Published in final edited form as:

Ann Appl Stat. 2019 ; 13(2): 1295–1318. doi:10.1214/18-AOAS1236.

EARLY DIAGNOSIS OF NEUROLOGICAL DISEASE USING PEAK DEGENERATION AGES OF MULTIPLE BIOMARKERS

Fei Gao,
Department of Biostatistics, University of Washington, Seattle, Washington 98195, feigao@uw.edu

Yüanjia Wang,
Department of Biostatistics, Columbia University, New York, New York 10032, yw2016@cumc.columbia.edu

Donglin Zeng,
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, dzeng@email.unc.edu

Alzheimer's Disease Neüroimaging Initiative

University of Washington, Columbia University and University of North Carolina at Chapel Hill

Abstract

Neurological diseases are due to the loss of structure or function of neurons that eventually leads to cognitive deficit, neuropsychiatric symptoms, and impaired activities of daily living. Identifying sensitive and specific biological and clinical markers for early diagnosis allows recruiting patients into a clinical trial to test therapeutic intervention. However, many biomarker studies considered a single biomarker at one time that fails to provide precise prediction for disease age at onset. In this paper, we use longitudinally collected measurements from multiple biomarkers and measurement error-corrected clinical diagnosis ages to identify which biomarkers and what features of biomarker trajectories are useful for early diagnosis. Specifically, we assume that the subject-specific biomarker trajectories depend on unobserved states of underlying latent variables with the conditional mean follows a nonlinear sigmoid shape. We show that peak degeneration age of the biomarker trajectory is useful for early diagnosis. We propose an Expectation-Maximization (EM) algorithm to obtain the maximum likelihood estimates of all parameters and conduct extensive simulation studies to examine the performance of the proposed methods. Finally, we apply our methods to studies of Alzheimer's disease and Huntington's disease and identify a few important biomarkers that can be used for early diagnosis.

SUPPLEMENTARY MATERIAL

Supplement A: Supplement to: Early Diagnosis of Neurological Disease Using Peak Degeneration Ages of Multiple Biomarkers (doi: 10.1214/00-AOASXXXXSUPP; .pdf). This supplement provides additional information on the theorem and proof on model identifiability, protocol for simulation studies, details on estimating the magnitude of measurement error, and residual plots of the examples.

Supplement B: R Codes. (doi: 10.1214/00-AOASXXXXSUPP; .pdf). Codes for the EM algorithm and simulation illustrating the proposed methods. The codes are available at http://columbia.edu/~yw2016/BPA.zip.

Keywords

Alzheimer’s disease; Huntington’s disease; inflection point; measurement error; nonlinear mixed effects model; sigmoid function

1. Introduction.

Neurological diseases, such as Huntington’s disease (HD), Alzheimer’s disease (AD), and Parkinson’s disease, involve the loss of structure or function of neurons that eventually leads to cognitive deficit, motor impairment, neuropsychiatric symptoms, and impaired activities of daily living. There are currently no disease-modifying treatments for these disorders since damaged neurons cannot be replaced or reproduced. The pathophysiological process of the diseases is thought to begin years before irremediable neuronal loss and cognitive deficits manifest (Sperling et al., 2011). Therefore, early diagnosis offers an opportunity for effective therapeutic intervention because the cognitive function might be preserved at the highest level possible before irreversible damage has occurred.

To develop effective therapeutics, it is important to identify biomarkers with the most rapid change at the earliest age and also associated with clinical diagnosis. Many subtle clinical features and biomarkers of preclinical pathological change can potentially serve as early diagnostic or prognostic indicators. For example, prognostic biomarkers in the motor, imaging, and cognitive domains are suggested to be useful for predicting early motor or cognitive abnormalities in HD (Paulsen, Long, Ross et al. 2014). For AD, various neurobiological measures, such as cerebrospinal fluid levels of $A\beta_{42}$ and total tau protein, show preclinical alterations that predict development of early AD symptoms (Hampel et al., 2008). However, all these findings are based on isolated analysis and it remains largely unknown which biomarkers manifest significant changes prior to disease onset and for how long before the onset.

To evaluate the relationship of changes in biomarkers and clinical diagnosis of AD, Hall et al. (2000, 2001, 2003) modeled longitudinal measurements of one or two biomarkers by change point polynomial mixed models, where the change point is associated with the age of clinical diagnosis that is assumed to be observed for all subjects. Later, Jacqmin-Gadda, Commenges and Dartigues (2006) extended the methods to jointly model measurements of a biomarker and right-censored age of clinical diagnosis. However, the change point only indicates the change of pattern of the biomarker over time and may not necessarily be the acceleration time of the biomarker change. Recently, an imputation-based analysis was used in Bateman et al. (2012). In this method, the biomarker measurements were first aligned by the age from the expected AD clinical diagnosis, and a cubic polynomial mixed effects model was used to model the biomarker trajectory retrospectively. The earliest time prior to the AD onset where a difference can be detected between mutation carriers and non-carriers and when the maximal difference is detected were considered as critical time points. There are several limitations with this analysis. First, participants (children of parents who had AD and carried mutations associated with AD) were recruited before being diagnosed with AD, thus their onset ages were censored. Bateman et al. (2012) imputed participant’s AD age at onset using their parents' age at onset since their approach does not handle censoring. This imputation may introduce inaccuracy into the analysis. Second, the analysis in Bateman et al. (2012) did not model multiple biomarkers simultaneously.

To model both longitudinal measurements and disease onset, joint modeling approaches, including selection models and pattern mixture models (Little, 1995; Hogan and Laird, 1997; Tsiatis and Davidian, 2004), have been extensively used. However, since these joint modeling approaches rely on some shared random effects to link longitudinal biomarkers with disease age at onset, they are not useful to identify any subject-specific biomarker features that are present prior to the disease onset. Furthermore, these methods do not handle the complication that the disease age at onset may be subject to measurement error, as commonly encountered in the studies of neurodegenerative diseases (Garcia, Marder and Wang, 2017).

In this paper, we model longitudinal measurements of multiple biomarkers and error-corrected clinical diagnosis age simultaneously. Our goal is to identify which biomarkers and what features of biomarker trajectories are useful for early diagnosis and characterization of disease progression. Specifically, to capture nonlinear sigmoid shape of the biomarker degeneration as observed in empirical studies (Jack et al., 2010; Jedynak et al., 2012), we assume that subject-specific trajectories of biomarkers are related to latent states of underlying neuron masses. This assumption is motivated by neural mass models (Hopfield, 1982), where neurons are considered as binary units in an active or inactive state and the population-level model of their activities is considered as aggregate activities of massive number of neurons. Furthermore, we allow biomarker-specific lead time between the disease onset and the peak degeneration ages of the biomarkers (inflection points where the maximal change of biomarker occurs) to vary across biomarkers and allow inflection points to depend on subject-specific covariates. We show that biomarker inflection points are useful for early diagnosis of neurological diseases. In addition, since biomarker at the peak degeneration age is most sensitive to change and easiest to be detected, inflection points indicate the optimal timing of intervention when designing clinical trials if the inflection point occurs prior to disease onset and closely monitoring is available. Furthermore, we show that the biomarker-specific lead time is an important feature to characterize disease progression.

To accommodate measurement error of the clinical diagnosis age, we assume an additive measurement error model. To bypass a difficult nonlinear optimization in our modeling, an EM algorithm with explicit solutions in the M-step is developed for maximum likelihood estimation. We conduct simulation studies to examine the performance of the proposed estimators and show that Bateman et al. (2012) approach to impute unobserved disease onset ages may lead to large bias in the biomarker trajectories and an increased variability in the estimation of parameters. Finally, we apply our methods to two studies of neurodegenerative diseases (HD and AD), where we identify biomarkers with peak degeneration ages occurring significantly earlier than clinical disease onset so that they can potentially serve as early diagnostic markers.

2. Motivating examples.

2.1. HD and Predictors of Huntington's Disease (PREDICT-HD) study.

HD is an autosomal dominant neurodegenerative disease caused by an expansion of the cytosine-adenine-guanine (CAG) in the first exon of *huntingtin* (*HTT*) gene (MacDonald et al., 1993). Whereas unaffected persons have a range of 6–35 CAG repeats, persons affected with HD have 36–121 CAG repeats length (Kremer et al., 1994; Rubinsztein et al., 1996). HD has a broad impact on a person's functional abilities and usually results in movement, cognitive and psychiatric impairments. Even though CAG repeats length and baseline age are recognized as important predictors of HD diagnosis, much effort is needed to refine the prediction of the age at motor onset.

The PREDICT-HD study is a prospective observational study of premanifest HD individuals who carry an expansion of CAG repeats (thus at risk of HD) but without a clinical diagnosis at the baseline (Paulsen, Long, Johnson et al. 2014). These pre-symptomatic, gene-positive individuals were recruited starting 2002 and followed for up to 12 years. During the follow-up period, various longitudinal measures in five domains (motor, cognitive, psychiatric, functional, and imaging) were collected. The onset of HD was determined by the motor symptoms evaluated on the Unified Huntington's Disease Rating Scale (UHDRS) by a trained neurologist. A subject rated as 4 on the diagnostic confidence level (DCL) is diagnosed with HD. However, the presence of variation in patients' motor symptoms and raters' diagnosis has made clinical diagnosis difficult (Garcia et al. 2017): a patient could receive a DCL of 4 (diagnosed with HD) at one visit, but fail to reach a DCL of 4 at the next visit if the patient expresses less motor symptoms (free of HD diagnosis). In the PREDICT-HD study, 63 (4.6%) patients had such reversion of diagnosis. Therefore, the observed HD age at onset determined by a neurologist is an approximation of a patient's true disease age at onset. Our proposed method will account for the random measurement errors in diagnosis age using a linear model with a known variance estimated from the incidences of disease status change in the PREDICT-HD study.

2.2. AD and Alzheimer's Disease Neuroimaging Initiative (ADNI) study.

AD is an irreversible neurodegenerative disease that results in a loss of cognitive function due to the deterioration of brain neuronal synapses. The progression of AD has been divided into three phases. The first phase is a pre-symptomatic phase where individuals are cognitively normal but some have AD pathological changes. The second prodromal phase, often referred to as mild cognitive impairment (MCI), is characterized by the onset of the earliest cognitive symptoms that do not meet the criteria for dementia. The final phase in the evolution of AD is dementia, defined as impairments in multiple domains that are severe enough to produce loss of function. To determine the sequence of pathological changes of AD, a sigmoid model was proposed and widely used for major AD biomarkers (Jack et al., 2010). Although some agreement between the temporal ordering of major biological cascade has been reached, there is no method to precisely estimate the lead time between when the peak biomarker degeneration occurs (inflection point) and dementia diagnosis, accounting for censoring and error in dementia diagnosis.

Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. In three phases of the study (ADNI1, ADNI GO, and ADNI2), early mild cognitive impairment (EMCI), MCI, mild AD and normal control subjects were recruited. Biomarkers, such as brain scans, genetic profiles, and biomarkers in blood and cerebrospinal fluid, were collected to track the progression of the disease. MCI was determined if the subject has Mini-Mental State Exam (MMSE) score between 24–30, a memory complaint, objective memory loss measured by education adjusted scores on Wechsler Memory Scale Logical Memory II, a Clinical Dementia Rating (CDR) of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia. Dementia was determined if the subject has MMSE score between 20–26, CDR of 0.5 or 1.0, and meets NINCDS/ADRDA (National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association) criteria for probably AD.

Similar to HD, random variations of the clinical diagnosis of MCI and dementia were observed. Sources of the variations include normal aging independent of AD, "cognitive reserve" due to education-linked factors, and disease heterogeneity (Nelson et al., 2012). In the ADNI study, 75 (4.3%) patients had received a diagnosis of MCI or AD at one visit, but was then diagnosed as normal at the next visit. Similarly to the PREDICT-HD study, the known variance in the measurement error model can be estimated using the observations of disease status change in the ADNI study.

3. Method.

3.1. Latent suppression state model for progression markers.

We consider K neurological disease markers measured over time from n independent subjects. For subject i, we let $Y_{ik}(t)$ be the measurement from the kth marker at age t for $k = 1, \dots, K$ and let W_i denote the underlying unobserved true disease age at onset. Additionally, we let $\boldsymbol{Z}_i$ denote a vector of baseline covariates for subject i. Our first model is to assume that in the population the disease onset follows $W_i \sim N\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}_i, \sigma_W^2\right)$, where $\boldsymbol{X}_i = \left(1, \boldsymbol{Z}_i^T\right)^T$. Given W_i and $\boldsymbol{Z}_i$, our models for K disease markers are motivated by the neural mass models in Hopfield (1982). Neural mass model was used to describe the aggregate activities of massive number of neurons. This approach motivates the population-level model by considering neurons as binary units in an active or inactive state. Assuming neuronal responses rest on a threshold of activity, any unimodal distribution of thresholds results in a sigmoid activation function at the population, following trajectories similar to those observed empirically for many neurological disease progression markers (Jack et al., 2010).

Specifically, we assume that marker $Y_{ik}(t)$ reflects the activity levels of neuron mass at age t and such levels further depend on the latent suppression status as suggested in the neural mass model. The suppression status of the neuron mass may be permanent or instantaneous, where the former most likely associates with susceptibility to neurodegeneration and the latter most likely associates with progression of neurodegeneration. Let Q_{ik} indicate the presence of the permanent suppression of the neuron mass (for instance, due to genetic mutation, neuronal injury, or nerve damage) and let $H_{ik}(t)$ indicate the instantaneous suppression at age t (for instance, due to neurofibrillary tangles). When subject i has no permanent suppression (*i.e.*,$Q_{ik} = 0$), or does not experience any instantaneous suppression at age t (*i.e.*, $H_{ik}(t) = 0$), we assume a linear declination trend due to normal aging process as suggested in Fjell et al. (2009). That is, when $Q_{ik} = 0$ or $H_{ik}(t) = 0$, we assume a linear mixed effects model for $Y_{ik}(t)$:

$$Y_{ik}(t) = \alpha_{0k} + \beta_k t + \nu_{ik} + \epsilon_{ik}(t),$$

where ν_{ik} is the subject- and marker-specific random intercept following a mean-zero normal distribution with unknown variance $\sigma_{k\nu}^2$, and $\epsilon_{ik}(t)$ is a white noise process with variance $\sigma_{k\epsilon}^2$. When suppression is present at age t, either due to the permanent suppression (*i.e.*, $Q_{ik} = 1$) or the instantaneous suppression at age t (*i.e.*, $Q_{ik} = 0$, $H_{ik}(t) = 1$), a further reduction in $Y_{ik}(t)$ occurs due to disease degenerative process (Fjell et al., 2009). Thus, we assume that the marker level at age t is further reduced by a subject-specific value, $\boldsymbol{\alpha}_{1k}^{\mathrm{T}}\boldsymbol{X}_i$. In other words, depending on the latent suppression states, our progression model assumes

$$Y_{ik}(t) = \alpha_{0k} + \boldsymbol{\alpha}_{1k}^{\mathrm{T}}\boldsymbol{X}_i\left\{Q_{ik} + \left(1 - Q_{ik}\right)H_{ik}(t)\right\} + \beta_k t + \nu_{ik} + \epsilon_{ik}(t)$$

for $k = 1,\dots, K$.

To model the distribution of Q_{ik} and $H_{ik}(t)$, we first assume that Q_{ik} is independent of W_i and satisfies the following logistic regression model:

$$\operatorname{logit}\Pr\left(Q_{ik} = 1 \middle| \boldsymbol{X}_i\right) = \boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{X}_i .$$

Since the instantaneous suppression is most relevant to the disease progression, we let $H_{ik}(t)$ depend on disease age at onset W_i, through

$$\Pr\left(H_{ik}(t) = 1 \middle| Q_{ik} = 0, W_i\right) = \frac{1}{1 + \exp\left\{-b_k\left(t - \mu_k - W_i\right)\right\}},$$

where b_k is an unknown parameter. Since the above sigmoidal model has an inflection point at $t_i^* = \mu_k + W_i$, the risk of experiencing an instantaneous suppression of the neuron mass increases over age, accelerates near age t_i^* until reaching its peak at t_i^*, and then the risk remains to increase but at a decelerated speed afterwards. Moreover, if $\mu_k < 0$, the peak suppression age has a lead time of $|\mu_k|$ prior to the disease onset. This suggests that the marker degeneration peaks before the disease onset, so it can potentially be used for early diagnosis. On the contrary, if $\mu_k > 0$, the inflection point age is after W_i, so the marker degeneration peaks after the disease onset, suggesting that this marker may be more likely to manifest a post-disease onset effect. Clearly, $|\mu_k|$ gives a magnitude of the lead time or lag time. For the purpose of early diagnosis, we aim to identify the progression marker with $\mu_k < 0$ and estimate the magnitude of $|\mu_k|$ to inform clinical trial design and recruitment.

Remark 3.1. From the proposed latent state models, the conditional mean for the progression marker $Y_{ik}(t)$ given W_i but marginalized over Q_{ik} and $H_{ik}(t)$ is given by

$$\alpha_{0k} + \frac{\boldsymbol{\alpha}_{1k}^{\mathrm{T}}\boldsymbol{X}_i}{1 + \exp\left(\boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{X}_i\right)}\left[\exp\left(\boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{X}_i\right) + \frac{1}{1 + \exp\left\{-b_k\left(t - W_i - \mu_k\right)\right\}}\right] + \beta_k t\,.$$

Thus, the smoothed trend of the marker measurement $Y_{ik}(t)$ is a sigmoid function with a linear drift over age. The peak degeneration age, $t_i^* = \mu_k + W_i$, coincides with the inflection point of the smoothed marker trajectory, which is the age of the maximal deterioration of the trajectory. Therefore, by monitoring the marker values with $\mu_k < 0$ and identifying the peak age of deterioration, one can make early diagnosis with $|\mu_k|$ time units ahead of the disease onset in individuals. Note that existing literature suggests that many neurological biomarkers manifest a nonlinear sigmoid shape (Jack et al. 2012; Jedynak et al. 2012; Samtani et al. 2012; Paulsen, Long, Ross et al. 2014), which is consistent with our model of $Y_{ik}(t)$ given W_i.

3.2. Likelihood-based estimation and inference.

In our applications of HD and AD studies, the biomarkers are collected longitudinally at discrete time points and some biomarkers may not be measured at the same time as the others. We assume that for $i = 1,\dots, n$, biomarker k ($k = 1,\dots, K$) is measured at $\left\{t_{i1k}, \dots, t_{i,n_{ik},k}\right\}$, where n_{ik} is the number of measurements. We use Y_{ijk} for $Y_{ik}(t_{ijk})$.

Another complication commonly encountered in the studies of neurological diseases is that the disease diagnosis relies on clinical assessments which are known to be imprecise. Therefore, the clinically diagnosed age at onset, denoted by T_i, is the true age at onset measured with error. Particularly, we assume that the measurement error δ_i is additive and normally distributed with known constant variance σ_δ^2 that can be determined apriori using observed data of clinical diagnosis or from existing literature, i.e.,

$$T_i = W_i + \delta_i, \quad \delta_i \sim N\left(0, \sigma_\delta^2\right).$$

Additionally, we assume that T_i is subject to right censoring due to the end of the study or patient's loss of follow-up. Let C_i denote the censoring age, such that we observe $\tilde{Y}_i \equiv \min\left(T_i, C_i\right)$ and $\Delta_i \equiv I(T_i \le C_i)$. The observed data from subject i consist of

$$\mathcal{O}_i = \left\{t_{ijk}, Y_{ijk}, \mathbf{Z}_i, \tilde{Y}_i, \Delta_i : k = 1, \dots, K;\ j = 1, \dots, n_{ik}\right\}.$$

Let $\phi(\cdot;\, \sigma^2)$ and $\Phi(\cdot;\, \sigma^2)$ denote the density function and cumulative distribution function of $N(0,\, \sigma^2)$, respectively. Write $\boldsymbol{\alpha}_k = (\alpha_{0k},\, \boldsymbol{\alpha}_{1k})$. Define $g_{ijk}(W_i;\mu_k,b_k) = \exp\{-b_k(t_{ijk} - W_i - \mu_k)\}$,

$$M_{ijk} = Y_{ijk} - \alpha_{0k} - \beta_k t_{ijk},$$

$$A_{ijk}\left(\nu_{ik};\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2\right) = \phi\left(M_{ijk} - \nu_{ik} - \alpha_{1k}^{\mathrm{T}} X_i;\sigma_{k\epsilon}^2\right),$$

$$B_{ijk}\left(\nu_{ik};\alpha_{0k},\sigma_{k\epsilon}^2\right) = \phi\left(M_{ijk} - \nu_{ik};\sigma_{k\epsilon}^2\right),$$

and

$$D_{ijk}\left(\nu_{ik},W_i;\mu_k,b_k,\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2\right) = \frac{g_{ijk}\left(W_i;\mu_k,b_k\right)B_{ijk}\left(\nu_{ik};\alpha_{0k},\sigma_{k\epsilon}^2\right) + A_{ijk}\left(\nu_{ik};\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2\right)}{1 + g_{ijk}\left(W_i;\mu_k,b_k\right)}.$$

Assuming that C_i is independent of T_i, W_i, and Y_{ijk} given $\boldsymbol{Z}_i$, the observed-data likelihood function concerning the parameters $\left(\boldsymbol{\alpha}_k,\beta_k,\sigma_{k\nu}^2,\sigma_{k\epsilon}^2,\boldsymbol{\eta}_k,\mu_k,b_k\right)$ $(k = 1,\ldots,K)$ and $\left(\boldsymbol{\theta},\sigma_W^2\right)$ is given by

$$L_n = \prod_{i=1}^{n}\int_{W_i}\left\{\prod_{k=1}^{K} q_k\left(W_i;\eta_k,\mu_k,b_k,\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2,\sigma_{k\nu}^2\right)\right\}h_i\left(W_i;\sigma_W^2,\sigma_\delta^2\right)dW_i,$$

where

$$\begin{aligned} & q_k\left(W_i;\eta_k,\mu_k,b_k,\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2,\sigma_{k\nu}^2\right) \\ & = \int_{\nu_{ik}}\frac{\exp\left(\boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{X}_i\right)\prod_{j=1}^{n_{ik}} A_{ijk}\left(\nu_{ik};\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2\right) + \prod_{j=1}^{n_{ik}} D_{ijk}\left(\nu_{ik},W_i;\mu_k,b_k,\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2\right)}{1 + \exp\left(\boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{X}_i\right)} \times \phi\left(\nu_{ik};\sigma_{k\nu}^2\right)d\nu_{ik}, \end{aligned}$$

and

$$h_i\left(W_i;\sigma_W^2,\sigma_\delta^2\right) = \phi\left(W_i - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Z}_i;\sigma_W^2\right)\phi\left(\tilde{Y}_i - W_i;\sigma_\delta^2\right)^{\Delta_i}\Phi\left(W_i - \tilde{Y}_i;\sigma_\delta^2\right)^{1-\Delta_i}.$$

We propose to maximize the likelihood function for parameter estimation. To compute the maximum likelihood estimates, we apply an EM algorithm treating Q_{ik}, V_{ik},

$H_{i1k}, \ldots, H_{i, n_{ik}, k}$, and W_i ($i = 1, \ldots, n$; $k = 1, \ldots, K$) as missing data, where $H_{ijk} = H_{ik}(t_{ijk})$. The details of the EM algorithm are described in the Appendix A.

Asymptotically, all parameter estimators are consistent and efficient following the standard maximum likelihood theory, provided that the model parameters are identifiable and the Fisher information matrix is non-singular. In particular, we prove the identifiability in Section S.1 of the supplemental materials. Due to the lack of an analytical form, we estimate the covariance matrix of the estimators through the nonparametric bootstrap. Specifically, for each bootstrap, we sample n subjects with replacement. The covariance matrix is then estimated by the sample covariance matrix of the bootstrap estimators.

3.3. Early diagnosis of disease onset.

Given the fitted model, the identified biomarkers with peak degeneration ages occurring before the disease onset can be used for disease monitoring and contribute to early diagnosis. In addition, we are able to predict the precise disease age at onset given observations of biomarkers. For a future subject who has not been diagnosed at age t with biomarker measurements $\boldsymbol{Y}_k \equiv \left(Y_{1k}, \ldots, Y_{n_k, k}\right)$ ($k = 1, \ldots, K$) measured at $t_{1k}, \ldots, t_{n_k, k}$ prior to age t, the disease age at onset can be predicted given the biomarkers and the diagnosis information. That is, we predict the disease age at onset W by the posterior mean of W given the biomarker measurements and the diagnosis information, $E(W|\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_K, T \geq t)$, which is given by

$$\int w\psi(w)dw,$$

where $\psi(w)$ is the posterior density function of the disease age at onset W that is given by

$$\frac{\phi\left(w - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Z}; \sigma_W^2\right)\Phi\left(w - t; \sigma_\delta^2\right)\prod_{k=1}^{K} q_k\left(w; \eta_k, \mu_k, b_k, \boldsymbol{\alpha}_k, \sigma_{k\epsilon}^2, \sigma_{k\nu}^2\right)}{\int\phi\left(W - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Z}; \sigma_W^2\right)\Phi\left(W - t; \sigma_\delta^2\right)\prod_{k=1}^{K} q_k\left(W; \eta_k, \mu_k, b_k, \boldsymbol{\alpha}_k, \sigma_{k\epsilon}^2, \sigma_{k\nu}^2\right)dW},$$

and the integral can be evaluated by numeric integration with Gauss-Hermite quadratures.

4. Simulations.

We conducted simulation studies to examine the performance of the proposed methods. A detailed description of the simulation protocol is given in Section S.2 of the supplementary materials. We considered $K = 2$ biomarkers and generated two independent covariates $Z_{i1} \sim N(0,1)$ and $Z_{i2} \sim$ Bernoulli(0.5) for $i = 1, \ldots, n$. We generated the censoring age C_i from Uniform[0, 10]. For each biomarker k and each subject i, we randomly chose n_{ik} from {3,…, 10} with equal probabilities and randomly generated t_{ijk} (j = 1,…, n_{ik}) independently from Uniform[0, C_i].

We generated the data from the proposed models, with the values of the parameters given in the second column of Table 1 and $\sigma_{\delta}^{2}=0.2$. The censoring rate is about 30%. We set $n = 200$ or 400 and used 1,000 replicates. The algorithm was regarded as converged if the maximum of the norms of the parameter differences in adjacent iterations is smaller than 0.001. For each simulated dataset, 100 bootstrapped datasets were used for variance estimation.

Tables 1 summarizes the simulation results, where the algorithm converged for all simulated datasets. Bias and SE are the median bias and standard error, respectively, of the parameter estimator, SEE is the median of the standard error estimator, and CP is the coverage probability of the 95% confidence interval. The biases for all parameter estimators are small and decrease as n increases. The variance estimators for a_{0k}, $\boldsymbol{\alpha}_{1k}$, β_k, μ_k, b_k, $\boldsymbol{\theta}$, and σ_{W}^{2} are accurate, especially for large n. The variance estimator for $\boldsymbol{\eta}_k$ slightly overestimates the true variabilities, but it gets more accurate as sample size increases. The confidence intervals have satisfactory coverage probabilities when the sample size is large ($n = 400$).

To evaluate the performance of the proposed prediction procedure, for each simulation replicate, we generated an independent data set of sample size 2,000. The data were generated in the same manner, except that we included only censored subjects. We predicted the disease age at onset for the censored subjects in the new dataset using the parameter estimators from the original replicate and compared the predicted ages at disease on set with the true disease onset ages. In addition, we calculated the average logarithmic score (Good, 1952; Bernardo, 1979; Gneiting, Balabdaoui and Raftery, 2007), which is the average of the negative logarithm of the predictive density function evaluated at the true disease onset age, such that a smaller value indicates a better fit. We compared the results with the proposed models with both biomarkers and one biomarker only.

Table 2 shows the mean prediction error, adjusted standard deviation (adjusted SD), and the mean adjusted logarithmic score (adjusted LS), where the adjusted SD is calculated as the squared root of mean squared prediction error minus the intrinsic prediction error variability that is estimated as the mean squared prediction error using the conditional mean of the disease age at onset given the diagnosis, and the adjusted LS is calculated as the logarithmic score minus that from the two-biomarker model with the true parameter values. The biases from all models are small. The adjusted SD and adjusted LS decrease as n increases. The adjusted LS based on both biomarkers is lower than those based on the models with one biomarker. Compared to those from the models with one biomarker only, the prediction based on both biomarkers has smaller variability: for $n = 400$, the improvement in prediction efficiency of using both biomarkers is about 15%.

5. Applications.

5.1. HD and PREDICT-HD study.

We applied the proposed methods to the aforementioned PREDICT-HD study. We included three motor markers (Ocular, Brady, and Chorea) measuring impairment in movement and three cognitive markers (SDMT, Stroop-WO, and Smell-ID) measuring impairment in cognition. Ocular, Brady, and Chorea are the ocular, bradykinesia, and chorea subscales from the UHDRS, reflecting ratings of eye movement and tracking, abnormal slowness or rigidity of movement, and abnormal involuntary movement disorder, respectively (Huntington Study Group, 1996). SDMT is the symbol digit modalities test that measures working memory, complex scanning, and processing speed. Stroop-WO is the stoop word test that measures basic attention and processing speed. Smell-ID is the University of Pennsylvania smell identification test that measures the olfactory recognition. The covariates $\boldsymbol{Z}_i$ for HD age at onset include baseline age, years of education, gender, and length of CAG repeats.

We included 1,073 gene-positive subjects with more than 35 CAG repeats at *huntingtin* gene in the analysis. During the follow-up, 225 (21%) subjects developed HD and the age at disease onset is defined as the age of the first observation with DCL=4. For each marker, on average more than three measurements are available for each subject. We estimated the magnitude of measurement error σ_δ^2 of HD diagnosis from the PREDICT-HD study. In particular, we fitted the adjacent observations with status change (from DCL<4 to DCL=4, or reverse) by a generalized linear model to obtain $\sigma_\delta^2 = 0.324$. The details of the estimation procedure are given in Section S.3 of the supplemental materials.

Table 3 shows the estimation results for various parameters associated with the peak degeneration ages and HD age at onset, where 1,000 bootstrap samples were used for variance estimation. Male subjects have later HD age at onset than females. Longer years of education and shorter CAG repeats length are associated with later HD age at onset. The inflection of the three motor measures occur close to HD age at onset, with the 95% confidence intervals of the lead times containing zero. These results are expected since the motor scores measure a patient's motor symptoms and HD diagnosis is also mainly based on motor function. In addition, this finding is also consistent with the existing literature suggesting that subtle motor abnormalities accelerate just prior to diagnosis (Long et al., 2014). The symbol digit modalities and stroop word cognitive tests, which have respective lead times approximately 2 and 1.5 years before HD onset and significantly earlier than HD onset, can be candidate markers for early detection of HD diagnosis.

Next, we examined the differences of biomarker values and peak degeneration ages among subgroups of subjects. Figure 1 shows the average estimated biomarker values among the subgroups of subjects with different CAG repeats length. Subjects with a longer CAG expansion are associated with an earlier HD age at onset and an earlier peak degeneration age for all considered biomarkers. In particular, subjects with CAG expansion < 41, 41 ≤ CAG expansion < 43, and CAG expansion ≥ 43 have peak degeneration ages of symbol digit modalities test at approximate 57, 50, and 41 years old, respectively, with corresponding scores 46, 44, and 42. Those subjects have peak degeneration ages of stroop word cognitive test at approximate 58, 51, and 42 years old, with corresponding scores 90, 86, and 84.

Finally, we examined the performance of the proposed methods on the prediction of HD age at onset given the biomarker measurements. Figure 2 presents the difference of the predicted HD age at onset and the observation age for each individual. For the non-censored subjects, the difference between the predicted and observed HD age at onset is within the measurement variability of T_i (within the distance of $\sqrt{\sigma_\delta^2 + \sigma_W^2}$). For the censored subjects, most of the predicted HD age at onset is beyond the lower limit of the censoring age considering variability of the disease age at onset (i.e., beyond censoring age minus $\sqrt{\sigma_\delta^2 + \sigma_W^2}$). The proposed methods thus provide adequate fit to the PREDICT-HD data.

5.2. AD and ADNI study.

We applied the proposed methods to the aforementioned ADNI study. We analyzed the combined MCI and AD as a composite event, which serves as an alternative definition of early AD as suggested by Dubois et al. (2007). We considered four markers: the Montreal Cognitive Assessment (MOCA) that assesses several cognitive domains; the Clinical Dementia Rating Sum of Boxes (CDRSB) that measures the staging severity of dementia; the Functional Activities Questionnaire (FAQ) that serves a screening tool for evaluating activities of daily living; and the $A\beta_{42}$ protein level (ABETA) measured from the cerebrospinal fluid. We associated the markers and early AD age at onset to baseline age, gender, education, number of APOE ϵ4 alleles, baseline Alzheimer's Disease Assessment Scale 11 terms total scores (TOTAL11), and baseline FAQ.

We included 414 subjects who were cognitively normal at the baseline, out of whom 87 (21.0%) subjects developed early AD during the follow-up. For each marker, more than two measurements are available for each subject. We estimated the magnitude of the measurement error using the generalized linear model as described in Section S. 3 in the supplemental materials to obtain $\sigma_\delta^2 = 1.47$.

Table 4 shows the estimation results of various parameters associated with the peak degeneration ages and age at early AD onset. Carriers of APOE ϵ4 alleles have a younger early AD age at onset than non-carriers, and larger values of baseline TOTAL11 and baseline FAQ are associated with younger age at onset. The peak degeneration ages of MOCA, FAQ, and CDRSB occur later than early AD onset. For ABETA, the peak degeneration occurs approximately 12 years before onset, suggesting that it is a candidate for early detection of early AD. This finding agrees with the hypothesis that $A\beta$-plaque deposits are early events in the AD cascade occurring before the appearance of clinical symptoms (Jack et al., 2010; Bateman et al., 2012).

The estimated lag times also have implications on clinical trials design. The peak acceleration of MOCA, FAQ and CDRSB occurs within about 1.5 years after diagnosis. A clinical trial designed to test changes in these measures in response to a therapy may recruit newly diagnosed MCI or AD patients within about 1.5 years to improve power.

Figure 3 shows the average estimated biomarker values among carriers and non-carriers of APOE ϵ4 alleles. Carriers are associated with a younger age at onset and an earlier peak degeneration age for all considered biomarkers. In particular, carriers and non-carriers have a peak ABETA acceleration at approximate 74 and 76 years of age, respectively. Early AD onset occurs approximately at 82 and 84 years for the two groups. The corresponding Aβ_{42} cutoff values are 143 and 183 pg/mL, which are slightly lower than the recommended threshold for using Aβ_{42} to define AD in Shaw et al. (2009) (Aβ_{42} < 192 pg/mL defined as AD, estimated as the value that maximizes the area under the receiver operating characteristic curve for the detection of AD). However, since the diagnostic test based on this threshold has a relatively high sensitivity (96.4%) and low specificity (76.9%), the reported cutoff in Shaw et al. (2009) may be anti-conservative.

Lastly, to see the potential bias of using parent's disease age at onset to impute offspring's AD age at onset as the analyses performed in Bateman et al. (2012), we simulated parent's age at onset and fit the proposed model. In particular, we assumed that the parent's age at onset has the same mean as the child's age at onset estimated from the proposed approach with a correlation of 0.3 or 0.65. For censored subjects, we imputed their age at onset by their parents' early AD age at onset. The simulated parent's onset age is on average 5.5 and 4.3 years different from the child's onset age. The red solid and dashed curves in Figure 4 show the average estimated values of biomarkers with censored onset ages replaced by imputation as in Bateman et al. (2012), where the black curves show our proposed approach that handles censoring appropriately. The horizontal axis is anchored at the estimated age at onset of early AD (years to onset of early AD). For both scenarios of correlation, imputing censored ages at onset leads to a large bias of the trajectories of biomarkers, and the estimated biomarker lead times can be shifted.

6. Discussion.

In this paper, we proposed a latent suppression state model to identify useful biomarkers for early disease diagnosis and estimate lead time to disease onset or lag time post onset. The proposed model is motivated from biological models of neural masses, and facilitates inference for modeling nonlinear sigmoid shapes of biomarker trajectories observed empirically. Furthermore, we proposed a computationally efficient EM algorithm with explicit solutions in the M-step and the evaluation of conditional expectation for the latent variables conducted using Gaussian quadratures. The numerical integration is at most two-dimensional, even if a large number of biomarkers are included.

For the asymptotic theory to hold, we require at least two measurements per biomarker for each subject. Empirically, we found that two measurements per biomarker for each subject provided stable estimation results for $n = 400$ (99.5% of the simulated datasets converged in simulated settings). This requirement on the number of measurements usually holds for neurological disease studies with relatively closely monitoring, as for the PREDICT-HD and ADNI studies.

A number of parametric assumptions are suggested to model the disease onset age and biomarker measurements. For example, we assume a functional relationship between the biomarker and suppression as well as the age at disease onset and measurement age. This parametric assumption is in fact very simple and standard and it yields a sigmoid shape of the observed biomarker with a peak degeneration age that is consistent with empirical observations and existing literature. In addition, we assume that measurement error for the disease onset age is normal distributed with known variance. In practice, some of these parametric assumptions may be violated and further investigation may be needed to study the performance of the proposed methods under mis-specified models.

In the PREDICT-HD study, we visualize the fit of the proposed model through comparing the predicted HD age at onset with the observation age graphically. We also examine the goodness of fit for the model of the biomarkers by plotting the residuals of the biomarker measurements against the ages at measurements (Rizopoulos, 2012, Chapter 6) in Figures S. 1 and S.2 in the supplementary materials. The proposed model is regarded as adequate since the predicted HD age at onset is consistent with the observation age, allowing for the existence of measurement errors, and the residuals are approximately randomly dispersed. A better model checking procedure may be developed to assess the goodness of fit of the proposed model.

In the ADNI study, we examined the performance of the imputation analyses in Bateman et al. (2012). Since the disease onset ages were observed in non-censored subjects, imputation was only applied to approximate disease onset for right-censored subjects. Even if the mean of the early AD age at onset was correctly specified, the trajectories of biomarkers were estimated with bias, and the inflection points were shifted (especially for Aβ_{42}). Our proposed methods make use of the observed diagnosis ages in non-censored subjects, appropriately handle censoring for those who were not diagnosed, and yield biomarker trajectories and peak degeneration ages with better accuracy and precision than Bateman et al. (2012).

The proposed approach, which assumes a normal distribution for the disease age at onset, can be extended to accommodate other parametric distributions, semiparametric distributions, or nonparametric distributions. For example, a proportional hazards model may be assumed for the age of disease onset. In addition, we may extend the proposed approach to accommodate interval-censored disease age at onset.

We assumed that the lead times or lag times between the peak degeneration of the biomarkers and the disease onset are the same for all subjects. This assumption can be easily relaxed to allow for subject-specific lead or lag times. For example, the biomarker model of AD proposed by Jack et al. (2010) hypothesized that the lag period between $A\beta$-plaque formation and neurodegenerative cascade may vary among subjects, indicating differences in $A\beta$ processing, brain resilience, or cognitive reserve. We may introduce subject-specific fixed effects and random effects to the sigmoid function to accommodate this general case, but with increased computational complexity.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This research is supported by U.S. NIH grants NS082062, NS073671, GM124104. The authors wish to thank the NIH dbGap data repository (accession number phs000222.v3.p2) and the PREDICT-HD study investigators. Data collection and sharing for the ADNI study in this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

This research is supported by U.S. NIH grants NS082062, NS073671, GM124104. ADNI Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

APPENDIX A:: DETAILS OF THE EM ALGORITHM

Denote $R_{ijk} = Q_{ik} + (1 - Q_{ik})\, H_{ijk}$. The complete-data log-likelihood concerning the parameters is given by

$$\sum_{i=1}^{n}\Bigg\{\log\phi\left(W_i - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Z}_i; \sigma_W^2\right) + \Delta_i \log\phi\left(\tilde{Y}_i - W_i; \sigma_\delta^2\right) + \left(1 - \Delta_i\right)$$
$$\times \log\Phi\left(W_i - \tilde{Y}_i; \sigma_\delta^2\right) + \sum_{k=1}^{K}\Bigg(\sum_{j=1}^{n_{ik}} \log\phi\left(M_{ijk} - \nu_{ik} - \boldsymbol{\alpha}_{1k}^{\mathrm{T}}\boldsymbol{X}_i R_{ijk}; \sigma_{k\epsilon}^2\right)$$
$$+ \log\phi\left(\nu_{ik}; \sigma_{k\nu}^2\right) + Q_{ik}\boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{X}_i - \log\left(1 + e^{\boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{X}_i}\right) + \left(1 - Q_{ik}\right)$$
$$\times \sum_{j=1}^{n_{ik}}\left[\left(1 - H_{ijk}\right)b_k\left(W_i + \mu_k - t_{ijk}\right) - \log\left\{1 + e^{b_k\left(W_i + \mu_k - t_{ijk}\right)}\right\}\right]\Bigg)\Bigg\}.$$

Since the complete-data log-likelihood can be factorized into pieces concerning disjoint subsets of parameters, we obtain the estimates for subsets of the parameters separately in the M-step. Specifically, we update $(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k)$ by

$$\left\{\sum_{i=1}^{n}\sum_{j=1}^{n_{ik}}\begin{pmatrix} 1 & \hat{E}\left(R_{ijk}\right)\boldsymbol{X}_i^{\mathrm{T}} & t_{ijk} \\ \hat{E}\left(R_{ijk}\right)\boldsymbol{X}_i & \hat{E}\left(R_{ijk}\right)\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}} & \hat{E}\left(R_{ijk}\right)t_{ijk}\boldsymbol{X}_i \\ t_{ijk} & \hat{E}\left(R_{ijk}\right)t_{ijk}\boldsymbol{X}_i^{\mathrm{T}} & t_{ijk}^2 \end{pmatrix}\right\}^{-1} \times \sum_{i=1}^{n}\sum_{j=1}^{n_{ik}}\begin{pmatrix} Y_{ijk} - \hat{E}\left(\nu_{ik}\right) \\ \left\{Y_{ijk}\hat{E}\left(R_{ijk}\right) - \hat{E}\left(\nu_{ik}R_{ijk}\right)\right\}\boldsymbol{X}_i \\ \left\{Y_{ijk} - \hat{E}\left(\nu_{ik}\right)\right\}t_{ijk} \end{pmatrix},$$

where $\hat{E}(\cdot)$ is the conditional expectation with respect to the observed data. We update $\sigma_{k\epsilon}^2$ by

$$\frac{1}{\sum_{i=1}^{n} n_{ik}} \sum_{i=1}^{n} \sum_{j=1}^{n_{ik}} \left\{ M_{ijk}^2 - 2M_{ijk}\hat{E}\left(\nu_{ik}\right) + \hat{E}\left(\nu_{ik}^2\right) + \boldsymbol{\alpha}_{1k}^{\mathrm{T}}\boldsymbol{X}_i\left(\boldsymbol{\alpha}_{1k}^{\mathrm{T}}\boldsymbol{X}_i - 2M_{ijk}\right)\hat{E}\left(R_{ijk}\right) + 2\boldsymbol{\alpha}_{1k}^{\mathrm{T}}\boldsymbol{X}_i\hat{E}\left(\nu_{ik}R_{ijk}\right) \right\}$$

and update $\sigma_{k\nu}^2$ by $\sum_{i=1}^{n}\hat{E}\left(\nu_{ik}^2\right)/n$. We update $\boldsymbol{\eta}_k$ by solving the equation

$$\sum_{i=1}^{n}\left\{\hat{E}\left(Q_{ik}\right) - \frac{\exp\left(\boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{X}_i\right)}{1+\exp\left(\boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{X}_i\right)}\right\}\boldsymbol{X}_i = 0$$

and update $\mu_k^* \equiv \mu_k b_k$ and b_k by solving the equations

$$\sum_{i=1}^{n}\hat{E}\left[\sum_{j=1}^{n_{ik}}\left(1-R_{ik}\right) - \left(1-Q_{ik}\right)\sum_{j=1}^{n_{ik}}\left\{\frac{g_{ijk}\left(W_i;\mu_k,b_k\right)}{1+g_{ijk}\left(W_i;\mu_k,b_k\right)}\right\}\right] = 0$$

and

$$\sum_{i=1}^{n}\hat{E}\left[\sum_{j=1}^{n_{ik}}\left(W_i - t_{ijk}\right)\left(1-R_{ijk}\right) - \left(1-Q_{ik}\right)\sum_{j=1}^{n_{ik}}\left(W_i - t_{ijk}\right)\left\{\frac{g_{ijk}\left(W_i;\mu_k,b_k\right)}{1+g_{ijk}\left(W_i;\mu_k,b_k\right)}\right\}\right] = 0.$$

We update $\boldsymbol{\theta}$ by $\left(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i\hat{E}\left(W_i\right)$, and update σ_W^2 by

$$n^{-1}\sum_{i=1}^{n}\left\{\hat{E}\left(W_i^2\right) - 2\hat{E}\left(W_i\right)\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}_i + \left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}_i\right)^2\right\}.$$

In the E-step, we evaluate the conditional expectations of $\hat{E}\left(R_{ijk}\right)$, $\hat{E}\left(\nu_{ik}\right)$, $\hat{E}\left(\nu_{ik}^2\right)$, $\hat{E}\left(\nu_{ik}R_{ijk}\right)$, $\hat{E}\left(Q_{ik}\right)$, $\hat{E}\left(W_i\right)$, $\hat{E}\left(W_i^2\right)$, $\hat{E}\left\{\left(W_i - t_{ijk}\right)\left(1-R_{ijk}\right)\right\}$, and

$$\hat{E}\left[\left(1-Q_{ik}\right)\sum_{j=1}^{n_{ik}}\left(W_i - t_{ijk}\right)^{m_1}\frac{g_{ijk}\left(W_i;\mu_k,b_k\right)}{\left\{1+g_{ijk}\left(W_i;\mu_k,b_k\right)\right\}^{m_2}}\right]$$

given the observed data $\mathcal{O}_i$ for m_1 =0,1,2 and m_2 = 1,2. Specifically, the conditional expectation of Q_{ik} given ν_{ik} and W_i is given by

$$\frac{\exp\left(\boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{X}_i\right)\prod_{j=1}^{n_{ik}}A_{ijk}\left(\nu_{ik};\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2\right)}{\exp\left(\boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{X}_i\right)\prod_{j=1}^{n_{ik}}A_{ijk}\left(\nu_{ik};\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2\right) + \prod_{j=1}^{n_{ik}}D_{ijk}\left(\nu_{ik},W_i;\mu_k,b_k,\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2\right)},$$

and the conditional expectation of R_{ijk} is given by

$$\frac{\exp\left(\boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{X}_i\right)\prod_{j'=1}^{n_{ik}}A_{ij'k}\left(\nu_{ik};\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2\right)}{\exp\left(\boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{X}_i\right)\prod_{j'=1}^{n_{ik}}A_{ij'k}\left(\nu_{ik};\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2\right)+\prod_{j'=1}^{n_{ik}}D_{ij'k}\left(\nu_{ik},W_i;\mu_k,b_k,\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2\right)} + \frac{\left\{\frac{A_{ijk}\left(\nu_{ik};\alpha_k,\sigma_{k\epsilon}^2\right)}{g_{ijk}\left(W_i;\mu_k,b_k\right)B_{ijk}\left(\nu_{ik};\alpha_{0k},\sigma_{k\epsilon}^2\right)+A_{ijk}\left(\nu_{ik};\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2\right)}\right\}}{\exp\left(\boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{X}_i\right)\prod_{j'=1}^{n_{ik}}A_{ij'k}\left(\nu_{ik};\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2\right)+\prod_{j'=1}^{n_{ik}}D_{ij'k}\left(\nu_{ik},W_i;\mu_k,b_k,\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2\right)}$$

$$\times\prod_{j'=1}^{n_{ik}}D_{ij'k}\left(\nu_{ik},W_i;\mu_k,b_k,\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2\right).$$

Note that the joint density of (ν_{ik}, W_i) given $\mathcal{O}_i$ is proportional to

$$h_i\left(W_i;\sigma_W^2,\sigma_\delta^2\right)\phi\left(\nu_{ik};\sigma_{k\nu}^2\right)\left\{\frac{\prod_{k'=1}^{K}q_{k'}\left(W_i;\eta_k,\mu_k,b_k,\alpha_{1k},\sigma_{k\epsilon}^2,\sigma_{k\nu}^2\right)}{q_k\left(W_i;\eta_k,\mu_k,b_k,\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2,\sigma_{k\nu}^2\right)}\right\}$$

$$\times\frac{\exp\left(\boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{X}_i\right)\left\{\prod_{j=1}^{n_{ik}}A_{ijk}\left(\nu_{ik};\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2\right)\right\}+\prod_{j=1}^{n_{ik}}D_{ijk}\left(\nu_{ik},W_i;\mu_k,b_k,\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2\right)}{1+\exp\left(\boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{X}_i\right)},$$

and the density of W_i given $\mathcal{O}_i$ is proportional to $\prod_{k=1}^{K}q_k\left(W_i;\eta_k,\mu_k,b_k,\boldsymbol{\alpha}_k,\sigma_{k\epsilon}^2,\sigma_{k\nu}^2\right)h_i\left(W_i;\sigma_W^2,\sigma_\delta^2\right)$. We evaluate the conditional expectations through numerical integration over ν_{ik} and W_i with two-dimensional Gauss-Hermite quadratures. We iterate between the E-step and M-step until convergence.

REFERENCES

Bateman RJ, Xiong C, Benzinger TL, Fagan AM, Goate A, Fox NC, Marcus DS, Cairns NJ, Xie X, Blazey TM et al. (2012). Clinical and biomarker changes in dominantly inherited Alzheimer's disease. N. Engl. J. Med,. 367 795–804. [PubMed: 22784036]

Bernardo JM (1979). Expected information as expected utility. Ann. Stat 7 686–690.

Dubois B, Feldman HH, Jacova C, DeKosky ST, Barberger-Gateau P, Cummings J, Delacourte A, Galasko D, Gauthier S, Jicha G, Meguro K, O'Brien J, Pasquier F, Robert P, Rossor M, Salloway S, Stern Y, Visser PJ and Scheltens P (2007). Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. Lancet Neurol 6 734–746. [PubMed: 17616482]

Fjell AM, Walhovd KB, Fennema-Notestine C, McEvoy LK, Hagler DJ, Holland D, Brewer JB and Dale AM (2009). One-year brain atrophy evident in healthy aging. J. Neurosci 29 15223–15231. [PubMed: 19955375]

Garcia T, Marder K and Wang Y (2017). Statistical modeling of Huntington disease onset. Handbook of Clinical Neurology 144 47–61. [PubMed: 28947125]

Gneiting T, Balabdaoui F and Raftery AE (2007). Probabilistic forecasts, calibration and sharpness. J. R. Stat. Soc. Ser. B 69 243–268.

Good IJ (1952). Rational decisions. J. R. Stat. Soc. Ser. B 14 107–114.

Huntington Study Group (1996). Unified Huntington's Disease Rating Scale: reliability and consistency. Mov. Disorders 11 136–142.

Hall CB, Lipton RB, Sliwinski M and Stewart WF (2000). A change point model for estimating the onset of cognitive decline in preclinical Alzheimer's disease. Stat. Med 19 1555–1566. [PubMed: 10844718]

Hall CB, Ying J, Kuo L, Sliwinski M, Buschke H, Katz M and Lipton RB (2001). Estimation of bivariate measurements having different change points, with application to cognitive ageing. Stat. Med 20 3695–3714. [PubMed: 11782027]

Hall CB, Ying J, Kuo L and Lipton RB (2003). Bayesian and profile likelihood change point methods for modeling cognitive function over time. Comput. Statist. Data Anal 42 91–109.

Hampel H, Bürger K, Teipel SJ, Bokde AL, Zetterberg H and Blennow K (2008). Core candidate neurochemical and imaging biomarkers of Alzheimer's disease. Alzheimer's & Dementia 4 38–48.

Hogan JW and Laird NM (1997). Mixture models for the joint distribution of repeated measures and event times. Stat. Med 16 239–257. [PubMed: 9004395]

Hopfield JJ (1982). Neural networks and physical systems with emergent collective computational abilities. PNAS 79 2554–2558. [PubMed: 6953413]

Jack CR, Knopman DS, Jagust WJ, Shaw LM, Aisen PS, Weiner MW, Petersen RC and Trojanowski JQ (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol 9 119–128. [PubMed: 20083042]

Jacqmin-Gadda H, Commenges D and Dartigues J-F (2006). Random change-point model for joint modeling of cognitive decline and dementia. Biometrics 62 254–260. [PubMed: 16542253]

Jedynak BM, Lang A, Liu B, Katz E, Zhang Y, Wyman BT, Raunig D, Jedynak CP, Caffo B, Prince JL et al. (2012). A computational neurodegenerative disease progression score: method and results with the Alzheimer's Disease Neuroimaging Initiative cohort. Neuroimage 63 1478–1486. [PubMed: 22885136]

Kremer B, Goldberg P, Andrew SE, Theilmann J, Teleniüs H, Zeisler J, Sqüitieri F, Lin B, Bassett A, Almqvist E et al. (1994). A worldwide study of the Huntington's disease mutation: the sensitivity and specificity of measuring CAG repeats. N. Engl. J. Med 330 1401–1406. [PubMed: 8159192]

Little RJ (1995). Modeling the drop-out mechanism in repeated-measures studies. J. Amer. Statist. Assoc 90 1112–1121.

Long JD, Paulsen JS, Marder K, Zhang Y, Kim J-I and Mills JA (2014). Tracking motor impairments in the progression of Huntington's disease. Mov. Disorders 29 311–319.

MacDonald ME, Ambrose CM, Duyao MP, Myers RH, Lin C, Srinidhi L, Barnes G, Taylor SA, James M, Groot N et al. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72 971–983. [PubMed: 8458085]

Nelson PT, Alafuzoff I, Bigio EH, Bouras C, Braak H, Cairns NJ, Castellani RJ, Crain BJ, Davies P, Tredici KD et al. (2012). Correlation of Alzheimer disease neuropathologic changes with cognitive status: a review of the literature. J. Neuropathol. Exp. Neurol 71 362–381. [PubMed: 22487856]

Paulsen J, Long J, Ross C, Harrington D, Erwin C, Williams J, WesterVELT J, Johnson H, Aylward E and Zhang Y (2014a). Prediction of manifest Huntington's disease with clinical and imaging measures: a prospective observational study. Lancet Neurol 13 1193–1201. [PubMed: 25453459]

Paulsen JS, Long JD, Johnson HJ, Aylward EH, Ross CA, Williams JK, Nance MA, Erwin CJ, Westervelt HJ, Harrington DL et al. (2014b). Clinical and biomarker changes in premanifest Huntington disease show trial feasibility: a decade of the PREDICT-HD study. Front. Aging Neurosci 6 1–11. [PubMed: 24478697]

Rizopoulos D (2012). Joint Models for Longitudinal and Time-to-Event Data: With Applications in R Chapman and Hall/CRC, New York.

Rubinsztein DC, Leggo J, Coles R, Almqvist E, Biancalana V, Cassiman J-J, Chotai K, Connarty M, Craufurd D, Curtis A et al. (1996). Phenotypic characterization of individuals with 30–40 CAG repeats In the Huntington disease (HD) gene reveals HD cases With 36 repeats and apparently normal elderly individuals with 36–39 repeats. Am. J. Hum. Genet 59 16–22. [PubMed: 8659522]

Samtani MN, Farnum M, Lobanov V, Yang E, Raghavan N, DiBernardo A and Narayan V (2012). An improved model for disease progression in patients From the Alzheimer's Disease Neuroimaging Initiative. J. Clin. Pharmacol 52 629–644. [PubMed: 21659625]

Shaw LM, Vanderstichele H, Knapik-Czajka M, Clark CM, Aisen PS, Petersen RC, Blennow K, Soares H, Simon A, Lewczuk P et al. (2009). Cerebrospinal fluid biomarker signature in Alzheimer's Disease Neuroimaging Initiative subjects. Ann. Neurol 65 403–413. [PubMed: 19296504]

Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, Iwatsubo T, Jack CR, Kaye J, Montine TJ et al. (2011). Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia 7 280–292.

Tsiatis AA and Davidian M (2004). Joint modeling of longitudinal and time-to-event data: an overview. Stat. Sin 14 809–834.

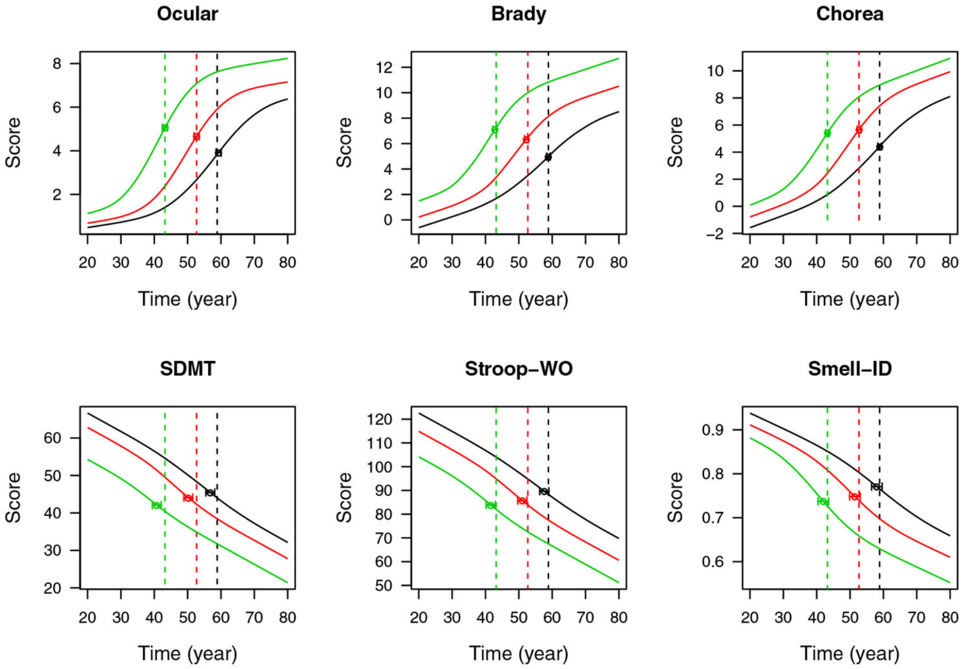

Fig 1.

Average estimated values of biomarkers over age among subgroups of subjects with different lengths of CAG expansion. The black, red, and green curves pertain to the subgroups of subjects with CAG expansion < 41, 41 ≤ CAG expansion < 43, and CAG expansion ≥ 43, respectively. The circles and bars indicate the average inflection points and their 95% confidence intervals. The dashed lines indicate the average HD age at onset. SDMT and Stroop-WO are identified as prognostic biomarkers using the proposed approach.

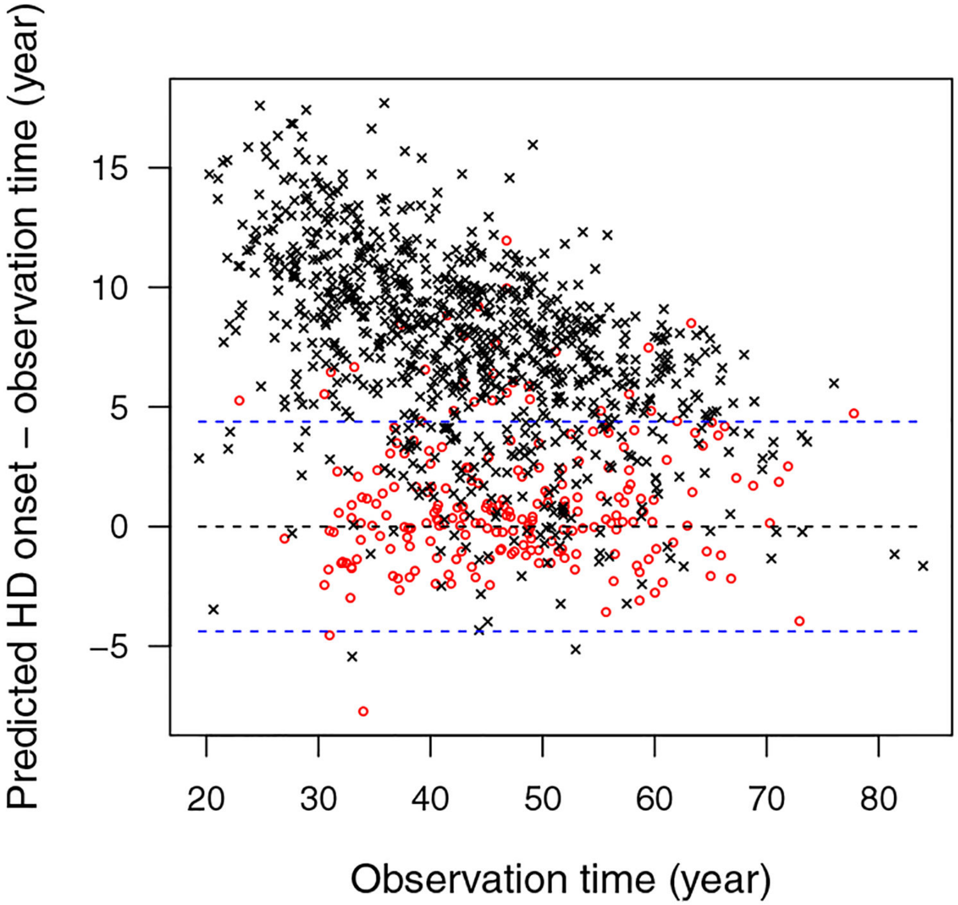

Fig 2.

Difference of the predicted HD age at onset and the observation age versus the observation age in the PREDICT-HD study. The red circles and black crosses pertain, respectively, to the uncensored and censored subjects. The blue dashed lines indicate variability $\pm\sqrt{\sigma_\delta^2 + \sigma_W^2}$.

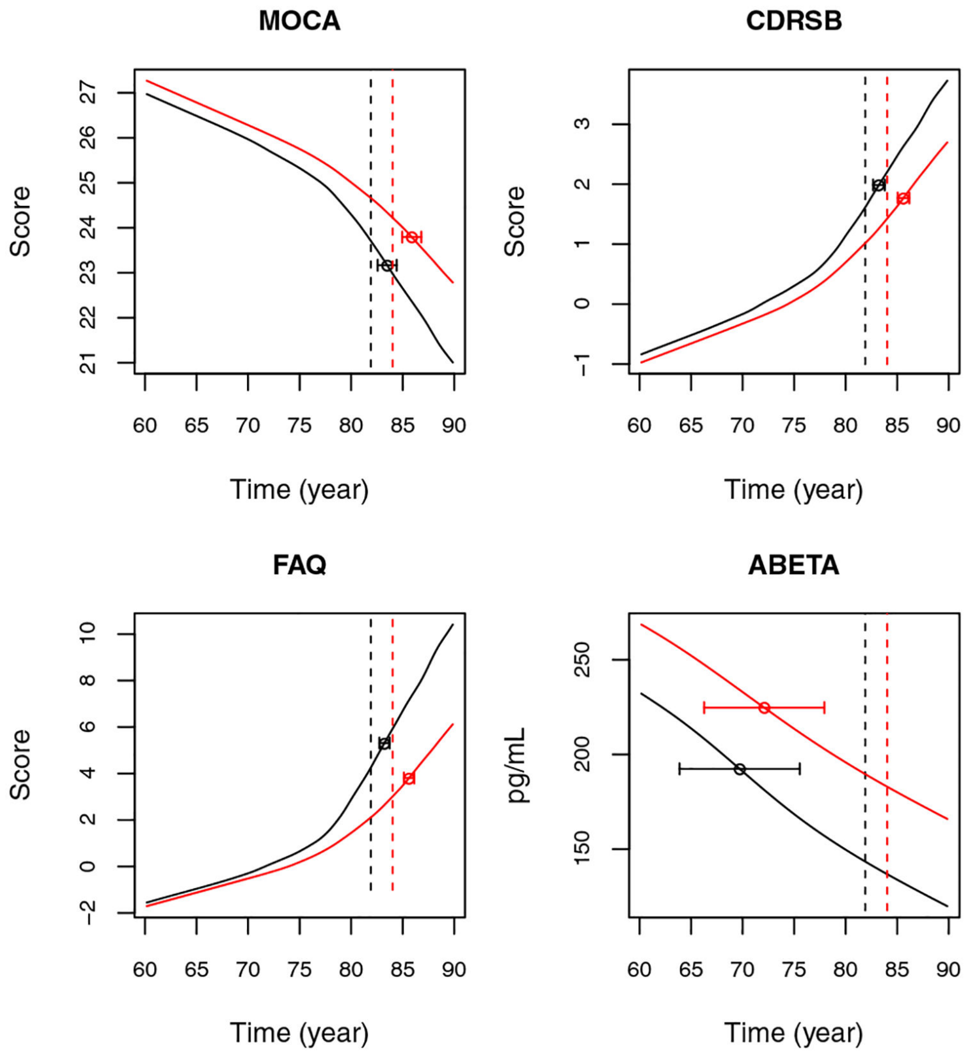

Fig 3.

Average estimated values of biological and clinical markers over age among carriers and non-carriers of APOE ∊4 alleles. The black and red curves pertain to the subgroups of APOE carriers and non-carriers, respectively. The circles and bars indicate the average inflection points and their 95% confidence intervals. The dashed lines indicate the average early AD onset ages. Aβ_{42} is identified as a prognostic biomarker and MOCA, FAQ, and, CDRSB are confirmed as diagnostic markers.

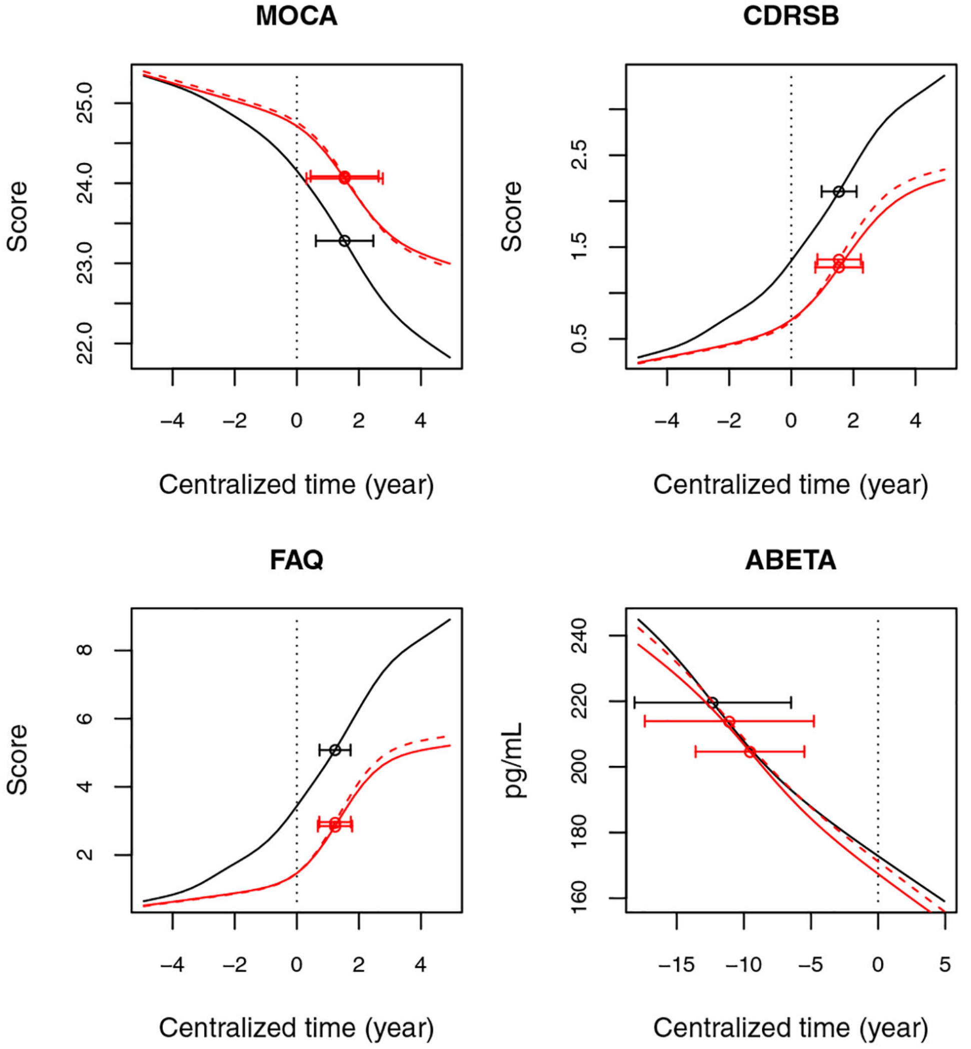

Fig 4.

Average estimated values of biological and clinical markers over centralized age (years to age at onset of early AD). The black curves pertain to the proposed approach with the observed data. The circles and bars indicate the population average peak degeneration ages and their 95% confidence intervals. The red solid and dashed curves pertain, respectively, to imputing censored age at early AD by parent's AD age at onset with a correlation of 0.3 or 0.65 between child's and parent's onset ages.

Table 1

Summary statistics for the proposed estimators in simulations

Parameter	True Value	***n*** **= 200**				***n*** **= 400**			
		Bias	**SE**	**SEE**	**CP**	**Bias**	**SE**	**SEE**	**CP**
α_{01}	0.4	−0.016	0.115	0.122	0.958	0.004	0.086	0.083	0.944
α_{02}	0.6	0.005	0.106	0.107	0.945	0.000	0.076	0.072	0.928
$\boldsymbol{\alpha}_{11}$	1.0	0.020	0.123	0.130	0.957	−0.003	0.089	0.090	0.944
	0.8	−0.005	0.059	0.062	0.956	−0.002	0.043	0.043	0.945
	0.7	−0.006	0.114	0.118	0.954	0.003	0.083	0.082	0.939
$\boldsymbol{\alpha}_{12}$	1.0	0.004	0.125	0.127	0.947	−0.004	0.089	0.088	0.933
	1.2	−0.001	0.063	0.064	0.943	−0.001	0.046	0.044	0.933
	0.8	0.003	0.121	0.125	0.958	0.000	0.084	0.087	0.947
β_1	0.8	0.001	0.017	0.017	0.950	0.000	0.012	0.012	0.953
β_2	−0.4	0.001	0.014	0.014	0.951	0.000	0.010	0.010	0.958
$\sigma_{1\epsilon}^2$	0.5	−0.005	0.031	0.031	0.950	0.000	0.021	0.022	0.956
$\sigma_{2\epsilon}^2$	0.5	−0.003	0.025	0.026	0.957	−0.001	0.018	0.018	0.949
$\sigma_{1\nu}^2$	0.5	−0.009	0.066	0.067	0.956	−0.005	0.048	0.047	0.945
$\sigma_{2\nu}^2$	0.5	−0.014	0.070	0.067	0.933	−0.005	0.049	0.048	0.944
$\boldsymbol{\eta}_1$	−0.5	0.006	0.628	0.751	0.989	−0.014	0.423	0.442	0.974
	0.5	0.004	0.383	0.459	0.989	−0.003	0.259	0.266	0.967
	0.0	−0.012	0.550	0.642	0.990	0.018	0.393	0.395	0.958
$\boldsymbol{\eta}_2$	0.0	−0.011	0.462	0.483	0.979	−0.001	0.300	0.304	0.970
	−0.5	−0.021	0.298	0.320	0.985	−0.001	0.198	0.196	0.966
	0.5	0.000	0.489	0.509	0.971	0.009	0.330	0.329	0.956
μ_1	−1.0	0.003	0.416	0.431	0.972	−0.002	0.290	0.292	0.951
μ_2	1.6	−0.030	0.466	0.455	0.948	0.001	0.292	0.316	0.969
b_1	−0.5	−0.006	0.115	0.116	0.962	−0.003	0.074	0.075	0.954
b_2	0.5	0.006	0.088	0.090	0.964	0.001	0.056	0.061	0.962
$\boldsymbol{\theta}$	3.0	0.001	0.071	0.072	0.953	0.002	0.052	0.051	0.942
	−0.2	0.000	0.052	0.052	0.946	−0.001	0.037	0.036	0.938
	0.2	0.001	0.099	0.103	0.953	−0.002	0.074	0.073	0.946
σ_W^2	0.2	−0.010	0.047	0.044	0.943	−0.008	0.033	0.031	0.942

Table 2

Summary statistics on prediction in simulations

	n = 200			*n* = 400		
Prediction	**Bias**	**Adjusted SD**	**Adjusted LS**	**Bias**	**Adjusted SD**	**Adjusted LS**
Both Biomarkers	0.001	0.079	0.036	−0.002	0.052	0.020
Biomarker 1	0.001	0.084	0.038	−0.002	0.060	0.022
Biomarker 2	0.001	0.085	0.037	−0.002	0.061	0.022

Table 3

Estimation results for selected, parameters in the PREDICT-HD study

	Parameter	Est	SEE	*p*-value
μ_k	Ocular	0.208	0.390	0.593
	Brady	−0.158	0.278	0.570
	Chorea	−0.008	0.275	0.977
	SDMT	−2.194	0.676	0.001
	Stroop-WO	−1.535	0.697	0.028
	Smell-ID	−0.963	0.807	0.232
θ	Intercept	64.23	5.533	<0.0001
	Baseline age	0.738	0.025	<0.0001
	Years of education	0.182	0.071	0.010
	Sex (Male)	0.881	0.416	0.034
	CAG repeats length	−1.090	0.107	<0.0001
σ_W^2		18.89	1.754	<0.0001

Table 4

Estimation results for selected parameters in the ADNI study

	Parameter	Est	SEE	*p*-value
μ_k	MOCA	1.622	0.470	0.0006
	FAQ	1.558	0.287	<0.0001
	CDRSB	1.488	0.255	<0.0001
	ABETA	−12.09	2.973	<0.0001
θ	Intercept	13.14	4.795	0.006
	Baseline age	0.956	0.061	<0.0001
	Gender	−0.310	0.570	0.587
	Education	0.136	0.118	0.248
	APOE ϵ4 allele	−1.339	0.511	0.009
	Baseline Total11	−0.357	0.087	<0.0001
	Baseline FAQ	−1.233	0.326	0.0002
σ_W^2		13.58	2.028	<0.0001