{b}_i||^2/40N$, where $\tilde{\mathbf{b}}_i^{(k)}$ was the predicted value for the ith subject in the testing set from the kth simulation. The classification error rate was the proportion of the subjects whose sign of the predicted random slope effect was different from that of the true parameter in the testing data.

In Table 1, the random forest outperforms the support vector machine and the decision tree in terms of the prediction error and the classification error rate. The classification error rate and the prediction error based on the proposed method were smaller than those of Wang, Tsai and Qu (2012) and the penalized quasi-likelihood approaches in all cases.

5. Data Analysis for a HIV Study

In this section, we use data from the Harvard AIDS clinical trial group (ACTG) 116A, 116B/117 studies to identify subgroups of patients with HIV who might benefit from a certain treatment. These patients were randomly assigned to one of three treatments (500 mg didanosine, 750 mg didanosine, or zidovudine), and each patient was measured nine times at weeks 0, 2, 8, 12, 16, 24, 32, 40, and 48. To incorporate the serial correlations over time, we selected 410 patients having at least the first 7 measurements. The objective of our study was to determine whether a certain treatment can improve a patient's condition and significantly slow down the progression of HIV. Here we used the CD4 cell counts as response measurements since the decrease in CD4 cell counts indicates a deterioration of the immune system.

A marginal regression model provides the association between the CD4 cell counts and the drug effect:

$$CD4_{it} = \beta_0 + \beta_1 ddi500_i + \beta_2 ddi750_i + \beta_3 CD4_{0i} + e_{it}, \tag{5.1}$$

where $CD4_{it}$ are the CD4 cell counts, $ddi500_i = 1$ if the ith patient receives 500 mg didanosine treatment and 0 otherwise, $ddi750_i = 1$ if the ith patient receives 750 mg didanosine treatment and 0 otherwise, and $CD4_0$ are the CD4 cell counts at the initial visit. We estimated the fixed-effects parameters, using the generalized estimating equation approach, for all patients to compare the three treatment effects. Here the AR(1) working correlation structure was assumed for the generalized estimating equation approach. Table 2 provides estimators for the fixed effects, their standard errors, the Wald test statistics, and the corresponding p-values. The conditional mean of the CD4 cell counts with 750mg didanosine is higher than those of either 500mg didanosine or zidovudine. However, the treatment effect is not significant for the marginal model using all patients, indicating that there is no difference among treatments in controlling the progression of CD4 cell counts.

Table 2. For the HIV disease study, coefficients estimated by the generalized estimating equation approach along with the standard errors (s.e.), Wald test statistics, and the p-values for all subjects (Overall) and three identified subgroups.

Group	Effect	Estimator	s.e.	Wald	p-value
Overall	*intercept*	11.21	5.02	4.99	0.025
	*ddi*500	-13.86	7.90	3.08	0.079
	*ddi*750	1.57	7.85	0.04	0.841
	$CD4_0$	0.99	0.04	750.26	< 0.001
500mg ddi beneficiary	*intercept*	-1.75	9.42	0.03	0.852
	*ddi*500	30.32	15.04	4.06	0.043
	*ddi*750	4.37	17.17	0.06	0.799
	$CD4_0$	1.03	0.08	179.38	< 0.001
750mg ddi beneficiary	*intercept*	15.78	9.73	2.63	0.105
	*ddi*500	-31.80	12.51	6.47	0.010
	*ddi*750	5.11	13.98	0.13	0.714
	$CD4_0$	1.04	0.05	493.08	< 0.001
zidovudine beneficiary	*intercept*	18.76	7.08	7.02	0.008
	*ddi*500	-33.55	11.44	8.60	0.003
	*ddi*750	-1.82	10.72	0.03	0.865
	$CD4_0$	0.88	0.06	225.22	< 0.001

We applied the proposed method to investigate whether certain subgroups of patients are more likely to benefit from any of these treatments. The linear mixed-effects model was specified as

$$CD4_{it} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})ddi500_i + (\beta_2 + b_{2i})ddi750_i + \beta_3 CD4_{0i} + e_{it}, \quad (5.2)$$

where $\mathbf{b}_i = (b_{0i}, b_{1i}, b_{2i})'$ is a vector of the ith patient's treatment effect. We estimated the random effect $\mathbf{b}_i$ and applied the random forest to establish a relationship between the random-effects estimates and the patient's individual characteristics such as age, weight, gender, ethnic origin, sexual orientation, and positive/negative AIDs diagnosis. Based on the personalized treatment assignment in Section 3.1, we defined three beneficiary groups corresponding to each treatment, and estimated the parameter $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$ in (5.1) using the generalized estimating equation approach for them. The results are summarized in Table 2. The coefficient of *ddi*500 in the 500mg didanosine beneficiary group and the coefficient of *ddi*750 in the 750 mg didanosine group are positive, while both coefficients in the zidovudine group are negative.

To validate whether the strategy is effective or not, we split the HIV data as 2/3 training set and 1/3 testing set. We built the prediction model using the

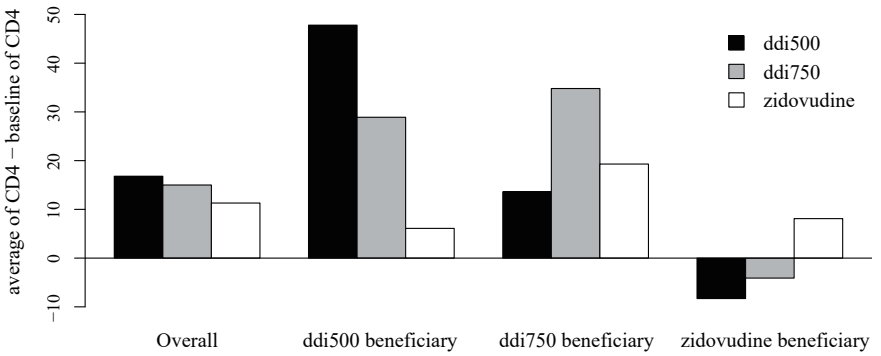

Figure 1. For the testing data, a comparison of the differences between average CD4 cell counts and the baseline CD4 cell counts corresponding to three treatments among all patients (Overall), the 500mg didanosine beneficiary group, the 750mg didanosine beneficiary group and the zidovudine beneficiary group.

training data and identified the 'best' treatment for individuals in the testing data. Figure 1 provides the differences between the average CD4 cell counts and its baseline for patients in the testing data. There was not much difference among the three treatment groups for the entire test data. However, once we divided the test data into three beneficiary subgroups based on our prediction model, the CD4 cell counts were the highest for patients receiving the 'best' treatment compared to the ones receiving other treatments in each beneficiary subgroup. This suggests that the proposed approach can effectively identify ideal treatments for individuals.

We further compared the performance of the proposed method with the mixed-effects model of Wang, Tsai and Qu (2012) and the penalized quasi-likelihood approaches. We utilized the random forest algorithm to establish the predictive model based on training data (randomly selecting 80% of the subjects) and computed the prediction error from the rest of the patients. Here the true random effect is unknown in practice. Thus, we took the prediction error for the testing data to be $\mathrm{PE}(\tilde{\mathbf{b}}) = \sum_{i=1}^{n} |\tilde{\mathbf{b}}_i - \hat{\mathbf{b}}_i|^2/n$, where $\tilde{\mathbf{b}}_i$ is the predicted value for the ith subject in the testing data, $\hat{\mathbf{b}}_i$ is its random-effect estimator, and n is the size of the testing data. The prediction error using the proposed method ($\mathrm{PE}(\tilde{\mathbf{b}}) = 6.8$) was smaller than those based on Wang, Tsai, and Qu's approach ($\mathrm{PE}(\tilde{\mathbf{b}}) = 12.1$) and the penalized quasi-likelihood method ($\mathrm{PE}(\tilde{\mathbf{b}}) = 15.0$). In summary, the proposed personalized treatment assignment approach was efficient in identifying the effective treatment for individuals.

6. Discussion

The proposed personalized treatment strategy for longitudinal data is useful when subjects show great variations in response to different treatments. We are able to assign an optimal treatment to new patients according to their own characteristics. This is achieved through a random-effects estimation method that does not require a specified distribution of the random effects. This approach is able to capture the subject-specific variations of treatment effects, and to distinguish treatment outcomes among different subjects. In general, the mixed-effects approaches yield more accurate estimators when the cluster size is large. Our simulation studies, however, indicate that our proposed method still performs well with relatively small cluster sizes.

A major advantage of the proposed strategy is that we can significantly reduce the error rate in classifying subjects to the wrong treatment group. Since our random-effects approach utilizes the correlation information for repeated measurements in random-effects estimation. In addition, we apply random forests to improve accuracy in predicting individual treatments. Asymptotically, we show that the efficiency of the proposed random-effects estimators is better than that of Wang, Tsai and Qu's (2012) random-effects estimators that assume independence.

The proposed approach is generally applicable when there is a common correlation structure of repeated measurements across subjects. If repeated measurements from different subjects are collected at different time points and there is no shared information of correlation structure among subjects, we can apply the nonparametric functional data approach. For example, Li (2011) models the covariance function nonparametrically using the kernel smoothing approach that does not require parametric modeling of the covariance structure. However, the functional data analysis approach typically requires more data collection from each subject and it does not take into account serial correlations for the errors. Further investigation on this topic is needed.

We can apply the proposed method to investigate the time-varying treatment effect through modeling the random effects as a function of time. Wu and Liang (2004) propose the varying-coefficient model to estimate coefficient functions for random effects. Their method can be applied for identifying time-varying personalized treatment, although their method assumes that the random effects are normal. Theoretical development and computational implementation for a time-varying random-effects model without normality for random effects could be quite challenging, but should be a valuable future research direction.

Supplementary Material

The R coding for our simulation studies is given in the online supplementary material available at `http://www.stat.sinica.edu/statistica`.

Acknowledgement

The authors are very grateful to the Co-Editor, two referees and an associate editor for their insightful comments and suggestions that have improved the manuscript significantly. Qu's research is supported by the National Science Foundation (DMS-1308227).

Appendix: Proofs of Theorems

Suppose $\boldsymbol{\beta}_0$ is the true parameter of fixed effects and that $\hat{\boldsymbol{\beta}}$ is the estimator of fixed effects obtained by minimizing (2.5) conditional on the random-effects estimators $\hat{\mathbf{b}}$. We need some regularity conditions for establishing its asymptotic properties:

(1) The parameter space is compact.

(2) If $\mathbf{U}$ is the weighting matrix, decomposed as $\mathbf{C}^{-1} = \mathbf{U}'\mathbf{U}$, $\mathbf{C}$ and $\mathbf{U}$ converge almost surely to constant matrices $\mathbf{C}_0$ and $\mathbf{U_0}$ that satisfy $\mathbf{C}_0^{-1} = \mathbf{U}_\mathbf{0}'\mathbf{U_0}$.

(3) $E\{\mathbf{g}^f(\boldsymbol{\beta}|\mathbf{b})\}$ and $E\{\mathbf{g}_i^C(\boldsymbol{\beta}|\mathbf{b}_i)\}$ are continuous and differentiable.

(4) There exists a unique $\boldsymbol{\beta}_0$ such that $E\{\mathbf{g}^f(\boldsymbol{\beta}_0|\mathbf{b}_0)\} = 0$.

(5) The expectation of $E\{\mathbf{g}^f(\boldsymbol{\beta}|\hat{\mathbf{b}})\}$ with respect to the random-effect parameters converges to 0 in probability as $N \to \infty$.

(6) The derivative of $\mathbf{g}^f(\boldsymbol{\beta}|\mathbf{b}_0)$ with respect to the fixed-effects parameter, conditional on the true parameter of random effects, is uniformly bounded in probability.

(7) If $e_{ij} = y_{ij} - \mu_{ij}(\boldsymbol{\beta}|\mathbf{b}_i)$ is the residual for the jth observation of subject i, the residuals within the same subjects $(e_{i1}, \ldots, e_{in})$ satisfy $||E(e_{ij}|e_{i,j-m})||_2 \leq c_j\varphi_m$, for $j = 1, \ldots, n$ and $m = 1, \ldots, j-1$, and $||e_{ij} - E(e_{ij}|e_{i,j+m})||_2 \leq c_j\varphi_{m+1}$, for $j = 1, \ldots, n$ and $m = 1, \ldots, n-j$, where $||\cdot||_2$ is the L_2 norm, ψ_m are some non-negative constants such that $\psi_m \to 0$ as $m \to \infty$, and the c_j, $j \geq 1$, satisfy $\overline{\lim}_{n\to\infty}(1/n)\sum_{j=1}^n c_j < \infty$, or $\{c_j\}$ can be given by $\{||e_{ij}||_2\}$.

Proof of Theorem 1. We write $\mathbf{A}_i^{-1/2}\mathbf{C}^{-1}\mathbf{A}_i^{-1/2}$ as $\mathbf{Q}_i$. Since each element of $\mathbf{Q}_i$ is bounded in probability, the order of $||\mathbf{Q}_i||^1$ is between n and n^2, where $||\cdot||^1$ is the sum of all matrix entries' absolute values. Let $r_n \in [n, n^2]$ denote the order of $||\mathbf{Q}_i||^1$, $||\mathbf{Q}_i||^1 = O_p(r_n)$. We take $g_i(\boldsymbol{\beta}|\mathbf{b}_i) = \dot{\boldsymbol{\mu}}_{i,b_i}(\boldsymbol{\beta}|\mathbf{b}_i)'\mathbf{Q}_i\{\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}|\mathbf{b}_i)\} = \sum_{k=1}^n\sum_{j=1}^n c_{ikj}\dot{\mu}_{ik,b_i}e_{ij}$ for $i = 1, \ldots, N$, where $\dot{\boldsymbol{\mu}}_{i,b_i}(\boldsymbol{\beta}|\mathbf{b}_i) = \frac{\partial}{\partial b_i}\boldsymbol{\mu}_i(\boldsymbol{\beta}|\mathbf{b}_i) =$

$(\dot{\mu}_{i1,b_i}, \dot{\mu}_{i2,b_i}, \ldots, \dot{\mu}_{in,b_i})'$, c_{ikj} is the $(k \times j)$th component of $\mathbf{Q}_i$, and $e_{ij} = y_{ij} - \mu_{ij}(\boldsymbol{\beta}|\mathbf{b}_i)$. The estimator $\hat{\mathbf{b}}_i$ is obtained by solving $g_i(\hat{\boldsymbol{\beta}}|\mathbf{b}_i) = 0$.

By a Taylor expansion, we have $\mathbf{b}_{0i} - \hat{\mathbf{b}}_i = \{\dot{g}_{i,b_i}(\hat{\boldsymbol{\beta}}|\tilde{\mathbf{b}}_i)\}^{-1} g_i(\hat{\boldsymbol{\beta}}|\mathbf{b}_0)$, where $\dot{g}_{i,b_i}(\hat{\boldsymbol{\beta}}|\tilde{\mathbf{b}}_i) = \frac{\partial}{\partial \mathbf{b}_i} g_i(\hat{\boldsymbol{\beta}}|\mathbf{b}_i) \mid_{\mathbf{b}_i = \tilde{\mathbf{b}}_i}$ and $\tilde{\mathbf{b}}_i$ is between $\hat{\mathbf{b}}_i$ and $\mathbf{b}_{0i}$. If $\hat{\boldsymbol{\beta}} \xrightarrow{\mathrm{p}} \boldsymbol{\beta}_0$, it follows that $\mathbf{b}_{0i} - \hat{\mathbf{b}}_i \to \{\dot{g}_{i,\mathbf{b}_i}(\boldsymbol{\beta}_0|\tilde{\mathbf{b}}_i)\}^{-1} g_i(\boldsymbol{\beta}_0|\mathbf{b}_0)$. Since $\dot{\mu}_{ij,\mathbf{b}_i}(\boldsymbol{\beta}|\tilde{\mathbf{b}}_i)$ is bounded in probability, we have $\{\dot{g}_{i,\mathbf{b}_i}(\boldsymbol{\beta}_0|\tilde{\mathbf{b}}_i)\}^{-1} = O_p(r_n^{-1})$. There then exists a constant K_1 such that

$$|g_i(\boldsymbol{\beta}_0|\mathbf{b}_0)| = \left| \sum_{k=1}^{n} \sum_{j=1}^{n} c_{ikj} \dot{\mu}_{ik,b_i} e_{ij} \right| \leq K_1 r_n \left| \frac{n}{r_n} \sum_{k=1}^{n} c_{ikj} \frac{1}{n} \sum_{j=1}^{n} e_{ij} \right| = O_p(r_n \bar{e}_i),$$

where $(n/r_n) \sum_{k=1}^{n} c_{ikj} = O_p(1)$ and $(1/n) \sum_{j=1}^{n} e_{ij} = \bar{e}_i$. It follows from $\{\dot{g}_{i,\mathbf{b}_i}(\boldsymbol{\beta}_0|\tilde{\mathbf{b}}_i)\}^{-1} = O_p(r_n^{-1})$ and $g_i(\boldsymbol{\beta}_0|\mathbf{b}_0) = O_p(r_n \bar{e}_i)$ that $\mathbf{b}_{0i} - \hat{\mathbf{b}}_i = O_p(\bar{e}_i)$. Therefore, it suffices to show that $\bar{e}_i = O_p(n^{-1/2})$, equivalently $E(|\bar{e}_i|^2) = O_p(n^{-1})$, since for any ϵ, there exists a constant K_2 such that $P\big[|\sqrt{n}\bar{e}_i| > K_2\big] \leq K_2^{-2} n E(|\bar{e}_i|^2) < \epsilon$. Under the condition that the sequence of random variables e_{ij} satisfies the L_2 mixingale condition and $\sum_{k=1}^{\infty} \varphi_k < \infty$,

$$\begin{aligned} E(|\bar{e}_i|^2) &\leq \frac{2}{n^2} \sum_{t=1}^{n} \sum_{j=1}^{t} E|e_{ij} e_{it}| \leq \frac{2}{n^2} \sum_{t=1}^{n} \sum_{j=1}^{t} E(|e_{ij}||E(e_{it}|e_{ij})|) \\ &\leq \frac{2}{n^2} \sum_{t=1}^{n} \sum_{j=1}^{t} ||e_{ij}||_2 ||E(e_{it}|e_{ij})||_2 \\ &\leq \frac{2}{n^2} \sum_{t=1}^{n} \sum_{j=1}^{t} v_j v_t \varphi_{t-j} = \frac{2}{n^2} \sum_{k=1}^{n} \varphi_k \sum_{j=1}^{n-k} v_j v_{j+k}, \end{aligned}$$

where $||e_{ij}||_2 = (E(|e_{ij}|^2))^{1/2}$ is bounded by v_j, and this implies $E(|\bar{e}_i|^2) = O_p(n^{-1})$.

Proof of Theorem 2. The inverse of the estimated correlation matrix $\mathbf{C}$ can be decomposed as $\mathbf{C}^{-1} = a_0 \mathbf{I} + \mathbf{D}$, where $\mathbf{I}$ is an identity matrix and a_0 is an unknown coefficient. Then, the random-effects estimator $\hat{\mathbf{b}}_i$ can be obtained by solving

$$\dot{\boldsymbol{\mu}}_{i,\mathbf{b}_i}(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)' \mathbf{A}_i^{-1/2} \mathbf{C}^{-1} \mathbf{A}_i^{-1/2} \{\mathbf{y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)\} = a_0 g_I(\hat{\boldsymbol{\beta}}|\mathbf{b}_i) + g_D(\hat{\boldsymbol{\beta}}|\mathbf{b}_i) = 0, \quad \text{(A.1)}$$

where $\dot{\boldsymbol{\mu}}_{i,\mathbf{b}_i}(\hat{\boldsymbol{\beta}}|\mathbf{b}_i) = \frac{\partial}{\partial \mathbf{b}_i} \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)$, $g_I(\hat{\boldsymbol{\beta}}|\mathbf{b}_i) = \dot{\boldsymbol{\mu}}_{i,\mathbf{b}_i}(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)' \mathbf{A}_i^{-1} \{\mathbf{y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)\}$ and $g_D(\hat{\boldsymbol{\beta}}|\mathbf{b}_i) = \dot{\boldsymbol{\mu}}_{i,\mathbf{b}_i}(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)' \mathbf{A}_i^{-1/2} \mathbf{D} \mathbf{A}_i^{-1/2} \{\mathbf{y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)\}$. In order to separate the contribution of $g_I(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)$ and $g_D(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)$ for random-effects estimation, we orthogonalize $g_D(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)$ from $g_I(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)$ as $g_D^*(\hat{\boldsymbol{\beta}}|\mathbf{b}_i) = g_D(\hat{\boldsymbol{\beta}}|\mathbf{b}_i) - \mathbf{W}_{21} \mathbf{W}_{11}^{-1} g_I(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)$, where $\mathbf{W}_{21} = cov(g_D(\hat{\boldsymbol{\beta}}|\mathbf{b}_i), g_I(\hat{\boldsymbol{\beta}}|\mathbf{b}_i))$ and $\mathbf{W}_{11} = \mathrm{Cov}\,(g_I(\hat{\boldsymbol{\beta}}|\mathbf{b}_i))$. Through the orthogonalization, $cov(g_D^*(\hat{\boldsymbol{\beta}}|\mathbf{b}_i), g_I(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)) = 0$.

It follows from the generalized method of moments that solving (A.1) is equivalent to minimizing $G(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)'\mathbf{W}_G^{-1}G(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)$, where $G(\hat{\boldsymbol{\beta}}|\mathbf{b}_i) = (g_I(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)', g_D^*(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)')'$ and $\mathbf{W}_G = \mathrm{Cov}\,(G(\hat{\boldsymbol{\beta}}|\mathbf{b}_i))$. The inverse of the asymptotic covariance for the random-effects estimator $\hat{\mathbf{b}}_i$ is proportional to

$$\dot{G}_{\mathbf{b}_i}(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)'\mathbf{W}_G^{-1}\dot{G}_{\mathbf{b}_i}(\hat{\boldsymbol{\beta}}|\mathbf{b}_i) = \dot{g}_{I,\mathbf{b}_i}(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)'\mathbf{W}_{11}^{-1}\dot{g}_{I,\mathbf{b}_i}(\hat{\boldsymbol{\beta}}|\mathbf{b}_i) + \dot{g}_{D,\mathbf{b}_i}^*(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)'\mathbf{W}_{22}^{*-1}\dot{g}_{D,\mathbf{b}_i}^*(\hat{\boldsymbol{\beta}}|\mathbf{b}_i),$$

where $\mathbf{W}_{22}^* = \mathrm{Cov}\,(g_D^*(\hat{\boldsymbol{\beta}}|\mathbf{b}_i))$, $\dot{G}_{\mathbf{b}_i}(\hat{\boldsymbol{\beta}}|\mathbf{b}_i) = \frac{\partial}{\partial \mathbf{b}_i}G(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)$, $\dot{g}_{I,\mathbf{b}_i}(\hat{\boldsymbol{\beta}}|\mathbf{b}_i) = \frac{\partial}{\partial \mathbf{b}_i}g_I(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)$, and $\dot{g}_{D,\mathbf{b}_i}^*(\hat{\boldsymbol{\beta}}|\mathbf{b}_i) = \frac{\partial}{\partial \mathbf{b}_i}g_D^*(\hat{\boldsymbol{\beta}}|\mathbf{b}_i)$. Here $\mathbf{W}_{22}^* = \mathbf{W}_2 - \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12}$, where $\mathbf{W}_2 = \mathrm{Cov}\,(g_D(\hat{\boldsymbol{\beta}}|\mathbf{b}_i))$. Under $\hat{\boldsymbol{\beta}} \xrightarrow{\mathrm{p}} \boldsymbol{\beta}_0$, it follows that

$$\dot{G}_{\mathbf{b}_i}(\boldsymbol{\beta}_0|\mathbf{b}_i)'\mathbf{W}_G^{-1}\dot{G}_{\mathbf{b}_i}(\boldsymbol{\beta}_0|\mathbf{b}_i) \to \dot{g}_{I,\mathbf{b}_i}(\boldsymbol{\beta}_0|\mathbf{b}_i)'\mathbf{W}_{11}^{-1}\dot{g}_{I,\mathbf{b}_i}(\boldsymbol{\beta}_0|\mathbf{b}_i) + \dot{g}_{D,\mathbf{b}_i}^*(\boldsymbol{\beta}_0|\mathbf{b}_i)'\mathbf{W}_{22}^{*-1}\dot{g}_{D,\mathbf{b}_i}^*(\boldsymbol{\beta}_0|\mathbf{b}_i).$$

Since $\mathbf{W}_{22}^*$ is a non-negative definite weighting matrix, $\dot{G}_{\mathbf{b}_i}(\boldsymbol{\beta}_0|\mathbf{b}_i)'\,\mathbf{W}_G^{-1}\dot{G}_{\mathbf{b}_i}(\boldsymbol{\beta}_0|\mathbf{b}_i) \geq \dot{g}_{I,\mathbf{b}_i}(\boldsymbol{\beta}_0|\mathbf{b}_i)'\mathbf{W}_{11}^{-1}\dot{g}_{I,\mathbf{b}_i}(\boldsymbol{\beta}_0|\mathbf{b}_i)$ in the sense of the Loewner ordering. Therefore, the efficiency of the random-effects estimator is improved by utilizing the estimated correlation matrix.

References

Biau, G. (2012). Analysis of a random forests model. *J. Mach. Learn. Res.* **13**, 1063-1095.

Biau, G., Devroye, L. and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.* **9**, 2015-2033.

Bonetti, M. and Gelber, R. (2000). A graphical method to assess treatment-covariate interactions using the Cox model on subsets of the data. *Statist. Medicine* **19**, 2595-2609.

Bonetti, M. and Gelber, R. (2004). Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics* **5**, 465-481.

Breiman, L. (2001). Random forests. *Mach. Learn.* **45**, 5-32.

Breiman, L. (2004). Consistency for a simple model of random forests. Technical report, University of California at Berkeley.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88**, 9-25.

Cai, T., Tian, L., Wong, P. H. and Wei, L. J. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* **12**, 270-282.

Denil, M., Matheson, D. and de Freitas, N. (2014). Narrowing the gap: Random forests in theory and in practice. *J. Mach. Learn. Res.: Workshop and Conference Proceedings* **32**, 665-673.

Diaz, F. J., Rivera, T. E., Josiassen, R. C. and Leon J. (2007). Individualizing drug dosage by using a random intercept linear model. *Statist. Medicine* **26**, 2052-2073.

Diaz, F. J., Yeh, H. W. and Leon, J. (2012). Role of statistical random-effects linear models in personalized medicine. *Current Pharmacogenomics and Personalized Medicine* **10**, 22-32.

Foster, J. C., Taylor, J. M. C. and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statist. Medicine* **30**, 2867-2880.

Jiang, J. (1999). Conditional inference about generalized linear mixed models. *Ann. Statist.* **27**, 1974-2007.

Jiang, J. and Zhang, W. (2001). Robust estimation in generalised linear mixed models. *Biometrika* **88**, 753-765.

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

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *J. Roy. Statist. Soc. Ser. B* **58**, 619-678.

Li, Y. (2011). Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation. *Biometrika* **98**, 355-370.

Lipkovich, I., Dmitrienko, A., Denne, J. and Enas, G. (2011). Subgroup identification based on differential effect search–recursive partitioning method for establishing response to treatment in patient subpopulations. *Statist. Medicine* **30**, 2601-2621.

Marlowe, D. B., Festinger, D. S., Dugosh, K. L., Lee, P. A. and Benasutti, K. M. (2007). Adapting judicial supervision to the risk level of drug offenders: Discharge and 6-month outcomes from a prospective matching study. *Drug and Alcohol Dependence* **88**, 4-13.

McCulloch, C. E. (1997). Maximum likelihood algorithm for generalized linear mixed models. *J. Amer. Statist. Assoc.* **92**, 162-170.

McCulloch, C. E., Searle, S. R. and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*. 2nd edition. John Wiley, New York.

Neuhaus, J. M., Hauck, W. W. and Kalbfleisch, J. D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* **79**, 755-762.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.* **39**, 1180-1210.

Qu, A., Lindsay, B. G. and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87**, 823-836.

Song, X. and Pepe, M. S. (2004). Evaluating markers for selecting a patient's treatment. *Biometrics* **60**, 874-883.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M. and Li, B. (2009). Subgroup analysis via recursive partitioning. *J. Mach. Learn. Res.* **10**, 141-158.

Vonesh, E. F., Wang, H., Nie, L. and Majumdar, D. (2002). Conditional second-order generalized estimating equations for generalized linear and nonlinear mixed-effects models. *J. Amer. Statist. Assoc.* **97**, 271-283.

Wang, P., Tsai, G. F. and Qu, A. (2012). Conditional inference functions for mixed-effects models with unspecified random-effects distribution. *J. Amer. Statist. Assoc.* **107**, 725-736.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalised linear models, and the Gauss-Newton method. *Biometrika* **61**, 439-447.

Wu, H. L. and Liang, H. (2004). Backfitting random varying-coefficient models with time-dependent smoothing covariates. *Scand. J. Statist.* **31**, 3-19.

Zhao, Y., Zeng, D., Rush, A. J. and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.* **107**, 1106-1118.

Zhao, L., Tian, L., Cai, T., Chaggett, B. and Wei, L. (2013). Effectively selecting a target population for a future comparative study. *J. Amer. Statist. Assoc.* **108**, 527-539.

Zhou, J. and Qu, A. (2012). Informative estimation and selection of correlation structure for longitudinal data. *J. Amer. Statist. Assoc.* **107**, 701-710.

Department of Statistics, Western Michigan University, Kalamazoo, MI 49008, USA.

E-mail: Hyunkeun.Cho@wmich.edu

Department of Operations, Business Analytics and Information Systems, University of Cincinnati, Cincinnati, OH 45220, USA.

E-mail: Peng.Wang@uc.edu

Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA.

E-mail: Anniequ@illinois.edu

(Received April 2015; accepted December 2015)