Published in final edited form as:

Biometrics. 2017 December ; 73(4): 1343–1354. doi:10.1111/biom.12663.

Penalized Nonlinear Mixed Effects Model to Identify Biomarkers that Predict Disease Progression

Huaihou Chen¹, Donglin Zeng², and Yuanjia Wang³

¹Department of Biostatistics, University of Florida, Gainesville, FL 32611

²Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599-7420

³Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032

Summary

Precise modeling of disease progression in neurodegenerative disorders may enable early intervention before clinical manifestation of a disease, which is crucial since early intervention at the premanifest stage is expected to be more effective. Neuroimaging biomarkers are indicative of the underlying disease pathology and may be used to predict future disease occurrence at the premanifest stage. As observed in many pivotal studies, longitudinal measurements of clinical outcomes, such as motor or cognitive symptoms, often present nonlinear sigmoid shapes over time, where the inflection points of the trajectories mark a meaningful time in disease progression. Therefore, to identify neuroimaging biomarkers predicting disease progression, we propose a nonlinear mixed effects model based on a sigmoid function to predict longitudinal clinical outcomes, and associate a linear combination of neuroimaging biomarkers with subject-specific inflection points. Based on an expectation-maximization (EM) algorithm, we propose a method that can fit a nonlinear model with many potentially correlated biomarkers for random inflection points while achieving computational stability. Variable selection is introduced in the algorithm in order to identify important biomarkers of disease progression and to reduce prediction variability. We apply the proposed method to the data from the Predictors of Huntington's Disease study to select brain subcortical regional volumes predictive of the inflection points of the motor and cognitive function trajectories. Our results reveal that brain atrophy in the striatum and expansion of the ventricular system are highly predictive of the inflection points. Furthermore, these inflection points may precede clinically defined disease onset by as early as a decade and thus may be useful biomarkers as early signs of Huntington's Disease onset.

Keywords

EM algorithm; inflection point; neuroimaging biomarkers; nonlinear mixed model; sigmoid function; variable selection

Supplementary Materials

Web Appendix, Tables, and Figures referenced in Sections 2, 3, 4, and 5 are available with this paper at the *Biometrics* website on Wiley Online Library.

1. Introduction

For many neurological disorders, including Huntington's disease (HD) and Alzheimer's disease, no biomarker exists to objectively and fully determine disease onset age (Reilmann et al., 2014; Biglan et al., 2013). Thus, the clinically defined disease diagnosis for these disorders relies on measurements of clinical symptoms, which may not be present in patients in the early stages of disease. Because of this fact, recent efforts have focused on identifying association between temporal clinical measurements and biomarkers that may reflect the underlying disease pathology and progression, with the eventual goal of early intervention (Dubois et al., 2014). The existing literature suggests that longitudinal measurements of many clinical symptoms take a nonlinear sigmoid shape (Jedynak et al., 2012; Samtani et al., 2012), in which there is an early phase of gradual change followed by a period of acceleration around the time of clinical diagnosis. For example, Paulsen et al. (2014) demonstrated that motor and cognitive outcomes of premanifest HD subjects showed a sigmoid degeneration shape over time. Nonlinear models have also been discussed by the European Medicines Agency (EMA) and the U.S. Food and Drug Administration (FDA) as clinical trial design tools for studying neurodegenerative disorders (Romero et al., 2015).

Existing literature on modeling progression of neurodegenerative disorders mostly describes the population average trend, which does not associate model parameters with subject-specific biomarkers, and thus limiting its applications for predicting progression in individuals. In practice, many biomarkers may be predictive of disease progression in individual subjects. For example, prognostic biomarkers, including neuroimaging measures, are suggested to be useful in predicting early motor or cognitive abnormalities in HD (Paulsen et al., 2014). Some of the most promising imaging biomarkers include volumetric measures derived from brain structural magnetic resonance imaging (MRI) (Paulsen et al., 2014). An important goal of current research on HD is to identify potentially useful biomarkers to predict disease progression as measured by repeated clinical measurements over time, which typically follow a nonlinear trajectory. The ability to predict these nonlinear trajectories using prodromal biomarkers will enable early disease intervention and prevention.

In this paper, we develop methods to model nonlinear disease progression using a nonlinear mixed effects model (NLMEM) with random inflection points. We model the long-term trajectory using a sigmoid function indexed by four parameters, namely, the minimal value, range, disease progression rate, and inflection point for the rate of progression. The inflection point is then modeled as a linear function of the baseline biomarkers. The inflection point is particularly important as it marks the critical time of maximal rate of deterioration, after which the disease progression continues but no longer accelerates. With Alzheimer's disease, empirical evidence shows that the acceleration of cognitive decline can begin months or even years before a clinical diagnosis is made (Amieva et al., 2005; Grober et al., 2008) and the inflection point where the rate of progression stops increasing may be a crucial landmark in disease abnormality (Jedynak et al., 2012). For HD, our preliminary analysis in Section 2 shows that these inflection points may occur earlier than clinical diagnosis and are correlated with diagnosis. Thus they may be used in clinical trial

recruitment to maximize the chance of observing a diagnosis during the trial and thus to improve the power.

The main contribution of our work is to propose a penalized NLMEM for identifying important biomarkers for disease progression characterized by the inflection points in the model. Although NLMEMs have been extensively studied in the literature (e.g., Lindstrom and Bates, 1990; Wolfinger, 1993; Davidian and Giltinan, 1995; Pinheiro and Bates, 1995, 2000), existing methods are restricted to fitting models with a small number of covariates, and the computational convergence may strongly depend on the choice of initial values. For high-dimensional settings, although variable selection has been considered for linear mixed effects model (Bondell et al., 2010; Fan and Li, 2012), there is little work for NLMEMs. An exception is Ke and Wang (2001), who proposed a semiparametric nonlinear mixed effects model together with Laplace approximations to the marginal likelihood and penalized marginal likelihood for estimation. Wang and Leng (2012) proposed an unified method using least squares approximation to the marginal likelihood; Wierzbicki et al. (2014) adopted a sparse semiparametric nonlinear marginal model for chromatographic fingerprint data. However, the parameters subject to variable selection are for fixed effects, while in our case the parameters being selected are for random effects and thus will enter the covariance matrix in the marginal model, rendering the methods in Wang and Leng (2012) and Wierzbicki et al. (2014) non-applicable.

In this paper, we simultaneously account for multiple biomarkers predicting random inflection points via penalization and propose an expectation-maximization (EM) algorithm-based approach for estimation, where initial values are chosen using a simple two-stage procedure. Due to the challenges in computing the integral over the random effects and optimizing the nonlinear objective function, directly fitting a nonlinear mixed effects model with a large number of predictors for the random effects leads to either unstable estimates or non-convergence, especially when the covariates for random effects may be correlated. To overcome these challenges, we transform a difficult nonlinear optimization with multivariate predictors to an easy problem of penalized least squares. Since the convergence of an NLMEM procedure relies heavily on the choice of initial values, we propose a simple and effective two-stage approach to obtain initial values. We show that the EM-algorithm based estimate is consistent and enjoys the oracle property for variable selection. We apply the method to our motivating example to identify biomarkers associated with inflection points and discuss implications on HD clinical trial planning.

2. The Motivating PREDICT-HD Study

HD is a monogenic disorder caused by the expansion of trinucleotide cytosine-adenine-guanine (CAG) in the *HTT* gene (MacDonald et al., 1993). Subjects with an expanded CAG repeats (> 36) in the huntingtin gene will eventually develop HD, but age at HD onset is not fully determined by CAG expansion status. The PREDICT-HD study (Paulsen et al., 2008) is a 32-site observational study to better characterize the natural history of HD so that preventive intervention can be initiated at the best time to maximize success. The follow-up period for each individual was up to 10 years and annual or biennial measurements were taken for the variables in important domains of HD phenotypes including motor, cognitive

and behavioral symptoms. The PREDICT-HD data we analyzed include 1,222 subjects with 278 subjects without CAG repeats expansion, 915 premanifest expanded subjects (CAG repeats ≥ 36) without HD diagnosis at the baseline, and 29 patients clinically confirmed to have an HD diagnosis at the baseline.

Since the CAG repeats length is inversely correlated with the age at onset of HD (Langbehn et al., 2004), to obtain less heterogeneous subgroups, we divide the sample into five groups according to a subject's CAG repeats length. Specifically, group 1 are non-expanded controls with CAG ≤ 35 who will not be affected by HD, and groups 2–5 are expanded subjects with CAG in [36, 40), [40, 42], (42, 46], and ≥ 46 , respectively. The first two groups are well defined in the literature (Langbehn et al., 2004), i.e., non-expanded controls and mild penetrated subjects. The last three groups are defined based on the sample 50th, 90th quantiles of those with CAG ≥ 40 (full penetration). Consequently, there are 278, 74, 456, 352, and 62 subjects in groups 1–5, respectively. Sensitivity analyses using other values to divide groups are presented in Supplementary Materials. Three longitudinal clinical outcomes for disease progression are of interest including a motor sign outcome, the total motor score (TMS); and two cognitive outcomes, the Stroop Word Test and the Symbol Digit Modalities Test (SDMT) (Kremer et al., 1996). The TMS is a summary of impairments of motor functions in ocular pursuit, tongue, finger, hands, arms, gait and so on (Kremer et al., 1996). The Stroop Word Test and SDMT are used as tools for evaluating cognitive executive functions.

As a preliminary analysis to examine the temporal patterns of the three disease progression outcomes, we fit locally weighted scatterplot smoothing (LOWESS) curves and sigmoid models without using any imaging biomarkers. We plot in Figures 1 and Web-2 the estimated sigmoid curves (the bold solid lines) superimposed on nonparametric LOWESS curves and stratified by CAG groups. Here we modeled the log-transformed total motor score, i.e., $\log(\text{TMS}+1)$, instead of the original scale. For groups 2, 3, and 4, the sigmoid model fits show good approximations to the nonparametric LOWESS fits. Some deviations are observed in group 5, which may be due to fewer available subjects and thus the LOWESS fit is highly variable. Thus, empirical observations are consistent with suggestions in the existing clinical literature (Jedynak et al., 2012; Wang et al., 2014; Samtani et al., 2012; Romero et al., 2015) that a sigmoid function is appropriate to model disease progression. Additionally, to study the relationship between the inflection points and clinical diagnosis, we use the vertical lines to indicate the average position of the inflection points obtained from each individual trajectories. The inflection point represents the age at maximal rate of deterioration. Since information on new HD diagnosis during the study is also available, we use bullet points in groups 2–5 to represent premanifest gene expanded subjects who did not have a clinical diagnosis but had HD gene mutation, diamond points for subjects who developed HD during the follow-up period, and star points for subject who developed HD before the baseline. The plots show a close inverse relationship between the inflection point lines and CAG repeats length, where a longer repeats (higher disease burden) is associated with an earlier inflection point. Figures 1 and Web-2 also show that the average inflection points appear to precede the average diagnosis age for those who developed HD during the follow-up (with an average diagnosis age of 47.0, $SD=9.8$). In this study, brain structural magnetic resonance imaging (MRI) volumetric measures were also

collected at baseline for 1,107 subjects. In Section 5, we further examine the relationship between individual inflection points, neuroimaging biomarkers, and the likelihood of developing HD.

3. Methodologies

3.1 NLMEM for disease progression

Let Y_{ij} be the longitudinal outcome (e.g., motor or cognitive outcomes in PREDICT-HD) measuring disease progression, and let Z_{ij} be the age for subject i at visit j for $i = 1, \dots, n$ and $j = 1, \dots, m_i$. Let X_i be a vector of p -dimensional neuroimaging biomarkers measured at baseline, among which are predictive of the subject-specific parameters. Based on the preliminary analysis from PREDICT-HD study (Section 2) and existing literature (Samtani et al., 2012; Wang et al., 2014), we specify each subject's individual trajectory as a sigmoid function and allow its inflection point to vary with baseline biomarkers X_i . Specifically, we assume

$$Y_{ij} = \theta_1 + \nu_{i1} + \theta_2 \{1 + \exp[-\theta_3(Z_{ij} - \tau_i)]\}^{-1} + \varepsilon_{ij}, \quad (1)$$

$$\tau_i = \theta_4 + X_i^T \beta + \nu_{i2}, \quad (\nu_{i1}, \nu_{i2}) \sim N(0, \Sigma_r), \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2).$$

Here, τ_i is a subject-specific random inflection point in the sigmoid trajectory, θ_1 is the minimum value, θ_2 is a scale parameter, θ_3 is the average rate of change, θ_4 is the population average inflection point, and ν_{i1} is a random intercept for subject i and is allowed to be correlated with the random effect ν_{i2} for the inflection point. Because the random inflection point depends on biomarkers X_i , the proposed model personalizes the disease degeneration trajectory according to each subject's biomarkers and accounts for the residual variability across subjects via the random effects. The inflection point τ_i captures the age when the symptom acceleration changes sign, which is a potentially important landmark of the disease process. Our goals are: (1) to estimate the average rate of progression and other parameters; to identify which neuroimaging biomarkers in X_i are informative of inflection points; and to assess the effects of all biomarkers X_i as a whole. Specifically, the second goal relates to variable selection for τ_i , while the last goal concerns the inference of the estimated β or the test $H_0: \beta = 0$.

3.2 Algorithm for the penalized NLMEM and theoretical justification

Denote by γ all parameters $(\beta^T, \theta^T, \xi^T)^T$, where $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)^T$ and ξ includes the parameters σ_ε^2 and Σ_r . Let $O_i = (X_i, Z_{ij}, Y_{ij}, j = 1, \dots, m_i)$ be the observed data and define $b_i = (\nu_{i1}, \tau_i - \theta_4)$ and

$$f(Z_{ij}, b_i; \theta) = \theta_1 + \nu_{i1} + \theta_2 \{1 + \exp[-\theta_3(Z_{ij} - \tau_i)]\}^{-1}.$$

We propose the following EM algorithm by treating b_i as missing data.

Note that the complete log-likelihood of the full data is

$$l_F(O_i, b_i) = -\frac{m_i}{2} \log \sigma_\varepsilon^2 - \sum_{j=1}^{m_i} \frac{[Y_{ij} - f(Z_{ij}, b_i; \theta)]^2}{2\sigma_\varepsilon^2} - \frac{1}{2} \log |\sum_r| - \frac{(b_i - (0, X_i^T \beta)^T)^T \sum_r^{-1} (b_i - (0, X_i^T \beta)^T)}{2}.$$

To fit the model with a large number of predictors X_i and perform variable selection, one approach is to maximize the penalized log-likelihood for the observed data

$$\sum_{i=1}^n \log [\int \exp \{l_F(O_i, b_i)\} db_i] + \phi_\lambda(|\beta|),$$

where $\phi_\lambda(\cdot)$ is a penalty function. However, direct maximization of the above penalized likelihood is a highly nonlinear optimization problem and computationally challenging. Instead, we propose an EM algorithm for estimation that transforms variable selection for nonlinear models into a much easier optimization of penalized least squares.

Specifically, in the E-step, we compute the conditional expectation of the log-likelihood evaluated at the current estimate of the parameters γ with respect to b_i given the observed data O_i . In the M-step we compute parameters γ by maximizing the expected log-likelihood in the previous E-step. After the E- and M-steps are iterated until convergence, we perform variable selection for the biomarkers using penalized generalized least squares. A major advantage of the EM algorithm is that the challenge presented by the large number of predictors X_i only appears in the generalized least squares estimation in the M-step, which is a much easier optimization problem. We give the details in the following steps.

Step 1 (EM algorithm). Treating b_i as missing data, the E-step is to evaluate the posterior expectation of some function of b_i given the observed data $O_i = (X_i, Z_{ij}, Y_{ij}, j = 1, \dots, m_i)$ as

$$\hat{E}[g(b_i)] = \frac{\int g(b_i) \exp\{-\sum_{j=1}^{m_i} R_{ij}^2 / (2\sigma_\varepsilon^2) - (b_i - (0, X_i^T \beta)^T)^T \sum_r^{-1} (b_i - (0, X_i^T \beta)^T) / 2\} db_i}{\int \exp\{-\sum_{j=1}^{m_i} R_{ij}^2 / (2\sigma_\varepsilon^2) - (b_i - (0, X_i^T \beta)^T)^T \sum_r^{-1} (b_i - (0, X_i^T \beta)^T) / 2\} db_i},$$

where g is a generic function of b_i and $R_{ij} = Y_{ij} - f(Z_{ij}, b_i; \theta)$. This calculation can be carried out using numerical integration method such as adaptive Gaussian quadratures (Pinheiro and Bates, 1995). In the M-step, we update parameters θ using one-step Newton-Raphson algorithm

$$\theta^{(l+1)} = \theta^{(l)} - \left[\sum_{i=1}^n \sum_{j=1}^{m_i} \hat{E}(R_{ij} \nabla_{\theta\theta}^2 f) - \sum_{i=1}^n \sum_{j=1}^{m_i} \hat{E}(\nabla_{\theta} f \nabla_{\theta} f^T) \right]^{-1} \times \left[\sum_{i=1}^n \sum_{j=1}^{m_i} \hat{E}\{\nabla_{\theta} f R_{ij}\} \right],$$

where $\nabla_{\theta} f = \nabla_{\theta} f(Z_{ij}, b_i; \theta)$ and $\nabla_{\theta\theta}^2 f = \nabla_{\theta\theta}^2 f(Z_{ij}, b_i; \theta)$ are the first and second derivatives, respectively, of f with respect to θ , and l indexes iteration. We update σ_ε^2 as

$$\left(\sum_{i=1}^n m_i\right)^{-1} \sum_{i=1}^n \sum_{j=1}^{m_i} \hat{E}[Y_{ij} - f(Z_{ij}, b_i; \theta^{(l+1)})]^2$$

and update β by minimizing

$$\sum_{i=1}^n \hat{E}[(b_i - (0, X_i^T \beta)^T)^T \sum_r^{-1} (b_i - (0, X_i^T \beta)^T)].$$

Under a special case when ν_{1l} and ν_{2l} are independent, we can update β by minimizing

$$\sum_{i=1}^n (\hat{E}[\tau_i] - \theta_4 - X_i^T \beta)^2,$$

since $\hat{E}[b_i] = (\hat{E}[\nu_{1l}], \hat{E}[\tau_i] - \theta_4)^T$. We further update Σ_r from the residual summed squares

$$\sum_r = n^{-1} \sum_{i=1}^n \hat{E}[(b_i - (0, X_i^T \beta)^T)(b_i - (0, X_i^T \beta)^T)^T].$$

As a note, when the dimension of X_i is high, it may be more accurate to replace n by $(n - p)$ in the above expression using a restricted maximum likelihood estimate, where p is the dimension of β .

Step 2 (variable selection). We perform variable selection after the EM algorithm converges. We apply the adaptive LASSO (Zou, 2006) to the following penalized generalized least squares

$$\sum_{i=1}^n \hat{E}[(b_i - (0, X_i^T \beta)^T)^T \sum_r^{-1} (b_i - (0, X_i^T \beta)^T)] + n \sum_{j=1}^p \phi_{\lambda_j}(|\beta_j|),$$

where p is the dimension of β , and $\phi_{\lambda_j}(|\beta_j|) = \lambda_j |\beta_j|$ is the adaptive LASSO penalty with $\lambda_j = \lambda / \hat{\beta}_j^{-1}$ and $\hat{\beta}_j$ the estimator from step 1. Particularly, this variable selection is implemented using the R glmnet package (Friedman et al., 2009).

Step 3 (post-selection inference). We apply the EM procedure again but using the selected subset of X_i , say, X_i^0 . From the EM procedure, we can estimate the effect of X_i^0 and its asymptotic covariance via the Louis formula (Louis, 1982). To apply it, we consider both the parameters of interest β , θ and the nuisance parameters ξ . The Louis formula (Louis, 1982) for the information calculation is given as

$$-\sum_{i=1}^n \hat{E}[\nabla_{\gamma}^2 l_F(O_i, b_i)] - \sum_{i=1}^n \left\{ \hat{E}[\nabla_{\gamma} l_F(O_i, b_i) \nabla_{\gamma} l_F(O_i, b_i)^T] - \left(\hat{E}[\nabla_{\gamma} l_F(O_i, b_i)] \hat{E}[\nabla_{\gamma} l_F(O_i, b_i)]^T \right) \right\}.$$

The inverse of this information matrix gives the asymptotic covariance for $\hat{\gamma}$.

Note the advantage of the proposed EM algorithm is that the challenge due to variable selection of X_i only appears in the penalized least squares estimation in Step 2, which is a much easier optimization problem. Thus, the algorithm avoids direct optimization of the penalized observed data likelihood for the proposed nonlinear models. In addition, the proposed method can be easily generalized to other nonlinear models, where subject-specific parameters b_i depend on a large number of predictors.

Finally, as in all NLMEM procedures, initial values play a crucial role to ensure convergence to a global maximum. To improve the convergence property, we further propose the following simple two-stage approach to select the initial values in the algorithm: at stage 1, we first fit an NLMEM without the biomarkers X_i in order to obtain the subject-specific random effects as $\hat{b}_i$ using the best linear unbiased estimator (BLUE); we then estimate β by the following penalized least square

$$\hat{\beta}^0 = \arg \min_{\beta^0 \in R^p} \sum_{i=1}^n (\hat{\tau}_i - \theta_4 - X_i^T \beta^0)^2 + n \sum_{j=1}^p \phi_{\lambda_j}(|\beta_j^0|), \quad (2)$$

where $\hat{b}_i = (\hat{\nu}_1, \hat{\tau}_i - \theta_4)$. In stage 2, we fit another NLMEM based on model (1) to estimate the other parameters by fixing $\beta = \hat{\beta}^0$. Essentially, a two-stage approach fits an NLMEM in the first stage, ignoring biomarkers X_i as predictors, and attempts to recover their effects using the predicted random effects. Such a procedure is easy to implement, but one can show that it leads to a biased estimate of β even when using a penalty function (e.g., adaptive LASSO, Zou, 2006) that guarantees consistency due to the shrinkage effect of the BLUE for b_i in the first step (see Web Appendix A).

We justify the proposed algorithm by proving its asymptotic properties in Web Appendix A. Specifically, in Theorem 1 (see Web Appendix A), we show that for a sequence of λ_n satisfying the conditions given in the appendix, the proposed estimator for $\hat{\gamma}_{\lambda_n}$ is $\sqrt{n}$ -consistent. Furthermore, $\hat{\beta}_{\lambda_n}$ has an oracle property that the non-zero component of β_0 can be identified with probability tending to one and it is asymptotically normally distributed. The proof follows the same arguments given in Johnson et al. (2008) and Liu and Zeng (2013). We provide a sketch of the proof in the Appendix. Finally, from the oracle property and the asymptotic normality in Theorem 1, it is clear that the estimator obtained in the post-variable selection Step 3 should also be asymptotically normal.

To predict inflection points for new patients, two methods can be considered. 1) Include the outcomes Y_{i0} observed at the baseline from the new subjects to re-fit the NLME, and use the posterior mean of the inflection point, $E(\tau_i | Y_{i0}, X_i, Z_i)$, as the predicted inflection point; 2) a more computationally efficient way is to use the fixed effects, $\hat{\theta}_4 + X_i^T \hat{\beta}$, as the predicted inflection point.

4. Simulation Studies

We performed simulation studies to investigate the performance of the following methods: (1) the proposed EM algorithm-based NLME with/without adaptive LASSO penalty (NLME-EM ALASSO/NLME-EM); (2) the two-stage nonlinear mixed effects model with adaptive LASSO penalty (NLME-2 ALASSO); and (3) NLMEM assuming known informative biomarkers (NLME-T) with the LME approximation algorithm implemented in the R pack-age *nlme*. For a fair comparison, we also include a third step in the two-stage method similar as the Step 3 of NLME-EM. We evaluated three properties: 1) the estimation accuracy of the population-level parameters and individual inflection points; 2) the variable selection performance of the penalized procedure for choosing informative neuroimaging variables; and 3) the type I error rate and power of the test for detecting the overall effect of the neuroimaging variables.

We generated X_i from $N(0, \Omega)$, where $p = 30$ or $p = 100$, and Ω has either a compound symmetric structure with between-marker correlation of $\rho = 0.1$ or 0.3 or an autoregressive correlation structure with $\rho = 0.7$. The baseline age Z_{i1} was generated from $N(40, 6^2)$. The individual inflection point τ_i given the biomarkers X_i was generated from

$\tau_i | X_i \sim N(X_i \beta + \theta_4, \sigma_\tau^2)$, where $(\beta_1, \beta_2, \beta_3, \dots, \beta_{30}) = (2, -1.5, 0, \dots, 0)$, and $\sigma_\tau = \sqrt{2}, 2$, or 3 . Finally, the longitudinal outcomes Y_{ij} were generated from the following sigmoid model

$$Y_{ij} = \theta_1 + \nu_{i1} + \theta_2 \{1 + \exp[-\theta_3(Z_{ij} - \tau_i)]\}^{-1} + \varepsilon_{ij},$$

$$\nu_{i1} \sim N(0, 5^2), \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \quad i = 1, \dots, n, \quad j = 1, \dots, m,$$

where $\sigma_\varepsilon = 10$, $(\theta_1, \theta_2, \theta_3, \theta_4) = (0, 100, 0.2, 45)$, and ν_{i1} is independent of τ_i . In the simulation study, we considered the number of subjects to be $n = 300$ and the number of the observations per subject to be $m = 3, 4$, or 5 .

Tables 1, Web-1, and 2 summarize the results for parameter estimation and variable selection properties using 200 replications. Table 1 shows that the mean absolute errors (MAE) of the estimated shape parameters $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4)$ are slightly higher for NLME-2 ALASSO compared to NLME-EM ALASSO. Moreover, MAE of the estimated coefficients of the biomarkers (e.g., $\hat{\beta}$) are larger for the NLME-2 ALASSO compared to NLME-EM ALASSO, especially when the variance of the random inflection points is large.

To evaluate the accuracy of the estimated individual inflection points, in Table Web-1, we report the mean absolute error (MAE) and correlation (COR) between the estimated inflection point $\hat{\tau}_i$ and the true τ_i . Table Web-1 gives the results based on NLME-2, NLME-2 ALASSO, NLME-EM, NLME-EM ALASSO methods, as well as NLME-T. From this table, we observe that the MAEs of NLME-EM ALASSO are lower than those of NLME-2 ALASSO, while the correlations with the true τ_i of the former are higher. We also see that the MAEs of NLME-EM ALASSO are lower than those of NLME-EM. Thus through variable selection the estimation accuracy for the inflection point is improved. NLME-EM

ALASSO is comparable to NLME-T, which incorporates the known informative biomarkers in the model.

In terms of variable selection, in Table 2, we compare the number of correctly and incorrectly selected non-null variables and the proportion of over-fit, correct-fit, and under-fit models. Both NLME-EM ALASSO and NLME-2 ALASSO appropriately select the informative biomarkers with large numbers of observations for each subject and moderate variability of τ_i . However, when σ_d^2 increases and the data are sparse ($m = 3, 4$), NLME-EM ALASSO outperforms NLME-2 ALASSO with a higher percentage of correct-fit and a lower percentage of under-fit models.

Finally, we investigated the performance of the Wald test based on NLME-EM and the F -test based on the first-step of NLME-2 without variable selection for the overall effect of the biomarkers. The type I error rate was obtained under $\beta = 0$, and the power was obtained under $(\beta_1, \beta_2, \beta_3, \dots, \beta_{30}) = \delta * (2, -1.5, 0, \dots, 0)$, where $\delta = 0.1, 0.2, 0.3, 0.4$, or 0.5 . The other settings were the same as in the previous simulations. The type I error rate was computed using 1000 simulations, and the power was computed using 400 simulations. The estimated type I error rates for the F -test are 0.057 and 0.056 for the two cases, respectively, which are close to the nominal level of 0.05. Figure Web-1 summarizes the power as a function of δ . We observe that both tests show reasonable power for detecting the effect of the biomarkers under the alternative hypothesis, and NLME-EM-based Wald test is more powerful than NLME-2-based F -test. A possible reason for this finding is that because of the shrinkage effect of BLUE for $\hat{b}_i$, the first-step of NLME-2 underestimates β , and therefore, the power of the F -test is lower compared to NLME-EM, which is a consistent approach. We thus demonstrate that the proposed NLME-EM provides much greater power than NLME-2 for detecting biomarker effects.

5. Application to PREDICT-HD Study

We applied the proposed methods to the PREDICT-HD study introduced in Section 2. In PREDICT-HD, the regional and global measures of volumes at cortical and subcortical regions of interest (ROIs) were obtained in neuroanatomical atlas, after preprocessing structural MRIs using Freesurfer 5.2 (<http://surfer.nmr.mgh.harvard.edu>; see, Paulsen et al., 2014). Previous clinical evidence (Paulsen et al., 2014) indicated a strong association between some of these biomarkers and HD age-at-onset. Furthermore, it has been shown that atrophy in the striatum and expansion of the lateral ventricles may manifest as early as 11 years before the estimated onset of HD (Walker, 2007). However, no work has been done to select which biomarkers are most predictive of the disease progression as measured by motor or cognitive outcomes in multivariate models, which is the main goal of this work.

First, we fit the sigmoid NLMEM without using the structural MRI regional brain volumes from a joint NLME model (Web-1) of three outcomes reflecting motor and cognitive impairment of HD. Because we are interested in disease progression, we fit separate models for the non-expanded control subjects (the first group in Figures 1 and Web-2) and the expanded subjects who carry the HD mutation. Figures 1 and Web-2 show the estimated sigmoid curves stratified by the CAG group for the log-transformed total motor score (TMS)

and two cognitive outcomes (Stroop Word Test and SDMT). The sigmoidal model fits data adequately as shown in the residual plots in Figure Web-3 and goodness-of-fit test. To check the independence assumption, we computed the within subject one-lag correlations of the errors which showed negligible correlations -0.08 , -0.15 , and -0.15 for TMS, Stroop Word and SDMT, respectively.

Table 3 summarizes the estimated progression rates and the average inflection points. Interestingly, we see a monotone relationship between the average inflection points and the CAG group. Specifically, the mean inflection points are around 60, 50, 40, and 30 years of age, respectively, for the four CAG groups. Groups with higher CAG repeats length have earlier inflection points (i.e., the curve is shifted to the left), indicating earlier disease acceleration and faster disease progression. The magnitude of the rate of change (parameter θ_3) also increases with CAG repeats length. These results are consistent with the literature that the CAG repeats length is inversely correlated with the age at onset of HD (Langbehn et al., 2004) and positively correlated with disease burden (Langbehn et al., 2010). For subjects in group 2, cognitive decline is similar to that in the normal aging group 1 (see Figures 1 and Web-2). The last column of Table 3 presents the p -values for detecting difference in the inflection point between the three outcomes. In Group 3, the average inflection points differ significantly for the three outcomes, with Stroop Word Test showing the earliest sign of decline compared to TMS and SDMT (i.e., about 4.5 and 6.1 years earlier than TMS and SDMT, respectively). For the other three groups, the inflection points for the three outcomes are not significantly different.

Next, we considered the following model to incorporate the structural MRI biomarkers and applied the proposed NLME-EM with adaptive lasso penalty to cope with correlated brain regional volumes in the variable selection. For a subject in the g th group ($g = 2, \dots, 5$),

$$Y_{ij} = \theta_{1g} + \nu_{i1} + \theta_{2g} \{1 + \exp[-\theta_{3g}(Z_{ij} - \tau_i)]\}^{-1} + \varepsilon_{ij}, \quad (3)$$

$$\text{where } \tau_i = \theta_{4g} + X_i^T \beta + \nu_{i2}, \quad (\nu_{i1}, \nu_{i2}) \sim N(0, \sum_r), \quad \text{and } \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2),$$

where $(\theta_{1g}, \theta_{2g}, \theta_{3g}, \theta_{4g})$ are the sigmoid parameters for the g th group, and X_i is the baseline neuroimaging biomarkers as listed in Table Web-3. The regional volumes were first normalized by dividing by the total intracranial volume (ICV) and then standardized by subtracting the mean and dividing by the standard deviation of MRI regional volumes for the controls. We tested the overall effects of the 39 normalized regional volumes using the F -test based on the first step of the NLME-2 and the Wald test based on the NLME-EM. The p -values for all three outcomes are less than 10^{-15} ; thus, the neuroimaging variables have significant effects on predicting the inflection points for the three outcomes. The updated inflection points after including neuroimaging biomarkers are in general earlier than without using these markers (Table 3, footnotes), demonstrating the utility of using them as prodromal markers to identify subjects with earlier inflection points.

The coefficients for the selected brain regions are listed in Table Web-3, which can be interpreted as a shift in the inflection point due to 1 standard deviation change in the regional volumes. For example, 0.1 standard deviation unit atrophy in the *right pallidum* area is associated with about 4.5 months (0.37 years) earlier inflection point for the TMS. Additionally, Figure 2 shows three representative slices for part of the selected regions of interest (ROIs) coefficients for the TMS. Note that we did not assume that the left and right brain ROIs are symmetric. Although the left and right brain ROIs are highly correlated, they do not have exactly the same effect. Overall, we see that atrophy in subcortical regions and expansion of the ventricle system and vessel predict the inflection points. Different ROIs tend to predict inflection points for motor and cognitive outcomes, suggesting distinct underlying pathology for the two clinical domains of HD.

From Table Web-3, we see that the penalized NLME-EM selected subcortical ROIs *left and right pallidum*, *left amygdala*, *right putamen*, and *CC-Mid-posterior* as predictors of the inflection point for motor function. *Right thalamus and right hippocampus* were also selected, but with negative coefficients due to their strong correlation with other selected ROIs (the correlation values ρ range from 0.53 to 0.66). *Pallidum*, *putamen* and *thalamus* belong to the basal ganglia system, with their primary function as to control and regulate activities of the motor and premotor cortical areas so that voluntary movements can be performed smoothly (Chakravarthy et al., 2010). Additionally, the *3rd ventricle*, *optic chiasm*, *right vessel*, *right choroid plexus* are also selected, and all show considerable expansion. This may be because that the intracranial volume is less likely to change, space from the lost cortical and subcortical volumes may be filling with fluid to maintain the overall structural integrity of the organ (Seidler et al., 2010).

For the cognitive outcome, Stroop Word Test, penalized NLME-EM chooses the subcortical ROIs *right pallidum*, *left amygdala*, and *left cerebellum white matter*. The ROIs *right amygdala*, *brain stem*, *CC-central* and *CC-Anterior* are also chosen, but with negative coefficients due to high correlations with other selected ROIs (the correlation values range from 0.4 to 0.67). CC stands for *Corpus Callosum*, which is the largest white matter that connects the left and right cerebral hemispheres to enable interhemispheric communication. *Cerebellum* plays an important role in motor control and is involved in some cognitive functions such as attention and language. For the cognitive measure SDMT, slightly different ROIs were chosen. The expansion of the ventricle system and vessel also significantly predict the inflection point for both the Stroop Word Test and SDMT.

To evaluate the predictive performance of the estimated inflection points, we computed their correlation with the observed age at onset of HD and lead time (the difference between the age at onset of HD and the inflection point). Group 5 with the highest CAG repeats tends to show the earliest inflection points. The correlations between the inflection points and age at onset of HD are 0.49 (95% bootstrap CI [0.32, 0.65]) for TMS, 0.45 (95% bootstrap CI [0.28, 0.60]) for Stroop Word, and 0.42 (95% bootstrap CI [0.27, 0.58]) for SDMT. The associations are shown in Figure 3. It can be seen that the majority of subjects with clinically confirmed HD during the follow-up period have positive lead times, and thus they manifest early signs of degeneration in the brain. The inflection points predicted from the neuroimaging biomarkers are much earlier than clinically defined HD onset, with mean lead

times of 8.8, 9.5, and 10.4 years for TMS, Stroop word test, and SDMT, respectively. Thus, neuroimaging measures are useful biomarkers for identifying subjects with earlier disease progression. Due to biological variation and measurement errors, the estimated inflection points may not be uniformly earlier than the diagnosis age in all subjects. Subjects with negative lead times tend to have early HD onset (< 50) but shorter instead of longer CAG repeats length. Thus, these patients may have a slightly different pathology than a typical patient and other prognostic factors not captured in brain atrophy measures may have contributed to disease progression. In addition, we used the Cox model to check the overall prediction performance. Figure Web-4 in the Supplementary Materials shows a negative association between the hazard rate of HD onset and the neuroimaging scores.

Lastly, in Figure Web-5, we superimposed fit individual sigmoid trajectories from six subjects on their observed measurements of the TMS. The six subjects are from groups 3, 4, and 5, respectively. The vertical dotted lines correspond to the individual inflection points. The first red dot of each subject corresponds to his/her age when HD was clinically diagnosed. We observe that the sigmoid trajectories fit the observed individual trends adequately, and the inflection point tends to be earlier than the age at diagnosis of HD.

6. Discussion

In this work, we propose NLMEMs for inferring subject-specific random inflection points as measures of disease progression. To incorporate many potentially correlated neuroimaging biomarkers, we developed an EM algorithm-based approach to overcome the computational challenges, stabilize estimation, and achieve variable selection. The proposed method simultaneously select the informative neuroimaging biomarkers and estimate individual inflection points based on the NLMEM. In our method, variable selection reduces to a penalized least squares form, which has a clear computational advantage. A simple two-stage approach is proposed to provide initial values for the EM algorithm, which reduces the number of iterations and improves computational convergence rate and performance of the algorithm. We apply the proposed methods to the PREDICT-HD study, and find that the inflection points tend to occur earlier than the age of HD onset defined clinically (mainly based on the motor impairment symptoms). Moreover, the use of MRI imaging biomarkers as predictors of individualized inflection point allows obtaining earlier detection of inflection points in subjects not yet diagnosed with HD.

Here, the effect of the structural variation in the *HTT* gene was modeled by stratification according to the CAG repeats length. An interesting future extension is to replace the stratification by a nonparametric piecewise linear or piecewise constant model. Another extension is to introduce different sparseness for neuroimaging variable effects $\beta(CAG)$ in different CAG stratification groups and perform group-LASSO for variable selection. A more flexible model will allow parameters in the nonlinear function to depend on subject-specific covariates, but may be more computationally challenging. Lastly, although our method is presented for sigmoid nonlinear function, which was motivated by the PREDICT-HD study data, it is not restricted to sigmoid nonlinear function. Other nonlinear models can be considered under the proposed framework.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to acknowledge the National Institute of Health Grants NS073671 and NS082062, the NIH dbGap data repository (phs000222.v3.p2) and the PREDICT-HD investigators.

References

- Amieva H, Jacqmin-Gadda H, Orgogozo J-M, Le Carret N, Helmer C, Letenneur L, et al. The 9 year cognitive decline before dementia of the Alzheimer type: a prospective population-based study. *Brain*. 2005; 128:1093–1101. [PubMed: 15774508]
- Biglan KM, Zhang Y, Long JD, Geschwind M, Kang GA, Killoran A, et al. Refining the diagnosis of Huntington disease: the PREDICT-HD study. *Frontiers in Aging Neuroscience*. 2013; 5
- Bondell HD, Krishna A, Ghosh SK. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*. 2010; 66:1069–1077. [PubMed: 20163404]
- Chakravarthy VS, Joseph D, Bapi RS. What do the basal ganglia do? A modeling perspective. *Biological Cybernetics*. 2010; 103:237–253. [PubMed: 20644953]
- Davidian, M., Giltinan, DM. *Nonlinear models for repeated measurement data*. Vol. 62. CRC Press; Boca Raton, Florida: 1995.
- Dubois B, Feldman HH, Jacova C, Hampel H, Molinuevo JL, Blennow K, et al. Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. *The Lancet Neurology*. 2014; 13:614–629. [PubMed: 24849862]
- Fan Y, Li R. Variable selection in linear mixed effects models. *Annals of Statistics*. 2012; 40:2043–2068. [PubMed: 24850975]
- Friedman J, Hastie T, Tibshirani R. *glmnet: Lasso and elastic-net regularized generalized linear models*. R package version. 2009; 1
- Grober E, Hall CB, Lipton RB, Zonderman AB, Resnick SM, Kawas C. Memory impairment, executive dysfunction, and intellectual decline in preclinical Alzheimer's disease. *Journal of the International Neuropsychological Society*. 2008; 14:266–278. [PubMed: 18282324]
- Jedynak BM, Lang A, Liu B, Katz E, Zhang Y, Wyman BT, et al. A computational neurodegenerative disease progression score: Method and results with the Alzheimer's disease neuroimaging initiative cohort. *NeuroImage*. 2012; 63:1478–1486. [PubMed: 22885136]
- Johnson BA, Lin D, Zeng D. Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association*. 2008; 103:672–680. [PubMed: 20376193]
- Ke C, Wang Y. Semiparametric nonlinear mixed-effects models and their applications. *Journal of the American Statistical Association*. 2001; 96:1272–1298.
- Kremer H, Group HS, et al. Unified Huntington's disease rating scale: reliability and consistency. *Movement Disorders*. 1996; 11:136–142. [PubMed: 8684382]
- Langbehn D, Brinkman R, Falush D, Paulsen J, Hayden M. A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clinical Genetics*. 2004; 65:267–277. [PubMed: 15025718]
- Langbehn DR, Hayden MR, Paulsen JS. CAG-repeat length and the age of onset in Huntington disease (HD): a review and validation study of statistical approaches. *American Journal of Medical Genetics Part B. Neuropsychiatric Genetics*. 2010; 153:397–408.
- Lindstrom M, Bates D. Nonlinear mixed effects models for repeated measures data. *Biometrics*. 1990; 46:673. [PubMed: 2242409]
- Liu X, Zeng D. Variable selection in semiparametric transformation models for right-censored data. *Biometrika*. 2013; 100:859–876.

- Louis TA. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B.* 1982; 44:226–233.
- MacDonald ME, Ambrose CM, Duyao MP, Myers RH, Lin C, et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell.* 1993; 72:971–983. [PubMed: 8458085]
- Paulsen J, Langbehn D, Stout J, Aylward E, Ross C, Nance M, et al. Detection of Huntington's disease decades before diagnosis: the Predict-HD study. *Journal of Neurology, Neurosurgery & Psychiatry.* 2008; 79:874–880.
- Paulsen JS, Long JD, Johnson HJ, Aylward EH, Ross CA, Williams JK, et al. Clinical and biomarker changes in premanifest Huntington disease show trial feasibility: a decade of the PREDICT-HD study. *Frontiers in Aging Neuroscience.* 2014; 6
- Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics.* 1995; 4:12–35.
- Pinheiro, JC., Bates, DM. *Mixed-effects models in S and S-PLUS.* Springer; Berlin, Germany: 2000.
- Reilmann R, Leavitt BR, Ross CA. Diagnostic criteria for Huntington's disease based on natural history. *Movement Disorders.* 2014; 29:1335–1341. [PubMed: 25164527]
- Romero K, Ito K, Rogers JA, Polhamus D, Qiu R, Stephenson D, et al. The future is now: Model-based clinical trial design for alzheimer's disease. *Clinical Pharmacology & Therapeutics.* 2015; 97:210–214. [PubMed: 25669145]
- Samtani MN, Farnum M, Lobanov V, Yang E, Raghavan N, DiBernardo A, et al. An improved model for disease progression in patients from the Alzheimer's disease neuroimaging initiative. *The Journal of Clinical Pharmacology.* 2012; 52:629–644. [PubMed: 21659625]
- Seidler RD, Bernard JA, Burutolu TB, Fling BW, Gordon MT, Gwin JT, et al. Motor control and aging: links to age-related brain structural, functional, and biochemical effects. *Neuroscience & Biobehavioral Reviews.* 2010; 34:721–733. [PubMed: 19850077]
- Walker FO. Huntington's disease. *The Lancet.* 2007; 369:218–228.
- Wang H, Leng C. Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association.* 2012; 102:1039–1048.
- Wang Y, Resnick SM, Davatzikos C. Analysis of spatio-temporal brain imaging patterns by hidden markov models and serial MRI images. *Human Brain Mapping.* 2014; 35:4777–4794. [PubMed: 24706564]
- Wierzbicki MR, Guo L-B, Du Q-T, Guo W. Sparse semiparametric nonlinear model with application to chromatographic fingerprints. *Journal of the American Statistical Association.* 2014; 109:1339–1349. [PubMed: 26388659]
- Wolfinger R. Laplace's approximation for nonlinear mixed models. *Biometrika.* 1993; 80:791–795.
- Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association.* 2006; 101:1418–1429.

Figure 1.

Descriptive plots of logarithm of motor (TMS) and cognitive outcome (Stroop Word) over age for subjects in the PREDICT-HD study. Bullet points in groups 2–5 represent premanifest gene expanded subjects who did not have a clinical diagnosis, diamond points represent subjects who developed HD during the follow-up period, and star points represent subject who developed HD before the baseline. The plot shows a close inverse relationship between the injection point lines and CAG repeats length, where a longer repeats (higher disease burden) associated with an earlier injection point. This figure appears in color in the electronic version of this article.

Figure 2.

Three representative slices for coefficients of part of the selected ROIs for predicting the injection point for the total motor scores. The MNI coordinates are $x = 0$, $y = 2$, $z = -7$. The green and red ROIs correspond to the left and right Pallidum; the blue ROI is the right Thalamus Proper, the yellow ROI is the right Putamen; and the dark blue ROI is the right choroid plexus.

Figure 3.

Association between lead time and age at onset of HD. Lead time is computed as observed age of HD onset minus injection points for total motor score (Top Left), stroop word test (Top Right), or SDMT (Bottom). This figure appears in color in the electronic version of this article.

Table 1

Average of mean absolute error (MAE) of estimated parameters and their standard deviations over 200 replications, $n = 300$, $p = 30$, $(\theta_1, \theta_2, \theta_3, \theta_4) = (0, 100, 0.2, 45)$, and $\beta = (2, -1.5, 0, \dots, 0)$.

NLME-2 ALASSO												
σ_ϵ	m	ρ	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
$\sqrt{2}$	5	0.1	2.14 (1.74)	4.81 (3.56)	0.01 (0.01)	0.36 (0.26)	0.52 (0.17)	0.61 (0.20)	1.95 (1.52)	4.47 (3.24)	0.01 (0.01)	0.34 (0.25)
	4	0.1	1.80 (1.47)	4.97 (4.14)	0.02 (0.01)	0.45 (0.32)	0.60 (0.18)	0.66 (0.20)	1.64 (1.31)	4.73 (3.78)	0.02 (0.01)	0.43 (0.32)
	3	0.1	2.28 (1.96)	6.99 (6.70)	0.02 (0.02)	0.53 (0.43)	0.68 (0.20)	0.75 (0.20)	2.15 (1.77)	6.47 (5.75)	0.02 (0.02)	0.51 (0.41)
$\sqrt{2}$	5	0.1	2.07 (1.40)	5.08 (3.89)	0.02 (0.02)	0.40 (0.30)	0.66 (0.22)	0.73 (0.22)	1.97 (1.33)	4.82 (3.63)	0.02 (0.01)	0.38 (0.30)
	4	0.1	2.34 (2.09)	5.64 (5.32)	0.02 (0.01)	0.40 (0.31)	0.72 (0.21)	0.82 (0.23)	2.08 (1.83)	4.99 (4.45)	0.02 (0.01)	0.38 (0.30)
	3	0.1	2.24 (2.02)	7.90 (7.96)	0.02 (0.02)	0.55 (0.53)	0.80 (0.25)	0.83 (0.23)	2.04 (1.73)	6.71 (6.17)	0.02 (0.02)	0.51 (0.49)
$\sqrt{2}$	5	0.1	2.87 (2.51)	5.96 (5.39)	0.03 (0.02)	0.45 (0.32)	0.88 (0.28)	0.89 (0.27)	2.34 (1.85)	4.97 (4.00)	0.02 (0.01)	0.39 (0.30)
	4	0.1	2.82 (2.19)	6.71 (5.51)	0.03 (0.02)	0.58 (0.37)	1.00 (0.28)	1.02 (0.28)	2.35 (1.78)	5.55 (4.53)	0.02 (0.02)	0.48 (0.34)
	3	0.1	2.45 (1.95)	5.88 (4.79)	0.06 (0.03)	0.86 (0.39)	1.11 (0.30)	1.06 (0.27)	2.28 (1.68)	5.60 (4.13)	0.04 (0.02)	0.60 (0.39)
$\sqrt{2}$	5	0.3	2.91 (2.44)	7.20 (6.44)	0.03 (0.02)	0.56 (0.35)	0.96 (0.26)	0.98 (0.28)	2.31 (1.89)	5.98 (5.05)	0.02 (0.02)	0.49 (0.33)
	4	0.3	3.02 (2.42)	6.99 (6.12)	0.03 (0.02)	0.69 (0.38)	1.04 (0.30)	1.07 (0.27)	2.36 (1.78)	5.60 (4.54)	0.02 (0.02)	0.55 (0.35)
	3	0.3	2.62 (2.08)	10.3 (10.0)	0.05 (0.03)	0.69 (0.49)	1.17 (0.31)	1.11 (0.25)	2.32 (1.66)	7.81 (6.17)	0.03 (0.02)	0.64 (0.43)
$\sqrt{2}$	5	0.7	1.81 (1.47)	4.68 (3.62)	0.02 (0.01)	0.32 (0.26)	0.89 (0.24)	0.91 (0.25)	1.80 (1.41)	4.53 (3.56)	0.01 (0.01)	0.32 (0.27)
	4	0.7	2.04 (1.67)	5.21 (4.08)	0.02 (0.01)	0.44 (0.32)	0.94 (0.24)	0.94 (0.24)	2.01 (1.62)	5.25 (4.16)	0.02 (0.01)	0.39 (0.30)
	3	0.7	2.07 (1.67)	5.71 (4.71)	0.02 (0.02)	0.53 (0.37)	1.09 (0.30)	1.06 (0.28)	2.09 (1.61)	5.78 (4.63)	0.02 (0.02)	0.46 (0.35)

NLME-2 ALASSO: two-stage NLME with adaptive LASSO penalty; NLME-EM ALASSO: EM algorithm-based penalized NLME with adaptive LASSO penalty. Number in the parenthesis is the standard deviation (SD).

Table 2

Variable selection for biomarkers, $n = 300$, $p = 30$, using 200 replications.

NLME-2 ALASSO										NLME-EM ALASSO									
σ_l	m	ρ	No. of non-zeros				Proportion of Correct-fit				No. of non-zeros				Proportion of Correct-fit				Under-fit
			C	IC	Over-fit	Under-fit	C	IC	Over-fit	Under-fit	C	IC	Over-fit	Under-fit					
$\sqrt{2}$	5	0.1	2.00	0.06	0.02	0.98	0.00	2.00	0.00	0.00	2.00	0.00	0.00	1.00	0.00	0.00			
	4	0.1	2.00	0.01	0.01	0.99	0.00	2.00	0.00	0.00	2.00	0.00	0.00	1.00	0.00	0.00			
	3	0.1	2.00	0.01	0.01	0.99	0.00	2.00	0.00	0.00	2.00	0.00	0.00	1.00	0.00	0.00			
$\sqrt{2}$	5	0.1	2.00	0.03	0.01	0.99	0.00	2.00	0.01	0.01	2.00	0.01	0.01	0.99	0.00	0.00			
	4	0.1	2.00	0.03	0.02	0.98	0.00	2.00	0.00	0.00	2.00	0.00	0.00	1.00	0.00	0.00			
	3	0.1	1.99	0.03	0.02	0.97	0.01	2.00	0.05	0.03	2.00	0.05	0.03	0.97	0.00	0.00			
$\sqrt{2}$	5	0.1	1.98	0.02	0.02	0.96	0.02	1.99	0.04	0.02	1.99	0.04	0.02	0.96	0.01	0.01			
	4	0.1	1.91	0.08	0.03	0.88	0.09	1.99	0.03	0.03	1.99	0.03	0.03	0.96	0.01	0.01			
	3	0.1	1.90	0.04	0.02	0.88	0.10	1.99	0.07	0.06	1.99	0.07	0.06	0.93	0.01	0.01			
$\sqrt{2}$	5	0.3	1.94	0.10	0.04	0.90	0.06	1.99	0.06	0.06	1.99	0.06	0.06	0.94	0.01	0.01			
	4	0.3	1.91	0.10	0.07	0.85	0.09	1.99	0.07	0.07	1.99	0.07	0.07	0.93	0.01	0.01			
	3	0.3	1.83	0.03	0.03	0.83	0.14	1.99	0.10	0.09	1.99	0.10	0.09	0.90	0.01	0.01			
$\sqrt{2}$	5	0.7	1.98	0.01	0.01	0.96	0.03	2.00	0.02	0.02	2.00	0.02	0.02	0.98	0.01	0.01			
	4	0.7	1.96	0.06	0.05	0.91	0.04	1.98	0.06	0.05	1.98	0.06	0.05	0.94	0.02	0.02			
	3	0.7	1.88	0.08	0.04	0.85	0.12	1.98	0.12	0.09	1.98	0.12	0.09	0.89	0.02	0.02			

C: Correctly selected non-zero coefficient variables; IC: incorrectly selected non-zero coefficient variables. NLME-2 ALASSO: two-stage NLME with adaptive LASSO penalty; NLME-EM ALASSO: EM algorithm-based penalized NLME with adaptive LASSO penalty.

Table 3

Results from the joint model (Web-1) for Total Motor Score, Symbol Digit Modalities Test and Stroop Word Test by CAG group.

Parameter	log(Total Motor Score)			Symbol Digit Modalities Test			Stroop Word Test			Group Difference*	
	Estimate	95% CI		Estimate	95% CI		Estimate	95% CI			p-value
$\theta_{3,2}$	0.04	[0.03, 0.06]		-0.08	[-0.14, -0.02]		-0.13	[-0.27, 0.01]			
$\theta_{4,2}$	61.19	[56.51, 65.87]		62.03	[53.06, 71.01]		61.16	[49.45, 72.86]			0.99
$\theta_{3,3}$	0.08	[0.07, 0.10]		-0.12	[-0.15, -0.09]		-0.18	[-0.24, -0.12]			
$\theta_{4,3}$	54.98	[53.11, 56.84]		56.60	[54.25, 58.95]		50.54	[47.95, 53.14]			0.002
$\theta_{3,4}$	0.17	[0.13, 0.20]		-0.29	[-0.37, -0.21]		-0.46	[-0.59, -0.33]			
$\theta_{4,4}$	44.20	[42.89, 45.52]		42.33	[40.79, 43.86]		43.30	[42.07, 44.53]			0.19
$\theta_{3,5}$	0.16	[0.11, 0.21]		-0.75	[-1.32, -0.18]		-0.14	[-0.21, -0.06]			
$\theta_{4,5}$	30.68	[29.12, 32.23]		32.57	[30.80, 34.33]		33.00	[30.46, 35.54]			0.23

$\theta_{3,g}$ and $\theta_{4,g}$ are the progression rate and injection point, respectively, for group g ; $g = 2, \dots, 5$. The updated injection points after including neuroimaging biomarkers for groups 2–5 are 58.15, 54.25, 46.32, and 22.62 for total motor score; 62.07, 54.22, 42.71, and 23.17 for SDMT, and 57.01, 50.10, 41.60, and 30.71 for Stroop Word.

* : p -value for the test of difference in $\theta_{4,g}$ for three outcomes in each CAG group.