ORIGINAL ARTICLE

Joint analysis for multivariate longitudinal and event time data with a change point anchored at interval-censored event time

Yue Zhan[1] | **Cheng Zheng**[1] | **Ying Zhang**[1]

[1]Department of Biostatistics, College of Public Health, University of Nebraska Medical Center, Omaha, Nebraska, United States

Correspondence
Ying Zhang
Email: ying.zhang@unmc.edu

Funding information
Funder One, Funder One Department, Grant/Award Number: 123456, 123457 and 123458; Funder Two, Funder Two Department, Grant/Award Number: 123459

Huntington's disease is an autosomal dominant neurodegenerative disorder, characterized by motor dysfunction, psychiatric disturbances, and cognitive decline. **The onset of Huntington's disease is diagnosed by severe motor impairment, which can be predicted by cognitive decline and may also consequently worsen cognitive impairment. However, clinical data are often collected at discrete times, and therefore, disease onset is subject to interval censoring.** We develop a joint model of multivariate longitudinal biomarkers with a change point anchored at an interval-censored event time. Our model allows us to simultaneously **study the effect of longitudinal biomarkers on the event time and the changes in trajectories of the longitudinal biomarkers post the event. We conduct a comprehensive simulation study to demonstrate the satisfactory finite-sample performance of the proposed method for making inferences. We apply the method** to the PREDICT-HD data from a multisite observational cohort study of prodromal Huntington's disease individuals to ascertain how cognitive impairment and motor dysfunction interact during disease progression.

Keywords — Joint model, Survival analysis, Longitudinal biomarker, Change point, Interval-censored data

Abbreviations: HD, Huntington's disease; MLE, Maximum Likelihood Estimation; DCL, Diagnose Confidence Level.

1 | INTRODUCTION

Huntington's disease (HD) is **an** inherited, autosomal dominant neurodegenerative disorder characterized by progressive motor dysfunction, cognitive decline, and psychiatric disturbances. The disease is caused by an abnormal expansion of CAG trinucleotide repeats in the *HTT* gene on chromosome 4, which leads to the production of a mutant huntingtin protein that disrupts normal cellular function [31, 25]. **Currently,** there is no **effective** treatment **available** for HD patients.

Huntington's disease encompasses motor, cognitive, and psychiatric manifestations, but the clinical diagnosis of onset is determined **primarily** by the emergence of motor symptoms [22]. **However,** cognitive impairment **has been found to appear** prior to HD motor diagnosis [8, 21, 23, 29, 36]. Zhang et al. [36] have **thoroughly** studied mild cognitive impairment (MCI) in various cognitive domains **characterized** by Harrington et al. [12], **which is regarded** as an early landmark for disease progression in prodromal HD individuals. **Although a rapid cognitive decline at the early stage, attributed to the onset of HD, has been observed [3], an overarching study of how cognitive impairment interacts with HD onset is lacking and is helpful to understand the HD progression in prodromal HD individuals. In this work, we propose a holistic model to uncover the relationship between cognitive decline in multiple domains and HD onset using joint modeling of multivariate longitudinal and event time data.**

Joint modeling of longitudinal and event time data has become increasingly important in biomedical studies, especially in chronic diseases, HIV/AIDS, cardiovascular research, etc, when understanding the interplay between longitudinal biomarkers and clinical events is crucial. The joint modeling framework was developed by simultaneously modeling the longitudinal and event time data under a structure with shared subject-specific random effects [33, 13, 28, 17, 1, 14, 16]. Recent development of the joint modeling method allows researchers to study the causal effect of the exposure on the survival outcome through the longitudinal marker [38, 6]. Previous work showed that by including both the random effect and the longitudinal marker in the survival model, the joint model framework allows the researcher to separate the direct causal effect of the longitudinal marker on the outcome and the unmeasured confounding between the marker and the outcome [37, 18]. The maximum likelihood method equipped with the EM algorithm has been a popular approach to estimate parameters in a joint model [24] for making statistical inferences. It is worth noting that most of the joint models were developed for longitudinal biomarkers and right-censored event time, emphasizing either unbiased inference for event time outcome using time-dependent longitudinal biomarkers [28, 17, 9] or unbiased inference of longitudinal trajectories of the biomarkers subject to informative drop-out [7, 11, 20]. These methods are not applicable to study the relationship between cognitive impairments and the HD onset in prodromal HD individuals, which is subject to interval censoring bracketed by two adjacent motor diagnostic times, where the first time shows negative and the second time yields the motor diagnosis of HD. Although joint models of longitudinal biomarkers and interval-censored event times have been studied more recently [10, 26, 2, 32], the methods did not concern the changes of longitudinal biomarkers triggered by the event time. Some two-phase changing-point analyses of longitudinal data around an interval-censored event time have been explored [35, 4, 5], but they did not address how the longitudinal data impacted the interval-censored event time.

In **this** work, we propose a **two-phase approach by extending the likelihood method for joint modeling of longitudinal and interval-censored event time data to incorporate a potential change point in longitudinal biomarkers anchored at an unobserved event time, depicted in Figure (1). The first phase of the model emphasizes how the longitudinal data impacts the event time, and the second phase investigates whether/how the event time changes the trajectories of longitudinal biomarkers.** We use **an** adaptive Newton-Raphson algorithm to **compute** the **maximum likelihood estimates (MLE)** of all coefficients in this joint model, **and the nonparametric** bootstrap method to **estimate** the standard error of all **estimated** coefficients.

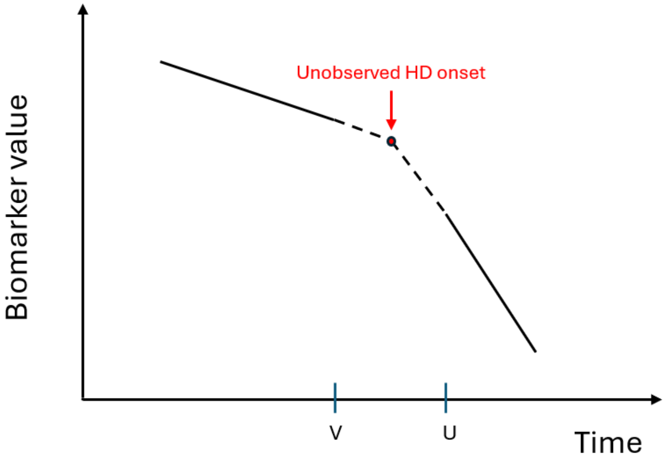

FIGURE 1 A hypothetical model for the HD disease progression

The rest of this paper is organized as follows: in Section 2, we **describe** our notation, models, and **the likelihood** method **for the two-phase joint modeling of changing-point longitudinal data and interval-censored event time and illustrate how to** calculate the MLE. In Section 3, we present simulation studies to evaluate the finite-sample performance of the proposed method. In Section 4, we apply our method to the **data from** the Neurobiological Predictors of Huntington's Disease Study (PREDICT-HD), a 12-year **multisite** prospective cohort study conducted between September 2002 and April 2014 [22] to **examine the interplay between cognitive decline and the HD onset in prodromal HD individuals that cognitive declines predict the HD onset and the later accelerates cognitive declines**. In Section 5, we summarize our findings and discuss potential extensions. The technical details are included in the supplemental materials.

2 | METHODS

2.1 | Notation

First, we define notation for the observed data. We consider K-dimensional longitudinal biomarker processes $M(t) = (M^1(t), M^2(t), \cdots, M^K(t))$ that are potentially related to the onset time of disease E, which is interval censored by $(V, U]$ via periodic diagnoses, where V and U can be zero and infinity, respectively. Let $X(t)$ be covariate processes that are potentially related to both longitudinal disease biomarkers and the onset time of disease. Assume the processes are assessed longitudinally at m random observation times $\underline{T} = (t_1, t_2, \cdots, t_m)$, and the data from the processes collected at $\underline{T}$ are denoted as $\underline{M} = (M(t_1), M(t_2), \cdots, M(t_m))$ and $\underline{X} = (X(t_1), X(t_2), \cdots, X(t_m))$. The observed data for an individual consist of $D = (m, \underline{T}; \underline{M}, \underline{X}; V, U)$, and we have a total of n i.i.d. copies of D, denoted by $\underline{D} = (D_1, D_2, \cdots, D_n)$.

2.2 | Two-phase joint model

In studying disease progress, we hypothesize that (1) the longitudinal disease biomarkers $M(\cdot)$ are predictive of the onset time, E; (2) the onset of the disease may alter the trajectories of the biomarkers. To accommodate this hypothesis, we propose to study the disease progression in two phases, described as follows:

$$M_1^k(t) = X(t)^\top \beta^k + Z(t)^\top a^k + \epsilon^k(t), \tag{1}$$

$$M_2^k(t) = M_1^k(t) + \gamma^k I(t > E)(t - E), \tag{2}$$

$$\lambda(t) = \lambda_0(t) \exp[\theta_x^\top X(t) + \theta_a^\top a^k + \theta_M^\top M_1(t)], \tag{3}$$

for $k = 1, 2, \cdots, K$, where $M_1^k(\cdot)$ and $M_2^k(\cdot)$ denote the first **(before E)** and second **(after E) phases** of **the** k**th longitudinal biomarker** process, $M^k(\cdot)$, respectively.

[The paragraph above is still a bit confusing since equation (2) requires a definition of $M_1(t)$ with t after E. I tried to modify this paragraph below:]

In studying disease progress, we hypothesize that (1) the longitudinal disease biomarkers $M(\cdot)$ are causally affect the onset time, E; and (2) the onset of the disease may alter the trajectories of the biomarkers. To accommodate this hypothesis, we propose to study the disease progression in two phases, described as follows:

$$M_\infty^k(t) = X(t)^\top \beta^k + Z(t)^\top a^k + \epsilon^k(t), \tag{4}$$

$$\lambda(t) = \lambda_0(t) \exp[\theta_x^\top X(t) + \theta_a^\top a^k + \theta_M^\top M_1(t)], \tag{5}$$

$$M_e^k(t) = M_\infty^k(t) + \gamma^k I(t > e)(t - e), \tag{6}$$

for $k = 1, 2, \cdots, K$, where $M_\infty^k(\cdot)$ denote the kth potential longitudinal biomarker process shall the disease never happened and $M_e^k(\cdot)$ denote the kth potential longitudinal biomarker process shall the disease happened at time e. By consistent assumption, the observed kth longitudinal biomarker process will satisfy $M^k(t) = M_E^k(t)$ and we can separate it into two phase for likelihood calculation, where $M_1^k(t), t \in (0, E)$ and $M_2^k(t), t \in [E, \infty)$ denote the first (before E) and second (after E) phases of the kth longitudinal biomarker process, $M^k(\cdot)$, respectively.

In the two-phase longitudinal model, β^k **represent** the effects of the covariates **on the** k**th biomarker**; $a = (a^1, a^2, \cdots a^K) \sim N(0, \Sigma_a)$ **represent the** random effects at the individual level, which are used to surrogate time-independent unmeasured confounding between the longitudinal **biomarkers** and the survival outcome; **$Z(\cdot)$ are the covariates of longitudinal processes corresponding to the random effects, which are often a part of $X(\cdot)$ in practice of the joint modeling; $\epsilon(t) = (\epsilon^1(t), \epsilon^2(t), \cdots, \epsilon^K(t)) \sim N(0, \Sigma_e)$ are i.i.d. time-independent random variations for observed longitudinal data; γ^k represents the changing-point effect anchored at the interval-censored disease onset time, E, on the kth longitudinal biomarker. For the survival model, $\lambda_0(t)$ represents** an unspecified baseline hazard function for the **onset time**; θ_x **represents the effects of the covariates on the onset time; θ_a represent effects of the random effects in the survival model, which reflect whether unmeasured baseline variables confound the association between longitudinal biomarkers and survival outcome; θ_M represent the** effects of longitudinal biomarkers at the first phase on the onset time.

The proposed two-phase joint model is a natural extension of the causal joint models for longitudinal and survival data [18, 37], which did not look at how disease onset impacts the trajectories of longitudinal biomarkers and only dealt with right-censored disease onset time. With an interval-censored disease onset time and two-phase longitudinal models, the likelihood of the observed data is much more complicated, with more involved numerical

challenges in computation.

2.3 | The likelihood for the observed data

In this subsection, we construct the likelihood function of the parameters $\Theta = (\lambda_0(\cdot), \beta, \gamma, \theta_X, \theta_a, \theta_M, \Sigma_a, \Sigma_e)$ of the two-phase joint models (1)-(3) for the observed data $\underline{D}$, where $\beta = (\beta^1, \beta^2, \cdots, \beta^K)$ and $\gamma = (\gamma^1, \gamma^2, \cdots, \gamma^K)$. As a conventional approach in joint modeling with shared random effects, we assume that given the shared random variables a, the observed data on the longitudinal biomarkers are independent within the same biomarker process and between different biomarker processes, as well as independent of the disease onset time.

First, we construct the likelihood for the data, in which the disease onset time is exactly known. We separate the observed biomarker profile $\underline{M}$ to two phases, $\underline{M}_1 = (M(t_1), M(t_2), \cdots, M(V))$ (before E) and $\underline{M}_2 = (M(U), \cdots, M(t_m))$ (after E). The joint model with shared random effects (1)-(3) gives

$$f(E, \underline{M}|m, \underline{T}, \underline{X}, a) = f(\underline{M}_2|E, m, \underline{T}, \underline{X}, a) f(E|\underline{M}_1, m, \underline{T}, \underline{X}, a) f(\underline{M}_1|m, \underline{T}, \underline{X}, a), \tag{7}$$

where

$$\begin{aligned} f(E|\underline{M}_1, m, \underline{T}, \underline{X}, a) =& \lambda_0(E) \exp(\theta_X^\top X(E) + \theta_a^\top a + \theta_M^\top M_1(E)) \\ & \times \exp(\{-\sum_{j:t_j \le E} \int_{t_{j-1}\wedge E}^{t_j \wedge E} \lambda_0(s) \exp(\theta_X^\top X(t_{j-1}) + \theta_a^\top a + \theta_M^\top M_1(t_{j-1}))ds\}), \end{aligned}$$

for which we extrapolate the values of the processes $X(\cdot)$ and $M_1(\cdot)$ between the two observation times t_{j-1} and t_j by their values at t_{j-1}, respectively, because the processes are only observed at the discrete times, $\underline{T}$,

$$\begin{aligned} & f(\underline{M}_1|m, \underline{T}, \underline{X}, a) = \\ & \qquad \prod_{j:t_j<E} \Big[\frac{1}{\sqrt{(2\pi)^K \det(\Sigma_e)}} \exp(-\frac{1}{2}(M_1(t_j) - X(t_j)^\top \beta - Z(t_j)^\top a)^\top \Sigma_e^{-1} \\ & \qquad \times (M_1(t_j) - X(t_j)^\top \beta - Z(t_j)^\top a)) \Big], \end{aligned}$$

and

$$\begin{aligned} & f(\underline{M}_2|E, m, \underline{T}, \underline{X}, a) = \\ & \qquad \prod_{j:t_j \ge E} \Big[\frac{1}{\sqrt{(2\pi)^K \det(\Sigma_e)}} \exp(-\frac{1}{2}(M_2(t_j) - X(t_j)^\top \beta - Z(t_j)^\top a - \gamma(t_j - E))^\top \Sigma_e^{-1} \\ & \qquad \times (M_2(t_j) - X(t_j)^\top \beta - Z(t_j)^\top a - \gamma(t_j - E))) \Big]. \end{aligned}$$

Since E is bracketed by the interval $(V, U]$ in our study of the disease progression for prodromal HD subjects, where U is possibly infinity for subjects who had not been diagnosed at the end of the follow-up, we introduce a binary indicator, Δ, such that $\Delta = 1$ for the case that E is within $(V, U]$ made by two subsequent observation time and $\Delta = 0$

for E being right censored at the end of follow-up, i.e. $V = t_m$. Hence,

$$f(\underline{M}, V < E \le U | m, \underline{T}, \underline{X}, a) = \\ \left[\int_V^U f(\underline{M}_2 | t, m, \underline{T}, \underline{X}, a) f(t | \underline{M}_1, m, \underline{T}, \underline{X}, a) dt\right]^{\Delta} S(V | \underline{M}_1, m, \underline{T}, \underline{X}, a)^{1-\Delta} f(\underline{M}_1 | m, \underline{T}, \underline{X}, a),$$

where

$$S(V | \underline{M}_1, m, \underline{T}, X, a) = \exp\left\{-\sum_{j: t_j \le V} \int_{t_{j-1}}^{t_j} \lambda_0(s) \exp(\theta_x^\top X(t_{j-1}) + \theta_a^\top a + \theta_m^\top M_1(t_{j-1})) ds\right\}.$$

Assuming that the number of observations, observation times, and the covariate processes are noninformative to the model parameters Θ, the likelihood for D as a function of Θ is

$$L(\Theta; D) = \int \left[\int_V^U f(\underline{M}_2 | t, m, \underline{T}, \underline{X}, a) f(t | \underline{M}_1, m, \underline{T}, \underline{X}, a) dt\right]^{\Delta} S(V | \underline{M}_1, m, \underline{T}, \underline{X}, a)^{1-\Delta} f(\underline{M}_1 | m, \underline{T}, \underline{X}, a) f(a; \Sigma_a) da,$$

where

$$f(a; \Sigma_a) = (2\pi)^{-K/2} (\det(\Sigma_a))^{-1/2} \exp(-\frac{1}{2} a^\top \Sigma_a^{-1} a),$$

and the likelihood for the observed data $\underline{D}$ is

$$L(\Theta; \underline{D}) = \prod_{i=1}^n L(\Theta; D_i) = \prod_{i=1}^n \int \left[\int_{V_i}^{U_i} f(\underline{M}_{i2} | t, m_i, \underline{T}_i, \underline{X}_i, a_i) f(t | \underline{M}_{i1}, m_i, \underline{T}_i, \underline{X}_i, a_i) dt\right]^{\Delta_i} \\ \times S(V_i | \underline{M}_{i1}, m_i, \underline{T}_i, \underline{X}_i, a_i)^{1-\Delta_i} f(\underline{M}_{i1} | m_i, \underline{T}_i, \underline{X}_i, a_i) f(a_i; \Sigma_a) da_i. \tag{8}$$

We implemented a sieve seminparametric maximum likelihood method to estimate the model parameters Θ by approximating the cumulative baseline hazard function $\Lambda_0(\cdot) = \int_0^t \lambda_0(u) du$ using the cubic monotone B-spline

$$\Lambda_0(t) = \sum_{j=1}^{q_n} \alpha_j B_j(t),$$

where $\alpha_j = \sum_{l=1}^{j} \exp(\xi_l)$ and $\{\xi_j\}_{j=1}^{q_n}$ without restrictions. In this way, we can enforce $\{\alpha_j\}_{j=1}^{q_n}$ to be non-negative and monotone increasing. The monotonicity of the B-spline coefficients warrants the monotone non-decreasing property of $\Lambda_0(\cdot)$ [27]. The spline-based semiparametric sieve maximum likelihood estimation method has been widely employed for many semiparametric models [19, 34, 15, 30].

2.4 | Numerical Algorithm

For the B-spline-based sieve semi-parametric maximum likelihood estimation, the model parameters become

$$\Theta = (\xi, \beta, \gamma, \theta_X, \theta_a, \theta_M, \Sigma_a, \Sigma_e)$$

with $\xi = (\xi_1, \xi_2, \cdots, \xi_{q_n})$ **to make** $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_{q_n})$ **non-negative and monotone. We adopted a general Fisher scoring algorithm with a line search procedure to compute the estimates:**

$$\Theta^{(p+1)} = \Theta^{(p)} - \eta \mathcal{I}^{-1}(\Theta^{(p)}) \dot{l}(\Theta^{(p)}),$$

where $\dot{l}(\Theta^{(p)}) = \nabla_\Theta l(\Theta^{(p)}) = \nabla_\Theta \log L(\Theta^{(p)}; \underline{D})$ **represents the score of the log likelihood function, and** $\mathcal{I}(\Theta^{(p)}) = \sum_{i=1}^{n} \dot{l}_i(\Theta^{(p)}) \dot{l}_i(\Theta^{(p)})^T$ **represents the observed information, where** $\dot{l}_i(\Theta^{(p)}) = \nabla_\Theta l_i(\Theta^{(p)}) = \nabla_\Theta \log L(\Theta; D_i)$**.** η **is an adaptive step length when updating the parameters. We use a line search strategy for deciding** η**. Stop criterion for this algorithm is that maximum of relative differences of parameters is less than** 10^{-3}**, i.e.** $\max_\Theta |\Theta^{(p+1)} - \Theta^{(p)}| / |\Theta^{(p)}| < 10^{-3}$**.**

Although the algorithm is clear and easy to understand, there are several numerical and computational issues when implementing it. The first challenge is that there is no closed form for integration between interval (V_i, U_i): $\int_{V_i}^{U_i} f(\underline{M}_{i2}^k | t, m_i, \underline{T}_i, \underline{X}_i, a_i) f(t | \underline{M}_{i1}, m_i, \underline{T}_i, \underline{X}_i, a_i) dt$, since the function to be integrated includes an exponential of cubic B-spline times a polynomial w.r.t. time t. To calculate this integral as accurately as possible, we choose to use Gauss-Legendre quadrature, which is developed for approximating the bounded integral of a smooth function. We use 20 nodes between (V_i, U_i), calculate values of this function at these nodes and then times the weights of these nodes to get a numerical approximation of this integral. The second problem is that there is also no explicit form of the integration over random effects a. We use Gauss-Hermite quadrature, which is developed to approximate the value of integrals of functions with form $e^{-x^2} f(x)$. We still use 20 nodes of random effects to approximate the likelihood. The third problem is that it's hard to get the analytical form of score function of each parameter. We use numerical differentiation with $\dot{l}_{\Theta_k}(\Theta^{(p)}) = \frac{l(\Theta^{(p)} + \delta \cdot e_k) - l(\Theta^{(p)} - \delta \cdot e_k)}{2\delta}$ as score function, where Θ_k is the k-th element of Θ, e_k is a vector with 1 as k-th element and 0 as other elements, δ is a very small number with magnitude 10^{-6}. The last problem is that the eigenvalues of observed information $\mathcal{I}(\Theta^{(p)})$ could be very small, making it hard to get the inverse of information. We use g-inverse matrix of $\mathcal{I}(\Theta)$ instead, which does not require the original matrix to be invertible.

For the initial guess of coefficients, we estimate initial coefficients of longitudinal model and survival model in order. First, we fit longitudinal models for each biomarker with fixed and random effects and the time effect after event. By using lme4 package in R we can get longitudinal coefficients and the value of estimated random effects. Then we fit a survival model treating $\frac{U+V}{2}$ as a right-censored event time and imputing random effects from longitudinal model to get coefficients of covariates in survival model. Then we use JM packages to combine longitudinal model and survival model to get an estimate of all coefficients except baseline hazard using estimated parameters we got previously. To give an initial guess of B-spline cumulative hazard, we first pre-specify the number of knots q_n. Then we find corresponding quantiles $t_1, t_2, \ldots, t_{q_n - 4}$ of set $\{U_i, V_i | i = 1, \ldots, n\}$ as interior knots. Next we use Maximum Likelihood Estimation (MLE) method for interval-censored survival data and *optim* function in R to get the initial guess of ξ.

To get estimated standard error of all parameters, we use Bootstrap method with $B = 50$ Bootstrap size. It's a powerful non-parametric way to estimate the standard error when there is no explicit form of it.

3 | SIMULATION

3.1 | Settings

We conduct a simulation study to demonstrate the performance of our two-phase model. First, the joint model for simulation can be written as following:

$$\begin{aligned} M_{1i}^k(t) &= \beta_0^k + \beta_1^k t + \beta_2^k X_i + a_i^k + \epsilon_i^k(t), \\ M_{2i}^k(t) &= M_{1i}^k(t) + \gamma^k I(t > E_i)(t - E_i), \\ \lambda_i(t) &= \lambda_0(t) \exp(\theta_x^\top X_i + \theta_a^\top a_i + \theta_m^\top M_{1i}(t)), \end{aligned}$$

where $k = 1, 2$ denotes the dimension of biomarkers. We use 2 biomarkers to simulate the longitudinal model. For fixed effect, we use $X_i \sim N(0, 0.8)$ as a time-independent one-dimensional covariate. For random effects a_i, we include one random intercept for each biomarker. The two random effects are from a joint normal distribution $a_i \sim N(0, \Sigma_a)$.

We use $\beta_0 = (0.1, 0.1)^\top, \beta_1 = (0.1, 0.1)^\top, \beta_2 = (0.1, 0.1)^\top, \gamma = (0.1, 0.1)^\top, \theta_x = 0.1, \theta_a = (0.1, 0.1)^\top, \theta_m = (0.1, 0.1)^\top, \Sigma_a = 0.8 \cdot I_2, \Sigma_e = 0.8 \cdot I_2$ as true parameters. And we use a nonlinear cumulative hazard $\Lambda_0(t) = \int_0^t \lambda_0(u)du = (0.2t)^{1.5}$ to generate the data. We generate 11 observations for each individual, starting from $t_0 = 0$. Observation times $\underline{T}_i = \{t_{ij}\}_{j=1}^{10}$ are independently generated from uniform distribution $U(j - 0.2, j + 0.2), j = 1, 2, \ldots, 10$. First-phase biomarkers $\underline{M}_{1i} = \{M_{1i}(t_{ij})\}_{j=0}^{10}$ are generated by the first equation above. Event time E_i is generated by inverse distribution of $F_i(t) = 1 - S_i(t)$, where $S_i(t) = \exp(-\Lambda_i(t)) = \exp(-\int_0^t \lambda_i(s)ds)$. Second-phase biomarkers $\underline{M}_{2i}$ are generated by the second equation above.

When approximating the cumulative hazard, we use 3-degree B-spline hazard with 6 and 10 interior knots (results of 6 knots to be added) to see if the number of interior knots will have effects on estimates. We run 1000 repeated trials with sample size = 400 and 800 respectively. Estimated bias, SD (standard deviation), ASE (Average Bootstrap Standard Error) and CP (Coverage Probability) are reported in the results. Since repeated trials and Bootstrap method are time-consuming, we run the simulation on Holland Computer Center (HCC) server. All codes are written by R module 4.4.

3.2 | Results

Table 1 are the results of using B-spline cumulative hazard with 10 interior knots. The bias of coefficients in longitudinal model are controlled at a very low level. However, the bias of coefficients in survival model are larger. The reason may be that we use piecewise constant hazard to approximate the non-linear true hazard. And this family of appximation function is not in the family of true hazard function. Standard deviation of coefficients and Bootstrap standard error are very close, indicating that the estimation of standard error is reasonable. And the coverage probability is close to 0.95, indicating our estimation is consistent. When sample size doubles, standard deviation and Bootstrap standard error both decreases in $\frac{1}{\sqrt{2}}$ scale, indicating that the estimation is consistent.

Figures 2 and 3 are the cumulative hazard plots using knots = 10. The red and blue lines are 2.5% and 97.5% percentiles of cumulative hazard respectively. The black curve is the true cumulative hazard. The green line is mean cumulative hazard averaged over 1000 trials. We can see that mean cumulative hazard is very close to true cumulative hazard and 95% confidence interval shrinks when sample size increases, indicating that our estimation is consistent.

From the simulation we can see that our model can estimate the coefficients of two-phase joint model with a low bias. The estimation is consistent and good for inference.

n=400	Parameters	Bias	SD	ASE	CP
Biomarker 1	β_0	-.000	0.055	0.057	0.944
	β_1	.000	0.009	0.009	0.953
	β_2	-.000	0.057	0.058	0.946
	γ	-.000	0.013	0.013	0.954
Biomarker 2	β_0	0.001	0.056	0.057	0.939
	β_1	.000	0.009	0.009	0.939
	β_2	-.000	0.056	0.058	0.938
	γ	0.001	0.013	0.013	0.944
Survival Model	θ_x	0.002	0.062	0.062	0.941
	θ_{m1}	-0.002	0.062	0.065	0.949
	θ_{m2}	0.002	0.064	0.065	0.956
	θ_{a1}	0.009	0.100	0.103	0.953
	θ_{a2}	0.003	0.101	0.103	0.960
n=800	Parameters	Bias	SD	ASE	CP
Biomarker 1	β_0	-0.001	0.041	0.040	0.940
	β_1	.000	0.006	0.006	0.942
	β_2	-0.001	0.040	0.040	0.941
	γ	-.000	0.009	0.009	0.933
Biomarker 2	β_0	-0.001	0.039	0.040	0.942
	β_1	.000	0.006	0.006	0.947
	β_2	0.002	0.038	0.041	0.954
	γ	-.000	0.009	0.009	0.955
Survival Model	θ_x	0.003	0.042	0.043	0.948
	θ_{m1}	-0.001	0.046	0.046	0.939
	θ_{m2}	0.001	0.045	0.046	0.941
	θ_{a1}	0.004	0.071	0.072	0.949
	θ_{a2}	0.003	0.070	0.072	0.947

TABLE 1 Simulation results for joint model using B-spline cumulative hazard with 10 interior knots. All the bias of coefficients are at a very low level. Non-parametric Bootstrap estimation for standard deviation is close to empirical standard deviation. And coverage probability (CP) is close to 0.95. When sample size increases from 400 to 800, the standard deviation becomes $\frac{1}{\sqrt{2}}$ times, demonstrating the consistency.

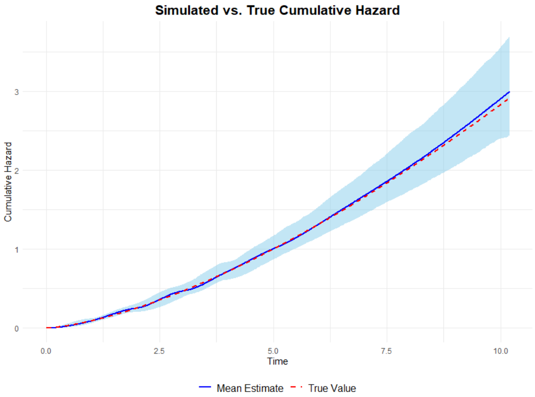

FIGURE 2 Estimated cumulative hazard with interior knots=10, n=400, true $\Lambda_0(t) = (0.2t)^{1.5}$

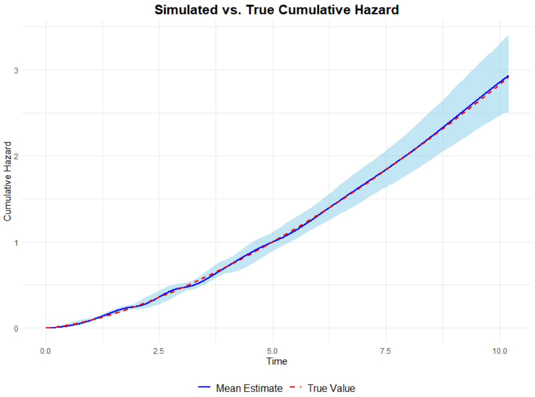

FIGURE 3 Estimated cumulative hazard with interior knots=10, n=800, true $\Lambda_0(t) = (0.2t)^{1.5}$

4 | APPLICATION ON PREDICT-HD DATA

In this section, we apply our model on PREDICT-HD data. We want to test two main hypotheses for this study: one is that cognitive decline predicts the HD onset, the other is that the HD onset accelerates the cognitive decline. First, let's introduce the structure of this data briefly. As longitudinal data, Cognitive biomarkers, time-dependent covariates and observation time, are collected at each observation. Cognitive biomarkers are many types, such as attention, information integration, speed and inhibition. There are also time-independent covariates, such as age, education years, gender, and the baseline HD disease burden (CAP). As survival data, DCL (Diagnose Confidence Level), an integer that can indicate the event, is also collected discretely at each observation. If DCL < 4 at one observation and ≥ 4 at its following observation, then we consider that the event happened between these two observations. Now we have longitudinal data and interval-censored survival data together.

The main challenge we mentioned previously is that we need to handle the change point of biomarkers and interval-censored event time at the same time. We choose 2 biomarkers, Symbol Digit Raw Score Total (sydigtot) and Stroop Word Reading Total (stroopwo) for our analysis. And we choose 4 time-independent covariates: age, education years, gender, and baseline HD disease burden (CAP). We divide CAP score by 100 to let the scale be similar as other covariates.

There are some missing data in either covariates' columns or biomarkers' columns. For those missing completely by individual, we directly delete the individual. For those missing at some observations, if missing happens before event, we use the value before the observation to fill in. If missing happens after event, we use the value after the observation to fill in. We drop individuals diagnosed with HD at the first observation, since we can't capture the change-point by these individuals. We also drop individuals with only 1 observation, since they don't contribute to our longitudinal model. By using this imputation and drop mechanism, we finally have 983 prodromal-HD individuals included, with the number of observations ranging from 2 to 12, resulting in a total of 5694 observations.

Table 2 is the result of longitudinal model using 3-degree cumulative hazard with interior knots = 10. Effects of all variables for two longitudinal biomarkers are similar. All variables except gender in both biomarkers and age in stroopwo are significant. Effects of time on the two biomarkers are both negative, indicating that the two biomarkers decrease as time goes on. Effects of time after diagnosis are also negative, indicating that HD onset accelerates the cognitive decline. Coefficients of age for sydigtot are significantly negative, indicating that larger age leads to lower Symbol Digit Raw Score Total. Coefficients of education years for both biomarkers are positive, indicating that more education years leads to higher cognitive scores. Coefficients of CAP score for both biomarkers are negative, indicating that higher CAP score leads to lower cognitive scores. These findings are corresponding to our expectation.

Table 3 is the result of survival model using 3-degree cumulative hazard with interior knots = 10. In survival model, gender, CAP score and sydigtot are significant, while age, education years and stroopwo are non-significant. However, the p-value of stroopwo is very close to 0.05 so we can consider stroopwo as marginally significant in our model. Coefficients of random effects are not our interests so they are not shown in the table. Hazard ratio of CAP score is larger than 1, indicating that larger CAP score leads to higher hazard to get HD. Hazard ratios of both biomarkers are smaller than 1, indicating that lower cognitive scores lead to higher hazard to get HD. These findings are also as our expected. Hazard ratio of gender is smaller than 1, indicating that males tend to be less likely to get HD than female holding other covariates constant.

	Coefficients	Estimate	Bootstrap 95% CI*		p-value
			Lower	Upper	
Longitudinal model for sydigtot	Intercept	57.807	51.287	64.327	<.0001
	Time	-0.550	-0.629	-0.470	<.0001
	Age	-0.141	-0.230	-0.051	0.0021
	Education	1.351	1.051	1.650	<.0001
	Gender	-1.812	-3.944	0.319	0.0956
	CAP/100	-5.560	-6.331	-4.789	<.0001
	Time after diagnosis	-1.288	-1.651	-0.925	<.0001
Longitudinal model for stroopwo	Intercept	102.676	93.154	112.198	<.0001
	Time	-0.846	-0.976	-0.716	<.0001
	Age	-0.063	-0.177	0.052	0.2836
	Education	1.414	0.858	1.971	<.0001
	Gender	-0.386	-3.170	2.399	0.7860
	CAP/100	-6.212	-7.525	-4.900	<.0001
	Time after diagnosis	-1.288	-2.827	-1.604	<.0001
*CI: Confidence Interval					

TABLE 2 Application results of longitudinal model using B-spline cumulative hazard with 10 interior knots. Gender is not significant for both biomarkers. Baseline age is not significant for stroopwo. Other variables are all significant for both two biomarkers.

	Coefficients	HR*	Bootstrap 95% CI		p-value
			Lower	Upper	
Survival model	Age	0.998	0.983	1.012	0.7447
	Education	1.010	0.944	1.087	0.7195
	Gender	0.735	0.541	0.999	0.0494
	CAP/100	2.174	1.594	2.965	<.0001
	sydigtot	0.956	0.922	0.991	0.0138
	stroopwo	0.981	0.961	1.001	0.0628
*HR: Hazard Ratio					

TABLE 3 Application results of survival model using B-spline cumulative hazard with 10 interior knots. Age, education are not significant in survival model. Stroopwo is marginally significant in survival model.

5 | CONCLUSION AND DISCUSSION

From simulation and application results, our two-phase joint model can successfully uncover the disease progression for prodromal-HD individuals. It models the change point of longitudinal biomarkers and deals with interval-censored survival data simultaneously. The model can control the bias of coefficients at a very low level by using 3-degree B-spline cumulative hazard function. The MLE of our model shows consistency with around 95% coverage probability and decrease of SD by proportion as sample size increases. The similar SD and ASE shows the rationality of Bootstrap method.

The application part shows that both cognitive functions (attention and information integration/speed and inhibition) deteriorate over time as the disease progresses and are significantly predictive of the HD onset. And the two cognitive functions deteriorate more rapidly after the motor onset of HD. As expected, baseline disease burden (CAP) are negatively associated with the cognitive functions, while education is positively associated with the cognitive functions. Baseline age has negative effect on Symbol Digit Raw Score Total, which is also corresponding to our common concept. These findings are new and informative to the HD research community for characterizing the HD progression.

There are some other advantages of our model. Although the covariates we use in simulation and application are all time-independent, the model can be easily generalized to handle time-dependent covariates. We just need to treat time-dependent covariates as constant between each observation, i.e. piecewise constant. Besides, the covariance structure of the random effect and error term can be essentially specified in any commonly-used types, such as unstructured, compound symmetry, AR(1), etc. In our simulation and application, we just use unstructured covariance. However, if we can prespecify a covariance structure by some prior knowledge, it may decrease the number of parameters in the model, then improve the performance and speed of our algorithm. Besides, the model of second phase can be generalized as the following form:

$$M_2(t) = M_1(t) + \gamma^\top I(t > E)g(t - E),$$

where $g()$ is a predetermined smooth function of time after event. Since the function to be integrated between interval (V, U) : $f(M_2|t, X, Z, a; \beta, \gamma, \Sigma_e)f(t|M_1, X, Z, a; \theta, \lambda))$ is still smooth, we can always use Gauss-Legendre quadrature to approximate the integration.

There are also some limitations of our model. We only add one random intercept into the linear mixed effect model for longitudinal biomarkers, which imposes a strong assumption that all individuals' cognitive decline at the same speed by time. Later we will solve this problem by adding a random slope of time in the first phase of the longitudinal model. Besides, while the fisher scoring algorithm works in this problem, it requires a fair amount of numerical effort. We use numerical differentiation in calculating score function, which may decrease the speed of calculation and introduce some bias. More research on efficient algorithms for computing such a complicated model is needed.

Another limitation is that we assume no measurement error in our longitudinal model for biomarkers so that the residual reflects the true biomarker variation over time that will potential cause the disease risk change. In real case, it is likely that only part of this residual term reflects the true variation while the remaining part reflect a measurement error. Then we shall exclude the measurement error part in the hazard model by using an error-free version $M_1^{k*}(t)$ to repalce the observed marker $M_1^k(t)$. This will generalize our model to the following:

$$\begin{aligned}
M_1^{k*}(t) &= X(t)^\top \beta^k + Z(t)^\top a^k + \epsilon^k(t),\\
M_1^k(t) &= M_1^{k*}(t) + e^k(t),\\
M_2^k(t) &= M_1^k(t) + \gamma^k I(t > E) g(t - E),\\
\lambda(t) &= \lambda_0(t) \exp[\theta_x^\top X(t) + \theta_a^\top a^k + \theta_M^\top M_1^*(t)],
\end{aligned}$$

for $k = 1, 2, \ldots, K$, where $M_1^{k*}(t)$ denotes the true value of k-th biomarker at time t in the first phase, which is unobserved in data. $M_1^k(t), M_2^k(t)$ denote the observed k-th biomarker's value at time t. $e^k(t)$ denotes independent measurement error of k-th biomarker at each time t, usually from a normal distribution. And the biomarker term in survival model will be the true but unobserved biomarkers' values. Repeated measures or external data will be needed to estimate the variance of the measurement error in order to separation of true biomarker variation and pure measurement error.

Acknowledgements

Acknowledgements should include contributions from anyone who does not meet the criteria for authorship (for example, to recognize contributions from people who provided technical help, collation of data, writing assistance, acquisition of funding, or a department chairperson who provided general support), as well as any funding or other support information.

Conflict of interest

You may be asked to provide a conflict of interest statement during the submission process. Please check the journal's author guidelines for details on what to include in this section. Please ensure you liaise with all co-authors to confirm agreement with the final statement.

References

[1] Elizabeth R Brown, Joseph G Ibrahim, and Victor DeGruttola. A flexible b-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, 61:64–73, 2005.

[2] Chyong-Mei Chen, Pao-sheng Shen, and Yi-Kuan Tseng. Semiparametric transformation joint models for longitudinal covariates and interval-censored failure time. *Computational statistics & data analysis*, 128:116–127, 2018.

[3] Y-S Chen, T-M Hu, Y-Y Wang, and C-L Wu. A case of huntington's disease presenting with psychotic symptoms and rapid cognitive decline in the early stage. *Eur. j. psychiatry*, pages 65–66, 2022.

[4] Chenghao Chu, Ying Zhang, and Wanzhu Tu. Distribution-free estimation of local growth rate around interval censored anchoring events. *Biometrics*, 75:809–826, 2019.

[5] Chenghao Chu, Ying Zhang, and Wanzhu Tu. Stochastic functional estimates in longitudinal models with interval-censored anchoring events. *Scandinavian Journal of Statistics*, 47(3):638–661, 2020.

[6] QL Coent, C Legrand, JJ Dignam, and Rondeau V. Validation of a longitudinal marker as a surrogate using mediation analysis and joint modeling: evolution of the psa as a surrogate of the disease-free survival. *Biometrical Journal*, page e70064, 2025.

[7] Peter J Diggle, Inês Sousa, and Amanda G Chetwynd. Joint modelling of repeated measurements and time-to-event outcomes: The forth armitage lecture. *Statistics in Medicine*, 27:2981–2998, 2008.

[8] Kevin Duff, Jane Paulsen, J Mills, LJ Beglinger, DJ Moser, MM Smith, Douglas Langbehn, Julie Stout, Sarah Queller, and DL Harrington. Mild cognitive impairment in prediagnosed huntington disease. *Neurology*, 75(6):500–507, 2010.

[9] Jean-François Dupuy and Mounir Mesbah. Joint modeling of event time and nonignorable missing longitudinal data. *Lifetime Data Analysis*, 8:99–115, 2002.

[10] Ralitza Gueorgruieva, Robert Rosenheck, and Haiqun Lin. Joint modeling of longitudinal outcome and interval-censored competing risk dropout in a schizophrenia clinical trial. *J R Stat Soc Ser A Stat Soc*, 172(2):417–433, 2012.

[11] Miao Han, Xinyuan Song, Sun Liuquan, and Lei Liu. Joint modelling of longitudinal data with informative observation time and dropouts. *Statistica Sinica*, 24:1487–1504, 2014.

[12] Deborah Lynn Harrington, Megan M Smith, Ying Zhang, Noelle E Carlozzi, Jane S Paulsen, PREDICT-HD Investigators of the Huntington Study Group, et al. Cognitive domains that predict time to diagnosis in prodromal huntington disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 83(6):612–619, 2012.

[13] Robin Henderson, Peter Diggle, and Angela Dobson. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480, 2000.

[14] Fushing Hsieh, Yi-Kuan Tseng, and Jane-Ling Wang. Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics*, 62:1037–1043, 2006.

[15] L Hua, Y Zhang, and W Tu. Spline-based semiparametric sieve likelihood method for over-dispersed panel count data. *The Canadian Journal of Statistics*, 42(2):217–245, 2014.

[16] Liang Li, Bo Hu, and Tom Greene. A semiparametric joint model for longitudinal and survival data with application to hemodialysis study. *Biometrics*, 65:737–745, 2009.

[17] Haiqun Lin, Charles E McCulloch, and Susan T Mayne. Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. *Statistics in Medicine*, 21:2369–2382, 2002.

[18] L Liu, C Zheng, and J Kang. Exploring causality mechanism in the joint analysis of longitudinal and survival data. *Statistics in Medicine*, 37:3733–3744, 2018.

[19] M Lu, Y Zhang, and J Huang. Semiparametric estimation methods for panel count data using monotone b-spline. *Journal of the American Statistical Association*, 104:1060–1070, 2009.

[20] Jung Yeon Park, Melanie M. Wallb, Irini Moustaki, and Arnold H. Grossman. A joint modelling approach for longitudinal outcomes and non-ignorable dropout under population heterogeneity in mental health studies. *Journal of Applied Statistics*, 49(13):3361–3376, 2022.

[21] Jane S Paulsen. Cognitive impairment in huntington disease: diagnosis and treatment. *Current neurology and neuroscience reports*, 11(5):474–483, 2011.

[22] Jane S Paulsen, Douglas R Langbehn, Julie C Stout, Elizabeth Aylward, Christopher A Ross, Martha Nance, Mark Guttman, Shannon Johnson, M MacDonald, Leigh J Beglinger, et al. Detection of huntington ’ s disease decades before diagnosis: the predict-hd study. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(8):874–880, 2008.

[23] Jane S Paulsen, Megan M Smith, Jeffrey D Long, Predict HD Investigators, Coordinators of the Huntington Study Group, et al. Cognitive decline in prodromal huntington disease: implications for clinical trials. *Journal of Neurology, Neurosurgery & Psychiatry*, 84(11):1233–1239, 2013.

[24] Dimitris Rizopoulos. *Joint models for longitudinal and time-to-event data: With applications in R*. CRC press, 2012.

[25] Christopher A Ross and Sarah J Tabrizi. Huntington disease: natural history, biomarkers and prospects for therapeutics. *Nature Reviews Neurology*, 7(10):204–216, 2014.

[26] Anaïs Rouanet, Pierre Joly, Jean-François Dartigues, Cècile Proust-Lima, and Hélène Jacqmin-Gadda. Joint latent class model for longitudinal data and interval-censored semi-competing events: application to dementia. *Biometrics*, 72:1123–1135, 2016.

[27] L Schumaker. *Spline Function: Basic Theory*. New York: Wiley, 1981.

[28] Xiao Song, Marie Davidian, and Anastasios A Tsiatis. A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics*, 58:742–753, 2002.

[29] Julie C Stout, Rebecca Jones, Izelle Labuschagne, Alison M O’Regan, Miranda J Say, Eve M Dumas, Sarah Queller, Damian Justo, Rachelle Dar Santos, Allison Coleman, et al. Evaluation of longitudinal 12 and 24 month cognitive outcomes in premanifest and early huntington’s disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 83(7):687–694, 2012.

[30] W Su, L Liu, G Yin, X Zhao, and Y Zhang. Semiparametric reversed mean model for recurrent event process with informative terminal event. *Statistica Sinica*, 34:1843–1862, 2024.

[31] F.O. Walker. Huntington’s disease. *The Lancet*, 369(9557):218–228, 2007.

[32] Di Wu and Chenxi Li. Joint analysis of multivariate interval-censored survival data and a time-dependent covariate. *Statistical Methods in Medical Research*, 30(3):768–784, 2021.

[33] Michael S Wulfsohn and Anastasios A Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, 53:330–339, 1997.

[34] Y Zhang, L Hua, and J Huang. A spline-based semiparametric maximum likelihood method for the cox model with interval-censored data. *Scandinavian Journal of Statistics*, 37:338–354, 2010.

[35] Ying Zhang, Gang Cheng, and Wanzhu Tu. Robust nonparametric estimation of monotone regression functions with interval-censored observations. *Biometrics*, 72(3):720–730, 2016.

[36] Ying Zhang, Junyi Zhou, Carissa R Gehl, Jeffrey D Long, Hans Johnson, Vincent A Magnotta, Daniel Sewell, Kathleen Shannon, and Jane S Paulsen. Mild cognitive impairment as an early landmark in huntington’s disease. *Frontiers in neurology*, 12:678652, 2021.

[37] C Zheng and L Liu. Quantifying direct and indirect effect for longitudinal mediator and survival outcome using joint modeling approach. *Biometrics*, 78:1233–1243, 2022.

[38] X Zhou and X Song. Causal mediation analysis for multivariate longitudinal data and survival outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, 30:749–760, 2023.