Submitted to the Annals of Applied Statistics
arXiv: arXiv:0000.0000

ROBUST MIXED-EFFECTS MODEL FOR CLUSTERED FAILURE TIME DATA: APPLICATION TO HUNTINGTON'S DISEASE EVENT MEASURES

By Tanya P. Garcia[¶], Yanyuan Ma[‖] Karen Marder[‡,**] and Yuanjia Wang[§,**]

Texas A&M University[¶], Pennsylvania State University[‖], and Columbia University[**]

An important goal in clinical and statistical research is properly modeling the distribution for clustered failure times which have a natural intra-class dependency and are subject to censoring. We handle these challenges with a novel approach that does not impose restrictive modeling or distributional assumptions. Using a logit transformation, we relate the distribution for clustered failure times to covariates and a random, subject-specific effect. The covariates are modeled with unknown functional forms, and the random effect may depend on the covariates and have an unknown and unspecified distribution. We introduce pseudo-values to handle censoring and splines for functional covariate effects and frame the problem into fitting an additive logistic mixed effects model. Unlike existing approaches for fitting such models, we develop semiparametric techniques that estimate the functional model parameters without requiring specifying or estimating the random effect distribution. We show both theoretically and empirically that the resulting estimators are consistent for any choice of random effect distribution and any dependency structure between the random effect and covariates. Lastly, we illustrate the method's utility in an application to a Huntington's disease study where our method provides new insights into differences between motor and cognitive impairment event times in at risk subjects.

[¶]T.P. Garcia is supported by the National Institute of Neurological Disease and Stroke (K01NS099343), Huntington's Disease Society of America Human Biology Project Fellowship, and Texas A&M School of Public Health Research Enhancement and Development Initiative (REDI-23-202059-36000).

[‖]Y. Ma is supported by National Science Foundation (DMS-1608540).

[‡]K. Marder is supported by NIH UL1TR001873, U10NS077267, U01NS100600, U01NS0822062, CHDI, Huntington's Disease Society of America, Michael J. Fox Foundation, Parkinson's Disease Foundation, TEVA (research support), Vaccinex (research support).

[§]Y. Wang is supported by National Institute of Neurological Disease and Stroke (NS073671, NS082062).

MSC 2010 subject classifications: Primary 62N01, 62N01; secondary 62P10

Keywords and phrases: Additive model, Clustered failure times, Logistic mixed model, Varying coefficient model, Semiparametric Estimator, Splines

1. Introduction. Clustered failure time data are commonly collected in biomedical research. Examples include the onset ages among family members for neurodegenerative disorders (Marder et al., 2003); and the time until first signs appear from an infectious disease in clusters of hospitals (Huang et al., 2010). In these examples and others, a key interest is properly modeling the clustered failure time distribution which has several challenges: within cluster dependency, right censoring, and the unknown relationship between covariates and failure times. We address these challenges with a new estimation framework that is simple and uses minimal assumptions to reduce the chance of model misspecification. The research focus is often on the failure time distributions themselves instead of hazard functions, so in this regard, we directly model the clustered failure time distribution. We use a time-varying, proportional odds model with functional covariates and a random effect. The random effect is free of distributional assumptions and is possibly correlated with some or all covariates. Over a range of time points, we cast the proportional odds model into an additive logistic mixed effect model using pseudo-values (Logan et al., 2011) to handle censoring and splines for the functional covariate effects. We then develop semiparametric methods to consistently estimate the model parameters without estimating or specifying working models for the random effect distribution. Our approach thus contributes a flexible new estimation framework that circumvents the challenges of clustered failure time data.

1.1. *Motivating example.* Our work is motivated by an observational study of Huntington's disease (HD) that evaluated failure-type events representative of the disease progression. HD is an autosomal dominant, neurodegenerative disease caused by an unstable expansion of the cytosine-adenine-guanine (CAG) trinucleotide repeat in the huntingtin gene (Huntington's Disease Collaborative Research Group, 1993). More CAG repeats lead to earlier onset of impairments (Ross and Tabrizi, 2010). In 2005-2011, the Cooperative Huntington's Observational Research Trial (COHORT) study was conducted on genetically predisposed individuals. For each participant, the study recorded (potentially censored) failure-type events representative of the disease course: the age when an individual first experienced a motor sign (i.e., chorea, dystonia, rigidity), and the age when cognitive impairments first impacted daily life. The data are an example of clustered failure times: for each subject, a cluster is formed by the two event times measured on that subject. A key interest is comparing the conditional odds of these events occurring by age t given the subject's CAG repeat-length and gender. Large conditional odds in favor of one event occurring before the other

helps to inform the natural history of the disease. This is critical for planning clinical trials, deciding the timing of intervention focus, and prognostic counseling.

For cluster $i = 1, \ldots, n$ and member $j = 1, \ldots, m_i$, we model the clustered failure time distribution. Let T_{ij} denote failure times, $X_{ij} \in \mathbb{R}$ and $\boldsymbol{Z}_{ij} \in \mathbb{R}^{p_1}$ denote covariates, and $R_i(\cdot)$ denote a random, cluster-specific effect. In the HD example, cluster i includes event times from the ith participant: age of first motor impairment (T_{i1}) and age when cognitive impairments first impact daily life (T_{i2}). Associated covariates are CAG repeat lengths (X) and gender (Z), and a random effect $R_i(\cdot)$ is associated with each subject. The clustered failure time distribution is then modeled as

$$\text{(1.1)}\quad \text{logit}[\text{pr}\{T_{ij} \leq t | X_{ij}, \boldsymbol{Z}_{ij}, R_i(t)\}] \quad = \quad \alpha(X_{ij}, t) + \boldsymbol{Z}_{ij}^{\mathrm{T}} \boldsymbol{\beta}(t) + R_i(t)$$

where $\text{logit}(p) = \log\{p/(1-p)\}$. The above is a time-varying, proportional odds model with random effect and the overall objective is to estimate the functional parameters $\alpha(X, t) \in \mathbb{R}$ and $\boldsymbol{\beta}(t) \in \mathbb{R}^{p_1}$. For the HD example, $\alpha(X, t)$ represents the time-varying effect of CAG repeats and $\beta(t)$ the gender effect. Estimating these functional parameters allows us to compare event times through conditional odds ratios. For example, with HD, given the subject covariates and random effect, we may compute the conditional odds of a motor impairment occurring by age t compared to a cognitive impairment occurring by age t via $\exp\{\widehat{\alpha}(X, t) + \widehat{\beta}(t)Z\}$ (see Section 4). The resulting quantity helps to identify which event in the disease course has better odds of occurring first.

A few remarks of the model in (1.1) are in order. First, the model is presented for scalar X_{ij}, but it can easily accomodate vector $\boldsymbol{X}_{ij} = (X_{ij1}, \ldots, X_{ijp})^T$ by replacing $\alpha(X_{ij}, t)$ with the summation $\sum_{k=1}^{p} \alpha(X_{ijk}, t)$. This generality introduces more computation but does not change the essence of the proposed method. Second, we separate covariates X_{ij} and $\boldsymbol{Z}_{ij}$ to distinguish between assumptions of their effects. Specifically, for X_{ij} we make nonparametric assumptions and for $\boldsymbol{Z}_{ij}$, we make parametric assumptions where the choice between assumptions is driven by the application and the flexibility desired. One could consider only nonparametric covariate effects (i.e., only $\alpha(X, t)$ terms) or only parametric effects (i.e., only $\boldsymbol{Z}^T\boldsymbol{\beta}(t)$ terms), but our method does not fundamentally change. Thus, because these two generalities do not fundamentally change our method, we proceed under the presentation in equation (1.1).

1.2. *Relationship to existing models and methods.* Model (1.1) differs from the existing proportional odds model for univariate censored data (Ben-

nett, 1983; Murphy et al., 1997) and for multivariate data with random effects (Zeng et al., 2005). These models are not designed for time-varying or functional parameters as is ours, and extensions are non-trivial. A significant challenge with the model in (1.1) is estimating $\alpha(X,t), \boldsymbol{\beta}(t)$ in the presence of the unobserved random effect $R_i(t)$. This same challenge exists for proportional hazards frailty models (Clayton, 1978) which is another model for clustered failure time distributions. The standard in proportional hazards frailty models is to specify a distribution for the random effect (i.e., frailty) such as inverse gaussian (Henderson and Oman, 1999), log normal (Ripatti and Palmgren, 2000), and gamma (Chen and Lio, 2008) for its mathematical convenience. Empirical studies have shown that minimal bias and efficiency loss occur under an assumed gamma frailty when the true frailty distribution is inverse Gaussian or positive stable (Hsu et al., 2007) or specific discrete distributions (Glidden and Vittinghoff, 2004). Despite these encouraging robustness results, they do not hold for population-averaged proportional hazards models when there is strong within-cluster dependency or when covariates and random effects are dependent (Heagerty and Kurland, 2001). In fact, because it is very difficult to model the distribution of the random effects conditional on the covariates, random effects are routinely assumed to be independent of covariates and modeled with a marginal model. But doing so can bias the subsequent estimation (Garcia and Ma, 2015).

Concerns for misspecification of the random effect (or frailty) distribution have motivated a range of graphical and numerical goodness-of-fit tests (Shih and Louis, 1995; Chen and Bandeen-Roche, 2005). Unfortunately, these tests are only applicable for certain cluster sizes and no test indicates the correct distribution when a poor fit is detected. Various approaches have been developed to directly address misspecification. These include modeling the random effect as a mixture of normals (Lesaffre and Molenberghs, 2001), a Student-t (Congdon, 1994), skew-t (Lee and Thompson, 2008), and other families of parametric distributions (Piepho and McCulloch, 2004). None of the parametric approaches, however, solves the problem completely since parametric forms do not span the entire range of possibilities. Semiparametric or nonparametric methods (e.g., Geerdens et al., 2013) better handle the misspecification problem, but also require more intense computation.

All aforementioned methods target at the issue caused by the misspecification of the random effect distribution. When population average effects are of interest, one way to bypass the random effect is to consider a different modeling approach and to work with marginal models (Chen et al., 2010). A marginal model does not impose any particular form of dependency. Hence, it is different from our model (1.1), which is a conditional

model that captures dependency by conditioning on the random effect. The interpretation of parameters in marginal models also differs from that of a conditional model with random effects, and the two are not comparable for nonlinear models (e.g. logit-link models as in equation (1.1)). In this work, our parameters of interest are conditional time-dependent log odds ratios instead of marginal parameters. Lastly, our approach has some resemblance with that of Efron (1988) in that we will use logistic regression techniques for survival curve estimation. However, our approach applies to clustered failure time data which that of Efron (1988) does not.

1.3. *Proposed method.* Given the aforementioned limitations of existing methods, we propose here a new, flexible approach that allows $R_i(\cdot)$ to depend on covariates and have a distribution that is unknown and unspecified. Our approach uses pseudo-values to handle censoring and splines for functional covariate effects. The combination leads to a simple semiparametric estimation framework that circumvents the challenges of having the random effect distribution be unknown and unspecified. The remainder of the paper is as follows. Section 2 describes the main technical results of the proposed method including asymptotic properties. Section 3 demonstrates the method's numerical effectiveness against competing approaches in terms of different clustering structures, random effect distributions, and dependencies between the random effect and model covariates. Section 4 provides a novel analysis of clinical differences between motor and cognitive impairment event times in individuals genetically predisposed to Huntington's disease. Section 5 concludes the paper. All proofs and additional simulations are deferred to Supplementary Material. An R implementation of the procedure is available upon request.

2. Main Estimation.

2.1. *Estimation setup.* Our main objective is to estimate $\alpha(x,t), \boldsymbol{\beta}(t)$ in equation (1.1) in the presence of unobserved random effects $R_i(t)$ with unknown distribution. At first glance, one may want to place restrictions on $\alpha(x,t), \boldsymbol{\beta}(t)$ and $R_i(t)$ to ensure that $\text{pr}\{T_{ij} \leq t | X_{ij}, \boldsymbol{Z}_{ij}, R_i(t)\}$ is a non-decreasing function of t. However, this does not necessarily hold since $R_i(t)$ is random at different t values and cannot be required to be monotone. Therefore, we do not impose any restrictions on $\alpha(x,t), \boldsymbol{\beta}(t)$ and $R_i(t)$.

We propose to estimate $\alpha(x,t), \boldsymbol{\beta}(t)$ at different $t = t_0$ values and then use linear interpolation. The t_0 values are chosen to spread evenly across the range of $S_{ij} = \min(T_{ij}, C_{ij})$, where T_{ij} denotes the failure time and C_{ij} the right-censoring time. (See Step 1 in the algorithm of Section 2.2 for how to

choose the number of t_0 values). Throughout, we assume C_{ij} is independent of T_{ij} and covariates $X_{ij}, \boldsymbol{Z}_{ij}$. Lastly, we let $\Delta_{ij} = I(T_{ij} \leq C_{ij})$ denote the censoring indicator.

We will transform the model in (1.1) to an additive logistic mixed effects model by introducing splines for functional covariate effects and jackknife pseudo-values to handle censoring as described next.

2.1.1. *Splines for functional parameters.* At each t_0, we approximate the unknown functional form $\alpha(x, t_0)$ using a B-spline of order r with N internal knots. We let

$$\xi_1 = \cdots = \xi_r < \xi_{r+1} < \cdots < \xi_{r+N} < \xi_{N+r+1} = \cdots = \xi_{N+2r} \tag{2.1}$$

where $\xi_{r+1}, \ldots, \xi_{N+r}$ is the sequence of internal knots. We also let the distance between neighboring knots be $h_k = \xi_{k+1} - \xi_k$ for $r \leq k \leq N + r$, and let $h = \max_{r \leq k \leq N+r} h_k$. In practice, the knots are often placed at equally spaced sample quantiles of the predictor X, and a common order is $r = 4$ corresponding to a cubic B-spline. In our empirical examples, we found that this knot selection and order worked well.

Based on the order and the number of internal knots, the number of B-spline basis functions is $p_2 = N + r$, and $\alpha(x, t_0)$ is approximated by

$$\widetilde{\alpha}(x, t_0) = \sum_{k=1}^{p_2} B_k(x) a_k(t_0) = \boldsymbol{B}^{\mathrm{T}}(x) \boldsymbol{a}(t_0), \tag{2.2}$$

where $\boldsymbol{a}(t_0)$ is a p_2-dimensional spline coefficient vector, and $\boldsymbol{B}(\cdot)$ are spline basis functions that do not include the intercept. We can ignore the intercept as it is common to all failure times and thus, by definition, is absorbed into the random intercept.

2.1.2. *Pseudo-value approach for censoring.* Pseudo-value regression (Logan et al., 2008, 2011) is a simple method to perform estimation for incomplete data due to right-censoring. In our case, the response of interest is the binary event status $Y_{ij}(t_0) \equiv I(T_{ij} \leq t_0)$ motivated by modeling the distribution function via logistic regression (Efron, 1988). The binary event $Y_{ij}(t_0)$ is observable when $\Delta_{ij} = 1$, or when $\Delta_{ij} = 0$ and $C_{ij} \geq t_0$ for which $I(T_{ij} \leq t_0) = 0$ since $t_0 \leq C_{ij} < T_{ij}$. Otherwise, when $\Delta_{ij} = 0$ and $C_{ij} < t_0$, $Y_{ij}(t_0)$ is unobservable.

To replace the unobservable $Y_{ij}(t_0)$, the idea is to construct jackknife pseudo-values that seemingly ignore dependencies in the data, but are later related to covariates and the random effect in a regression model that

re-captures the dependency. The construction of the pseudo-values uses Kaplan-Meier estimators designed for independent data, but are consistent even for dependent data (Ying and Wei, 1994). We show below that the pseudo-values satisfy properties which allow us to (i) relate the pseudo-values to covariates and random effect through a regression model; and (ii) use the regression model to unbiasedly estimate the model parameters even when pseudo-values are used in place of the unknown binary events.

We now define two types of jackknife pseudo-values depending on the nature of event type.

EXAMPLE 1. *Single event type. Suppose cluster i contains information about a common event. For example, cluster i corresponds to a family and T_{ij} represents the time to a common event (e.g., disease-onset age) for each family member j. Another example is when cluster i corresponds to an individual and T_{ij} are recurrent event times (e.g., tumor occurrences) for individual i.*

Let $M = \sum_{i=1}^{n} m_i$. The jackknife pseudo-value to substitute $Y_{ij}(t_0)$ is

$$Y_{ij}^*(t_0) = M\widehat{F}(t_0) - (M-1)\widehat{F}^{-(ij)}(t_0).$$

Here, $\widehat{F}(t_0) = 1 - \widehat{S}(t_0)$ with $\widehat{S}(t_0)$ the Kaplan-Meier estimator based on all M events, and $\widehat{F}^{-(ij)}(t_0)$ is a similar estimator after removing observation j from cluster i.

The pseudo-value in Example 1 is a special case of pseudo-values constructed for clustered data with competing risks but there is no competing outcome (Logan et al., 2011). When there is no censoring prior to t_0, $Y_{ij}^*(t_0)$ simplifies to $I(T_{ij} \leq t_0)$ (Logan et al., 2011, sec. 2.3). Otherwise, under censoring, $Y_{ij}^*(t_0)$ satisfies two properties:

(P1) For clusters, $i \neq k$, pseudo-values $Y_{ij}^*(t_0)$ and $Y_{k\ell}^*(t_0)$ are approximately independent as M tends to infinity.

(P2) The conditional expectation of $Y_{ij}^*(t_0)$ given $X_{ij}, \boldsymbol{Z}_{ij}$ and $R_i(t_0)$ satisfies $\lim_{M\to\infty} E\{Y_{ij}^*(t_0)|X_{ij}, \boldsymbol{Z}_{ij}, R_i(t_0)\} = \text{pr}\{T \leq t_0|X_{ij}, \boldsymbol{Z}_{ij}, R_i(t_0)\}$.

Justification of (P1) and (P2) is provided in Logan et al. (2011) and a summary of the key results is in Section S.1.1 (Supplementary Material). The properties imply that (asymptotically) the relationship between pseudo-values and the covariates and random effect is exactly the conditional distribution $\text{pr}\{T \leq t_0|X_{ij}, \boldsymbol{Z}_{ij}, R_i(t_0)\}$ in equation (1.1); see property (P2). As shown in Klein et al. (2014, chap. 10), this implies that one may construct unbiased estimating equations using the pseudo-values in place of the unobservable binary event indicators. The unbiased estimating equations

(see Proposition 3) will then lead to consistent estimators for the model parameters of interest.

EXAMPLE 2. *Multiple event types. Suppose cluster i contains information about multiple event types. For example, cluster i corresponds to an individual and T_{ij} represents measures of multiple event types j on the same individual (i.e., age of first motor impairment, age of first cognitive impairment as in the HD application, Section 4). The jackknife pseudo-value to substitute the unobservable $Y_{ij}(t_0)$ is*

$$Y_{ij}^{\dagger}(t_0) = n\widehat{F}_j(t_0) - (n-1)\widehat{F}_j^{-(i)}(t_0).$$

Here, $\widehat{F}_j(t_0) = 1 - \widehat{S}_j(t_0)$ with $\widehat{S}_j(t_0)$ the Kaplan-Meier estimator using only information for event j from all n clusters, and $\widehat{F}_j^{-(i)}(t_0)$ is a similar estimator after removing cluster i.

The setting of Example 2 resembles that for competing risks except that the occurrence of one event does not preclude the observation of another. This is exactly the setting of the HD application (Section 4). One observes the age of first motor impairment and age of first cognitive impairment as the disease progresses, but the occurrence of either impairment does not preclude the other. In Example 2, because the event types are different and non-competing, it does not make sense to combine information across event types when computing pseudo-values (as done in Example 1). Instead, when handling different, non-competing event types, pseudo-values are constructed using event-specific Kaplan-Meier estimators (i.e., $1 - \widehat{S}_j(t_0)$) as specified in Example 2.

Properties of $Y_{ij}^{\dagger}(t_0)$ are similar to those for $Y_{ij}^{*}(t_0)$ except for the notational changes to reflect the event-specific Kaplan-Meier estimators. First, when there is no censoring prior to t_0, $Y_{ij}^{\dagger}(t_0)$ is the binary indicator of whether event type j for person i occurred prior to t_0. Second, $Y_{ij}^{\dagger}(t_0)$ satisfies

(P1†) For clusters, $i \neq k$, the pseudo-values $Y_{ij}^{\dagger}(t_0)$ and $Y_{kj}^{\dagger}(t_0)$ are approximately independent as n tends to infinity.

(P2†) The conditional expectation of $Y_{ij}^{\dagger}(t_0)$ given $X_{ij}, \boldsymbol{Z}_{ij}$ and $R_i(t_0)$ satisfies $\lim_{n\to\infty} E\{Y_{ij}^{\dagger}(t_0)|X_{ij}, \boldsymbol{Z}_{ij}\} = \text{pr}\{T_j \leq t_0|X_{ij}, \boldsymbol{Z}_{ij}, R_i(t_0)\}$ where $\text{pr}\{T_j \leq t_0|X_{ij}, \boldsymbol{Z}_{ij}, R_i(t_0)\}$ denotes the conditional distribution for event type j.

Justification of properties (P1†) and (P2†) is in Section S.1.2 (Supplementary Material) and follows the proof in Logan et al. (2011). As with Exam-

ple 1, properties (P1†) and (P2†) mean that one may construct regression models relating the pseudo-values to model covariates with pseudo-values appropriately replacing the unobservable $I(T_{ij} \leq t_0)$. Estimating equations constructed from these regression models will also be unbiased and hence yield consistent estimators for parameters in the conditional distribution which is linked to pseudo-values by (P2†).

A few remarks about the pseudo-values in Examples 1 and 2 are in order. Properties (P2) and (P2†) follow because we assume that censoring does not depend on covariates. This assumption, however, can be relaxed by constructing pseudo-values that are covariate-dependent (Andersen and Perme, 2010). Suppose censoring depends on a discrete covariate U with values $1, 2, \ldots, m_u$. Then, for Example 1, in place of the Kaplan-Meier estimator $\widehat{F}(t_0) = 1 - \widehat{S}(t_0)$, one would use $1 - \widehat{S}_k(t_0)$ where $\widehat{S}_k(t_0)$ is the Kaplan-Meier estimate based on subjects with covariate $U = k$. Also, in place of M, one would use M_k corresponding to the number of subjects with covariate $U = k$. Likewise for Example 2, let $\widehat{S}_{jk}(t_0)$ be the Kaplan-Meier estimate for event type j based on subjects with covariate $U = k$, and n_k be the number of subjects with covariate $U = k$. Then, in Example 2, one replaces $\widehat{F}_j(t_0) = 1 - \widehat{S}_j(t_0)$ with $1 - \widehat{S}_{jk}(t_0)$ and replaces n with n_k. Andersen and Perme (2010) showed that this approach corrects the bias introduced when covariate-dependent censoring is ignored, but induces higher variability than pseudo-values based on the standard Kaplan Meier estimators. The approach of Andersen and Perme (2010) is easy to accommodate when U is discrete and can be adapted to continuous U via kernel weights. However, the method quickly becomes onerous when U is multivariate for both discrete and continuous cases. These cases require a careful and separate investigation that is beyond the scope of the current paper.

Lastly, the jackknife pseudo-values as defined in the Examples are not guaranteed to be in $[0, 1]$. This is important considering that they are ultimately used to model a conditional distribution function. When the pseudo-values fall outside this interval, we can round them to the nearest 0 or 1. In our empirical studies, the proportion of jackknife pseudo-values that fall outside $[0, 1]$ was less than 7% (see Table S.1, Supplementary Material), and consistency appears unaffected. Logan et al. (2008) made similar observations for identical pseudo-values as proposed here.

2.2. *Estimation procedure.* We now describe how we relate the pseudo-values to the covariates in an additive logistic mixed-effect model. Following Section 2.1.2, let $Y_{ij}(t_0)$ be $I(T_{ij} \leq t_0)$ when the binary indicator is observable and a pseudo-value otherwise. The computation of the pseudo-value

depends on the problem (see Examples 1 and 2).

For ease in notation, we describe the estimation procedure at a fixed $t = t_0$, so that the notation $Y_{ij}(t_0)$, $\alpha(x, t_0)$, $\boldsymbol{a}(t_0)$, $\boldsymbol{\beta}(t_0)$, simplifies to Y_{ij}, $\alpha(x)$, $\boldsymbol{a}$, $\boldsymbol{\beta}$, respectively. Let $\boldsymbol{\theta} = (\boldsymbol{a}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ be a vector of length $q = p_1 + p_2$; $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^{\mathrm{T}}$ and $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{im_i})^{\mathrm{T}}$ be m_i-dimensional vectors, and $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}, \ldots, \boldsymbol{Z}_{im_i})$ be a $p_1 \times m_i$ matrix. Then, under the B-spline model in (2.2), our model in (1.1) satisfies

$$E(Y_{ij}|X_{ij}, \boldsymbol{Z}_{ij}, R_i) = \frac{\exp\{\eta(X_{ij}, \boldsymbol{Z}_{ij}; \boldsymbol{\theta}) + R_i\}}{1 + \exp\{\eta(X_{ij}, \boldsymbol{Z}_{ij}; \boldsymbol{\theta}) + R_i\}}, \quad j = 1, \ldots, m_i,$$

where $\eta(X_{ij}, \boldsymbol{Z}_{ij}; \boldsymbol{\theta}) = \boldsymbol{B}^{\mathrm{T}}(X_{ij})\boldsymbol{a} + \boldsymbol{Z}_{ij}^{\mathrm{T}}\boldsymbol{\beta}$. The above expression is the conditional mean for a logistic mixed effects model, and holds whether Y_{ij} is an observed binary indicator or a pseudo-value. With f denoting (conditional) densities described by the subindices, the density for the ith cluster is

$$\begin{aligned} f_{\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta}) &= \int f_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},R}(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i, r_i; \boldsymbol{\theta}) f_{\boldsymbol{X},\boldsymbol{Z},R}(\boldsymbol{x}_i, \boldsymbol{z}_i, r_i) d\mu(r_i) \\ = \int \exp\left(\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta}) + \mathbf{1}_{m_i}^{\mathrm{T}} r_i\}\boldsymbol{y}_i - \sum_{j=1}^{m_i} \log[1 + \exp\{\eta(x_{ij}, \boldsymbol{z}_{ij}; \boldsymbol{\theta}) + r_i\}] \right) & \\ \times f_{\boldsymbol{X},\boldsymbol{Z},R}(\boldsymbol{x}_i, \boldsymbol{z}_i, r_i) d\mu(r_i). & \end{aligned} \tag{2.3}$$

Here, μ denotes the dominating measure, $\mathbf{1}_{m_i}$ is a m_i-dimensional vector of ones and $\boldsymbol{\eta}(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta}) = \{\eta(x_{i1}, \boldsymbol{z}_{i1}; \boldsymbol{\theta}), \ldots, \eta(x_{im_i}, \boldsymbol{z}_{im_i}; \boldsymbol{\theta})\}^{\mathrm{T}}$. We assume the joint density $f_{\boldsymbol{X},\boldsymbol{Z},R}(\boldsymbol{x}, \boldsymbol{z}, r)$ is a valid, yet unspecified distribution with $\boldsymbol{X}, \boldsymbol{Z}$ and R not necessarily independent.

An immediate advantage of the representation in (2.3) is that it reveals the connection between $f_{\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z})$ and generalized linear latent variable models (Huber et al., 2004; Conne et al., 2010) with latent variable R. We show in Section S.1.3 (Supplementary Material) that for such a model, a consistent estimator for $\boldsymbol{\theta}$ results from treating $f_{\boldsymbol{X},\boldsymbol{Z},R}(\boldsymbol{x}, \boldsymbol{z}, r)$ as a nuisance parameter and factoring out its effect with semiparametric projection. The result is summarized in the proposition below.

PROPOSITION 1. *For the joint density in (2.3), whether $\boldsymbol{Y}_i$ consists of observable binary indicators or pseudo-values, a consistent estimator for $\boldsymbol{\theta}$ is the root of $\sum_{i=1}^{n} \boldsymbol{S}_{\mathrm{eff}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) = \mathbf{0}$ where $\boldsymbol{S}_{\mathrm{eff}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta}) = \boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}) - E\{\boldsymbol{h}(\boldsymbol{X}, \boldsymbol{Z}, R)|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}\}$. The q-dimensional score vector $\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}) = \partial log f_{\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta})/\partial \boldsymbol{\theta}$ and $\boldsymbol{h}$ is an unknown q-dimensional function satisfying*

$$E\{\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}) - E\{\boldsymbol{h}(\boldsymbol{X}, \boldsymbol{Z}, R)|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}\}|\boldsymbol{X}, \boldsymbol{Z}, R] = \mathbf{0}. \tag{2.4}$$

The proof of Proposition 1 is in Section S.1.3 (Supplementary Material). The proposition indicates that forming the estimating equation requires solving for $\boldsymbol{h}$ in (2.4), but this is an ill-posed problem (Tsiatis and Ma, 2004). Fortunately, Proposition 1 *combined* with a simple decomposition of $\boldsymbol{Y}_i$ allows us to circumvent the ill-posed problem.

PROPOSITION 2. *Define* $W_i = \mathbf{1}_{m_i}^{\mathrm{T}} \boldsymbol{Y}_i = \sum_{j=1}^{m_i} Y_{ij}$, $\boldsymbol{V}_i = (\mathbf{0}, \boldsymbol{I}_{m_i-1})\boldsymbol{Y}_i = (Y_{i,2}, \ldots, Y_{i,m_i})^{\mathrm{T}}$,

$$\boldsymbol{A}_i = \begin{pmatrix} 1 & \mathbf{1}_{m_i-1}^{\mathrm{T}} \\ \mathbf{0} & \boldsymbol{I}_{m_i-1} \end{pmatrix}.$$

Under this transformation, $\boldsymbol{Y}_i = \boldsymbol{A}_i^{-1}(W_i, \boldsymbol{V}_i^{\mathrm{T}})^{\mathrm{T}}$ *and a simpler, consistent estimator for* $\boldsymbol{\theta}$ *is the root of* $\sum_{i=1}^{n} \boldsymbol{S}_{\mathrm{eff}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) = \mathbf{0}$, *where*

$$\begin{aligned} \boldsymbol{S}_{\mathrm{eff}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta}) = {} & E\{\boldsymbol{U}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, R) | W, \boldsymbol{V}, \boldsymbol{X}, \boldsymbol{Z}\} \\ & - E\{\boldsymbol{U}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, R) | W, \boldsymbol{X}, \boldsymbol{Z}\}. \end{aligned}$$

Here $\boldsymbol{U}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, R) = \partial \log f_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},R}(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{Z}, R; \boldsymbol{\theta})/\partial \boldsymbol{\theta}$ *and* $f_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},R}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{z}, r)$ *is the first product term of the integrand in (2.3).*

The construction of W and $\boldsymbol{V}$ comes from how one may isolate the random effect terms in (2.3). The isolation comes from two special properties (proofs are in Section S.1.4 of Supplementary Material). The first property is that given $(W, \boldsymbol{X}, \boldsymbol{Z})$, the terms R and $\boldsymbol{V}$ are conditionally independent. The second property is that for any q-dimensional function $\boldsymbol{g}(W, \boldsymbol{X}, \boldsymbol{Z})$ whenever $E\{\boldsymbol{g}(W, \boldsymbol{X}, \boldsymbol{Z}) | \boldsymbol{X}, \boldsymbol{Z}, R\} = \mathbf{0}$, we have that $\boldsymbol{g}(W, \boldsymbol{X}, \boldsymbol{Z}) = \mathbf{0}$. Applying these properties to the ill-posed equation in (2.4) removes the outer integral with the random effect and leads to the simplified estimating equation in Proposition 2. A proof of this result is in Section S.1.4 (Supplementary Material).

The results in Proposition 2 are fundamental to simplifying our method. They establish that the estimating equation solely involves conditional expectations of $\boldsymbol{U}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, R)$. While such a calculation normally requires integrating out the unknown random effect, we show next that we can actually bypass this step since the random effect drops out.

PROPOSITION 3. *With* W_i *and* $\boldsymbol{V}_i$ *as in Proposition 2, the estimating*

equations for the logistic mixed effects model are free of R_i and take the form

$$\sum_{i=1}^{n} \boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\{\boldsymbol{B}(X_{ij}) - \boldsymbol{B}(X_{i1})\}\{V_{ij} - E(V_{ij}|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})\}$$
$$\sum_{i=1}^{n} \boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}(\boldsymbol{Z}_{ij} - \boldsymbol{Z}_{i1})\{V_{ij} - E(V_{ij}|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})\}.$$

The proof for Proposition 3 (Section S.1.5 of Supplementary Material) follows from the form of $f_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},R}$, the first product term in the integrand in (2.3). Specifically, direct calculation shows that $\boldsymbol{U}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, R) = \partial \log f_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},R}(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{Z}, R; \boldsymbol{\theta})/\partial\boldsymbol{\theta}$ is composed of two terms: the first a function of $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})$ and the second a function of $(\boldsymbol{X}, \boldsymbol{Z}, R)$. This separation allows us to eliminate the contribution of R via the special properties of W and $\boldsymbol{V}$ mentioned above. Because R and $\boldsymbol{V}$ are conditionally independent given $(W, \boldsymbol{X}, \boldsymbol{Z})$, then for any function $\boldsymbol{k}(\boldsymbol{X}, \boldsymbol{Z}, R)$, we have $E\{\boldsymbol{k}(\boldsymbol{X}, \boldsymbol{Z}, R)|W, V, \boldsymbol{X}, \boldsymbol{Z}\} - E\{\boldsymbol{k}(\boldsymbol{X}, \boldsymbol{Z}, R)|W, \boldsymbol{X}, \boldsymbol{Z}\} = \boldsymbol{0}$. Hence, the second term in $\boldsymbol{U}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, R)$, which is the only one containing R, does not contribute at all to the estimating equations. For this reason, the estimating equations in Proposition 3 are free of the unknown R, and the need to integrate out the random effect is completely eliminated.

The only main computation in Proposition 3 is forming $E(V_{ij}|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})$. In Section S.1.6 (Supplementary Material), we show that this expectation is

$$E(V_{ij}|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) = \frac{\int_{\mathcal{R}(\boldsymbol{v}_i)} v_{ij} \exp\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})\boldsymbol{A}_i^{-1}(w_i, \boldsymbol{v}_i^{\mathrm{T}})^{\mathrm{T}}\} d\mu(\boldsymbol{v}_i)}{\int_{\mathcal{R}(\boldsymbol{v}_i)} \exp\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})\boldsymbol{A}_i^{-1}(w_i, \boldsymbol{v}_i^{\mathrm{T}})^{\mathrm{T}}\} d\mu(\boldsymbol{v}_i)},$$

where $\mathcal{R}(\boldsymbol{v}_i)$ denotes the range of possible values of $\boldsymbol{v}_i = (y_{i2}, \ldots, y_{im_i})^{\mathrm{T}}$ such that $\sum_{i=1}^{m_i} y_{ij} = w_i$. When the event times are not censored, Y_{ij} takes values in $\{0, 1\}$ and the expectation $E(V_{ij}|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})$ is a discrete sum. Otherwise, Y_{ij} is a pseudo-value and takes values in the interval $[0, 1]$; the expectation will then involve a mix of discrete sums and integrations to account for the proper range of Y_{ij}. Determining the appropriate range $\mathcal{R}(\boldsymbol{v}_i)$ can be cumbersome if handled by brute-force especially when m_i is large. However, in Section S.1.6 (Supplementary Material), we provide a systematic approach that uses the built in function adaptIntegrate in R (Johnson and Narasimhan, 2013) to handle this complex problem.

A sumary of our method to estimate $\alpha(x, t), \boldsymbol{\beta}(t)$ at each $t = t_0$ is provided below.

Algorithm for estimating $\alpha(x, t), \boldsymbol{\beta}(t)$.

1. Choose $t_{01}, \ldots, t_{0L}$ evenly spaced across the range of $S_{ij} = \min(T_{ij}, C_{ij})$ for failure times T_{ij} and censoring times C_{ij}. In general, we recommend choosing $L \geq 5$ as it worked well in our empirical examples. The exact choice of L for a particular application will influence how smoothly $\boldsymbol{\beta}(t)$, $\alpha(x,t)$ is approximated over time t, where larger L will generally lead to more wiggly estimates of $\boldsymbol{\beta}(t), \alpha(x,t)$ compared to smaller L.
2. For each t_0 in Step 1, do the following:
   (a) Set $Y_{ij}(t_0) = I(T_{ij} \leq t_0)$ when $\Delta_{ij} = 1$ or $\Delta_{ij} = 0$ and $C_{ij} \geq t_0$. Otherwise, when $\Delta_{ij} = 0$ and $C_{ij} < t_0$, let $Y_{ij}(t_0)$ be the pseudo-value $Y^*_{ij}(t_0)$ (Example 1) or $Y^\dagger_{ij}(t_0)$ (Example 2) depending on the nature of event type.
   (b) Choose a set of spline basis functions $\boldsymbol{B}(\cdot)$ that does not include the intercept and has its knots at equally spaced sample quantiles of the observed X_{ij} values, $i = 1, \ldots, n$, $j = 1, \ldots, m_i$.
   (c) Set $\eta(X_{ij}, \boldsymbol{Z}_{ij}; \boldsymbol{\theta}) = \boldsymbol{B}^{\mathrm{T}}(X_{ij})\boldsymbol{a} + \boldsymbol{Z}_{ij}^{\mathrm{T}}\boldsymbol{\beta}$.
   (d) Set $W_i = \sum_{j=1}^{m_i} Y_{ij}$ and $\boldsymbol{V}_i = (Y_{i2}, \ldots, Y_{i,m_i})^{\mathrm{T}}$. Compute $E(V_{ij}|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})$ for all i, j. See Section S.1.6 for a systematic implementation.
   (e) Obtain $\widehat{\boldsymbol{a}}$ and $\widehat{\boldsymbol{\beta}}$ as the roots of the estimating equation in Proposition 3. Then set $\widehat{\boldsymbol{\beta}}(t_0) = \widehat{\boldsymbol{\beta}}$ and $\widehat{\alpha}(x_k, t_0) = \boldsymbol{B}^{\mathrm{T}}(x_k)\widehat{\boldsymbol{a}}$ for different x_k values evenly spread along the range of x.

Repeat Step 2 separately for each $t_{0\ell}$, $\ell = 1, \ldots, L$ to obtain estimates $\widehat{\alpha}(x_k, t_{0\ell})$ and $\boldsymbol{\beta}(t_{0\ell})$. For estimates at other t values within the range of $t_{01}, \ldots, t_{0L}$, use linear interpolation along t.

A few remarks are in order. First, our model and algorithm currently assume common $\alpha(x,t)$ and $\beta(t)$. With minor modifications, we can generalize the method to different functional coefficients such as $\alpha_j(X_{ij}, t)$ and $\beta_j(t)$. In this case, at each t_0, Step 2(c) has $\eta(X_{ij}, \boldsymbol{Z}_{ij}, \boldsymbol{\theta}) = \boldsymbol{B}_j^{\mathrm{T}}(X_{ij})\boldsymbol{a}_j + \boldsymbol{Z}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_j$, where $\boldsymbol{B}_j(\cdot)$ are potentially different sets of B-splines for each event type j. Also, the estimating equations in Step 2(e) are now $2m$-many with $\boldsymbol{S}_{\mathrm{eff},\boldsymbol{a}_j}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) = \{\boldsymbol{B}(X_{ij}) - \boldsymbol{B}(X_{i1})\}\{V_{ij} - E(V_{ij}|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})\}$ and $\boldsymbol{S}_{\mathrm{eff},\boldsymbol{\beta}_j}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) = (\boldsymbol{Z}_{ij} - \boldsymbol{Z}_{i1})\{V_{ij} - E(V_{ij}|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})\}$.

Second, our method can also adapt to non-event specific covariates with minor modifications. Consider the case that $m_i = 2$ for all i. One interest is modeling the effects of non-event specific covariates such as baseline covariates. For example, we could have $X_{i1} = X_{i2} \equiv X_i, \boldsymbol{Z}_{i1} = \boldsymbol{Z}_{i2} \equiv \boldsymbol{Z}_i$, so the covariates are the same for $j = 1$ and $j = 2$. Because the covariates are non-event specific, common functions of $X_i, \boldsymbol{Z}_i$ form part of the random effect and must be absorbed into $R_i(t)$. In this case, our method identifies

and estimates the logit differences of covariate effects. That is, we model

$$\begin{aligned} \text{logit}[\text{pr}\{T_{i1} < t | X_i, \boldsymbol{Z}_i, R_i(t)\}] &= \alpha(X_i, t) + \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\beta}(t) + R_i(t), \\ \text{logit}[\text{pr}\{T_{i2} < t | X_i, \boldsymbol{Z}_i, R_i(t)\}] &= R_i(t), \end{aligned} \tag{2.5}$$

and estimate $\alpha(x,t)$, and $\boldsymbol{\beta}(t)$ which represent effects of a covariate on the conditional log odds ratio of events $T_{i1} < t$ and $T_{i2} < t$ given random effects and covariates. The parametrization in (2.5) is seen in some joint modeling of longitudinal outcomes and censored outcomes literature (Rizopoulos et al., 2011), where longitudinal model components enter a survival model through shared random frailty terms. Such parametrization allows flexible estimates of $\alpha(x,t), \boldsymbol{\beta}(t)$ which are important in comparing the odds of events as in the HD application (Section 4). In terms of estimation, our algorithm remains the same except Step 2(c) uses $\eta_{i1}(X_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) = \boldsymbol{B}^{\mathrm{T}}(X_i)\boldsymbol{a} + \boldsymbol{Z}_i\boldsymbol{\beta}$ and $\eta_{i2}(X_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) = 0$, and the estimating equations in Step 2(e) are $\boldsymbol{S}_{\text{eff},\boldsymbol{a}}(\boldsymbol{Y}_i, X_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) = \boldsymbol{B}(X_i)\{V_i - E(V_i|W_i, X_i, \boldsymbol{Z}_i; \boldsymbol{\theta})\}$ and $\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, X_i, \boldsymbol{Z}_i) = \boldsymbol{Z}_i\{V_i - E(V_i|W_i, X_i, \boldsymbol{Z}_i; \boldsymbol{\theta})\}$.

It is important to note that interpretations of marginal effects in equation (2.5) are not possible. Because we make no distributional assumptions about the random effect, we cannot integrate over it to obtain the marginal effect of the covariates. Therefore, $\alpha(x,t), \boldsymbol{\beta}(t)$ solely represent the effect of the logit-differences between the distributions.

Third, asymptotic properties of $\widehat{\alpha}(x,t)$ and $\widehat{\boldsymbol{\beta}}(t)$ are developed in Appendix A. In summary, $\widehat{\alpha}(x,t)$ and $\widehat{\boldsymbol{\beta}}(t)$ are shown to be asymptotically consistent and normally distributed. We establish the asymptotic variability of each, but in practice, we recommend a bootstrap variability as described in Section 2.3.

2.3. *Features of the proposed estimator.* A major advantage of our approach is that the construction of the score vectors $\boldsymbol{S}_{\text{eff},\boldsymbol{a}}$ and $\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}$ in Proposition 3 completely breaks free of the unknown density $f_{\boldsymbol{X},\boldsymbol{Z},R}(\boldsymbol{x}, \boldsymbol{z}, r)$. This means we can construct the score vectors without estimating the unknown random effect distribution or postulating potentially incorrect parametric forms. Doing so is useful in practice since it is almost impossible to know the random effect distribution a priori.

A second advantage is that our approach yields consistent estimators whether the random effect and covariates are independent or not. Traditionally, the random effect is considered independent of the covariates, but such an assumption can be invalid in biological studies. For example, in a model for repeatedly measured apathy responses, Heagerty and Kurland

(2001) showed that the variability of the random effect depended on the covariate gender. Govindarajulu et al. (2007) likewise demonstrated that the random effect in a frailty model for the Framingham Heart Study depended on patient covariates. An appropriate model should thus accommodate dependency between the random effect and covariates when necessary.

Testing for dependency between covariates and the random effect can be accomplished using the Hausman (1978) chi-squared test which tests the null hypothesis that the covariates and the random effect are independent. At a single time point t_0, the test involves comparing results from our proposed method which makes no restrictions on the dependency between $(X, \boldsymbol{Z})$ and $R(\cdot)$, and a method which imposes independence between $(X, \boldsymbol{Z})$ and $R(\cdot)$. A method that is a competitor to ours but assumes independence between $(X, \boldsymbol{Z})$ and $R(\cdot)$ is the generalized additive mixed model (Wood, 2008, GAMM). At each $t = t_0$, GAMM views the model in (1.1) as an additive mixed effects model and estimates parameters using a penalized likelihood and automatic selection of multiple smoothing parameters to capture the functional parameter shapes. We show in Section 3 that when $(X, \boldsymbol{Z})$ and $R(\cdot)$ are independent, GAMM estimates are consistent and otherwise, they are not. It is this flip between consistent and inconsistent estimates that drives the results of the Hausman chi-squared test.

To form the Hausman chi-squared test at $t = t_0$: (i) compute the estimates obtained from our method denoted as $\widehat{\boldsymbol{\psi}}(X, t_0) = \{\widehat{\alpha}(X, t_0), \widehat{\boldsymbol{\beta}}^T(t_0)\}^T$; and (ii) compute the estimates obtained from GAMM denoted as $\widehat{\boldsymbol{\psi}}_{\mathrm{IND}}(X, t_0) = \{\widehat{\alpha}_{\mathrm{IND}}(X, t_0), \widehat{\boldsymbol{\beta}}^T_{\mathrm{IND}}(t_0)\}^T$ where the notation IND emphasizes the independence assumption between $(X, \boldsymbol{Z})$ and $R(\cdot)$. Under the null hypothesis, both $\widehat{\boldsymbol{\psi}}(X, t_0)$ and $\widehat{\boldsymbol{\psi}}_{\mathrm{IND}}(X, t_0)$ are consistent, and under the alternative, $\widehat{\boldsymbol{\psi}}(X, t_0)$ is consistent and $\widehat{\boldsymbol{\psi}}_{\mathrm{IND}}(X, t)$ is not. Therefore, a statistically significant difference between $\widehat{\boldsymbol{\psi}}(X, t_0)$ and $\widehat{\boldsymbol{\psi}}_{\mathrm{IND}}(X, t_0)$ is evidence in favor of dependency between $(X, \boldsymbol{Z})$ and $R(\cdot)$. The Hausman chi-squared test statistic is $\{\widehat{\boldsymbol{\psi}}(X, t_0) - \widehat{\boldsymbol{\psi}}_{\mathrm{IND}}(X, t_0)\}^T [\mathrm{var}\{\widehat{\boldsymbol{\psi}}(X, t_0)\} - \mathrm{var}\{\widehat{\boldsymbol{\psi}}_{\mathrm{IND}}(X, t_0)\}]^{-1} \{\widehat{\boldsymbol{\psi}}(X, t_0) - \widehat{\boldsymbol{\psi}}_{\mathrm{IND}}(X, t_0)\}$, and it follows a chi-squared distribution with k degrees of freedom, where $k = \mathrm{rank}[\mathrm{var}\{\widehat{\boldsymbol{\psi}}(X, t_0)\} - \mathrm{var}\{\widehat{\boldsymbol{\psi}}_{\mathrm{IND}}(X, t_0)\}]$. See Hausman (1978) for the derivation of the test statistic.

Extending the Hausman chi-squared test over a range of time points t_0 can be achieved with graphical methods. Plot $\widehat{\boldsymbol{\psi}}(X, t)$ from our method and $\widehat{\boldsymbol{\psi}}_{\mathrm{IND}}(X, t)$ from GAMM over $t = t_{01}, \ldots, t_{0L}$ using linear interpolation to connect estimates between the $t_{0\ell}$'s, $\ell = 1, \ldots, L$. In addition, plot the 95% confidence band associated with each estimate. A confidence band for our method is obtained using a bootstrap approach. For GAMM, it is formed using estimated variances from Bayesian principles implemented in the mgcv

package in R (Wood, 2008). For our method, a bootstrap data set is obtained by randomly selecting n clusters (with replacement), keeping the cluster membership in tact. That is, we randomly select among the cluster groups, not among the individual cluster members. We then apply our algorithm to each bootstrap data set and obtain the corresponding parameter estimates at $t_{01}, \ldots, t_{0L}$. The 95% bootstrap confidence band is then formed by first computing the percentile bootstrap confidence interval at each $t_{0\ell}$ (i.e., the 2.5th and 97.5th percentiles of the bootstrap estimates at each $t_{0\ell}$), and then connecting the bootstrap confidence interval across the $t_{0\ell}$'s using linear interpolation. In our application, we found $B = 100$ bootstrap data worked well.

After forming the 95% confidence band for estimates from our method and from GAMM, we then compare the two to asses the null hypothesis that the covariates and random effect are independent. If the confidence bands overlap, the null hypothesis is not rejected. Otherwise, if the confidence bands do not overlap at least at one t, then the null hypothesis is rejected and thus there is evidence of dependency between the covariates and random effect. We apply this graphical test to our Huntington's disease application in Section 4.

Although the Hausman chi-squared test and its graphical version are helpful for determining dependency between covariates and the random effect, it does not specify how to model the dependency. Existing methods that do model such dependency involve multiple mixed effects models (Heagerty and Kurland, 2001) or intensive Monte Carlo Markov Chain computations (Govindarajulu et al., 2007), both of which are computationally burdensome. Our approach is advantageous in this respect in that it is computationally simple and does not require testing for dependency beforehand.

A last advantage is our method's simplicity in constructing estimating equations. The most involved computation is the expectation $E(V_{ij}|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})$ which is, at worst, a combination of discrete sums and numerical integrations that can be systematically carried out in R with the adaptIntegrate function (see Supplementary Material S.1.6).

3. Simulation Study.

3.1. *Simulation design.* We evaluated the performance of our method for different random effect distributions and different dependencies between the random effect and model covariates. Because at each $t = t_0$ we view the model in (1.1) as an additive logistic mixed effects model, a competitor is the generalized additive mixed model (Wood, 2008, GAMM) as described in Section 2.3. GAMM is well developed theoretically (Wood, 2008) and

computationally (i.e., mgcv package in R), but, compared to our method, assumes that the random effect is normally distributed and independent of model covariates. Our simulation study is designed to investigate the sensitivity of these assumptions. We show results for clustered failure times formed from single event types (Example 1) here, and from multiple event types (Example 2) in the Supplementary Material (Section S.2).

To assess sensitivity to non-normally distributed random effects, we considered different distributional forms for $R(t)$. We generated $R(t) = R$ where (i) R is Normal$(0, 1)$; (ii) R is a mixture with 50% from Normal$(-1, 1)$ and 50% from Normal$(1, 0.25^2)$; (iii) R is a mixture with 50% from Normal$(-1, 1)$ and 50% from Beta$(4, 2)$; and (iv) R is Uniform$[-2.5, 2.5]$. Setting (i) is the standard assumption in GAMM, while the others strongly deviate from the normality assumption. Covariates X_{ij} were generated from a Uniform[0,1] distribution and Z_{ij} from a Uniform[1,2] distribution.

To assess sensitivity to dependence between covariates and the random effect, we generated R_i using the distributions (i)-(iv) above and added another complexity. We set $R_i(t) = R_i + b_i$, and $X_{ij} = X^*_{ij} + b_i$ where b_i is Normal$(0, 0.05^2)$ and X^*_{ij} is Uniform$[0, 1]$. Lastly, we generated Z_{ij} from Uniform$[R_i - 1, R_i + 1]$. Under this setup, both X_{ij} and Z_{ij} depend on R_i.

In all settings, we set $n = 500, m_i = 3$ and simulated 1000 data sets from model (1.1) with $\alpha(x, t) = 3\sin(\pi x)\log(t/50)$, $\beta(t) = 2\log(t/50)$, both of which are nonlinear. At each t, the generated random intercept $R_i(t)$ were centered, in accordance with the assumptions of GAMM, and event times are 40% censored from an independent, uniformly distributed censoring time C_{ij}. We applied our method and GAMM to estimate $\alpha(x, t), \boldsymbol{\beta}(t)$ across 18 equally spaced times t in [40,50] and 100 equally spaced x-values in [0,1].

We evaluated bias, empirical variance, estimated variance and 95% coverages at specific t- and x-choices ($t = 46$, $x = 0.50$) and averaged across t in $[40, 50]$ and x in $[0, 1]$. For bias, we report pointwise bias at $t = 46$, $x = 0.50$. We also report bias over a range of t (or x values) through average absolute bias, calculated via $\sum_{\ell=1}^{L} |\overline{\widehat{\beta}(t_{0\ell})} - \beta_0(t_{0\ell})|/L$, and similarly for $\alpha(x, t)$. Here, $\beta_0(t_{0\ell})$ is the truth and $\overline{\widehat{\beta}(t_{0\ell})}$ is the average estimate based on 1000 simulations.

Lastly, estimated variances for our method are bootstrap-based as described in Section 2.3. For GAMM, estimated variances are obtained using the implemented Bayesian variance calculations in the mgcv package in R.

3.2. *Simulation results.* Regardless of the random effect's distribution or dependency between covariates and random effect, our method unbiasedly estimates $\alpha(x, t)$ and $\beta(t)$. This is evident from the negligible average

absolute bias of parameter estimates (Table 1), negligible pointwise bias (Table 2), and Figure 1. In Figure 1, estimates from our method (red dashed) overlap the true curves (black solid) when $R \sim \text{Normal}(0, 1)$. Similar unbiasedness is observed for different distributions of R. The estimated bootstrap variances of our method also closely match the empirical variances (Table 2) and the 95% coverage probabilities match the nominal levels in all settings. These numerical and graphical results exemplify that our method is robust to the true and unknown properties of the random effect and its distribution.

Interestingly, GAMM estimates are not sensitive to deviations from non-normally distributed random effects. Average absolute biases (top-half of Table 1) and pointwise biases (left-half of Table 2) of the estimated parameters remain negligible for all distributional assumptions of the random effect. The empirical robustness to non-normal random effect distributions is similar to that seen in Glidden and Vittinghoff (2004), Hsu et al. (2007) and Gorfine et al. (2012). They observed robustness in gamma frailty models when the true frailty was not gamma distributed.

Despite this observed robustness to distributional deviations, GAMM is quite sensitive when X and R are dependent, and Z and R are dependent. (In simulations not shown, dependence between X and Z, but not with R did not appear to affect the performance of GAMM.) The bias is visually evident in Figure 1 (right half) where the GAMM estimates (blue dashed-dotted curve) completely miss the true $\alpha(x, t)$ and $\beta(t)$ curves (black solid curves) across x and t. The strong biases lead to 95% coverage probabilities that are far from the nominal level and lead to inflated mean squared errors (MSE). In all settings where the covariates and random effect are dependent, the MSE for GAMM estimates are nearly twice as large as the MSE for estimates from the proposed method (Tables 1 and 2). These results illustrate that GAMM is limited to situations where covariates and random effects are independent. However, such an assumption is not always valid in biological studies (Heagerty and Kurland, 2001; Govindarajulu et al., 2007).

Evident in all settings (Tables 1, 2) is that our method has more variability than GAMM. This larger variability is expected and is an artifact of our assumption that the random effect has an unknown and unspecified distribution. The fact that about 7% of jackknife pseudo-values fall outside [0,1] (Table S.1 in Supplementary Material) does not contribute to the larger variability. As evidence of this, we re-performed our simulation study at 0% censoring in which case, all event times are observed and no pseudo-values are introduced into the estimation process. Results in Table S.4 (Supplementary Material) show that even without pseudo-values, our proposed method has larger variability than does GAMM. This indicates that the unspecified

distribution of the random effects drives the increased variability.

Stronger model assumptions such as those in GAMM always reduce the variability of the parameter estimates. Though GAMM's efficiency initially appears advantageous, its effect is actually the opposite, especially when the estimates are biased. When estimates are biased and have small variability, this leads to a false conclusion that the researcher treats with great confidence as correct. This is worse than any method which yields confidence intervals that indeed have a promised chance of covering the truth. In addition, when GAMM estimates are biased, the resulting mean squared errors are always larger than those from our method, indicating that the high bias outweighs our method's larger variability.

4. Application to a Huntington's Disease (HD) Study.

4.1. *Clinical research problem.* We applied our method and GAMM to COHORT, a large observational study of HD that evaluated failure-type events representative of the disease course. From 2005-2011, COHORT collected information from 1,293 symptomatic or at-risk adults, including gender, number of CAG repeats, and the ages when certain events occurred that most impacted a person's normal life. These events include

1. Age when a subject first experiences a motor sign (i.e., chorea, dystonia, rigidity). Reported ages are either from (i) a trained rater, or (ii) the subject if the rater did not observe a motor sign but the subject did, or (iii) a family member if neither the rater nor the subject observed a motor sign but a family member did. Among the 1,293 subjects, 774 subjects (59.8%) experienced a first motor sign during the COHORT study, and 519 did not (i.e., age of first motor sign had 40.1% censoring). For those who experienced a first motor sign, 75.1% had the ages of first motor sign determined by a rater, 19.8% were self-determined, and 5.2% were determined by a family member.
2. Age when cognitive impairments first impact daily life. This age is patient-reported in response to "At what age did cognitive impairment impact your daily life?", and we report the first age of occurrence. Among the 1,293 subjects, 385 subjects (29.8%) experienced impacting cognitive impairment and 908 subjects did not (i.e., age of impacting cognitive impairment had 70.2% censoring).

The data are an example of clustered failure times where, for each subject, a cluster is formed by the two event types measured on that subject. For subject i, we let T_{i1} be the age of first motor sign, and T_{i2} be the age of impacting cognitive impairment. Censoring for both events is largely ad-

ministrative (i.e., the study period ends before the event of interest occurs), so assuming covariate independent censoring, as our method does, is appropriate here. We let Z_i denote gender (Z_i=1 corresponds to a male) and X_i denote the subject's CAG repeat-length. Forty-four percent of the subjects were males, and we focused on those individuals with CAG repeat-lengths 39 to 50 since very few had repeats outside this range (81% of subjects had less than 45 CAG repeats).

Given that the covariates are non-event specific, the data are modeled using model (2.5). The model will help determine whether a motor sign or cognitive sign has higher odds of occurring first. Knowing which sign occurs first facilitates prioritizing these features as endpoints in clinical trial planning and assists in disease management.

The parameters $\alpha(x,t)$, $\boldsymbol{\beta}(t)$ in (2.5) represent the logit-differences between the distributions for T_1, T_2. That is, $\alpha(x,t)$ represents the difference in how CAG repeat-length affects the log odds of the dichotomized time T_1 compared to that of T_2. Likewise, $\beta(t)$ represents the difference in how gender affects the same two log odds of dichotomized survival times. We discuss the importance of these functional parameters in relation to HD in Section 4.2. Using our approach and GAMM, we estimated $\alpha(x,t), \boldsymbol{\beta}(t)$ over the range of CAG repeat-lengths and for t in 35 to 60 (age measured in years) at $t_0 = 35, 40, \cdots, 60$. CAG repeats (X) were standardized to be in [0,1] and we estimated $\alpha(x,t)$ at 11 equally spaced points in this interval. Estimated variances for $\widehat{\boldsymbol{\beta}}(t)$ and $\widehat{\alpha}(x,t)$ from our method were obtained using 100 bootstrap replicates. We found that less than 5% of the pseudo-values computed for the HD data fell outside [0,1] (Table S.5, Supplementary Material). This observation is similar to pseudo-values computed for simulated data in Section 3.

4.2. *Results.* Prior to applying our method to the COHORT study, we confirmed that our method performed well at 70% censoring similar to that observed in COHORT (see Section S.2, Supplementary Material).

We evaluated for dependence between the random effect and covariates by comparing results from our proposed method and GAMM. As described in Section 2.3, dependence is evident if the 95% confidence bands of the estimates from our method and GAMM do not always overlap. Figure 2 shows results from both methods for $\widehat{\beta}(t)$ and $\widehat{\alpha}(x,t)$ over t in [35,60] and at $x = 40$ CAG repeats. While the 95% confidence bands clearly overlap for $\widehat{\beta}(t)$, they do not for $\widehat{\alpha}(x,t)$ when $t \geq 40$. The non-overlapping confidence bands is evidence of dependence between the covariates and random effect. From our simulation results (Section 3), this means that results from GAMM

will be biased as it is not designed to handle dependence between covariates and the random effect. We now discuss and compare results from our method and GAMM.

Our method found that $\beta(t)$ is not significantly different from zero (Table 3, left upper half). This implies that gender has the same effect on the likeliness of a first motor sign occurring before age t ($35 \leq t \leq 60$) as it does on impacting cognitive impairment occurring before age t. The result agrees with earlier studies where no gender effect was found in the mean survival time of HD patients (Harper, 1996) and HD progression (Marder et al., 2000).

Our method estimated $\alpha(x,t)$ to be positive and, for the most part, significantly different from zero (Table 3, left lower half). Thus, it is more likely that a first motor sign occurs before a patient self-reports impacting cognitive impairment. The implication of the positive $\alpha(x,t)$ is best understood through the conditional odds ratio comparing the odds of observing a first motor sign before age t to the odds of impacting cognitive impairment before t given CAG repeat length, gender, and random intercept. The conditional odds ratios, computed as $\exp\{\widehat{\alpha}(X,t) + \widehat{\beta}(t)Z\}$, are given in Table 4 (first column) and indicate odds in favor of observing a first motor sign. For example, for a male with 40 CAG repeats, the conditional odds of a first motor sign occurring before age 50 is 4.264 (95% CI: 1.671, 12.013) times the odds of impacting cognitive impairment occurring before age 50. For a male with 46 CAG repeats and at age 50, the conditional odds ratio increases to 14.171 (95% CI: 4.61, 53.88). This conditional odds in favor of a first motor sign is similarly observed with females.

These conditional odds highlight the challenges of relying on self-reported cognitive signs. Some clinical studies suggest that cognitive impairments emerge years before a motor-diagnosis and, perhaps, even before first motor impairments (Stout et al., 2011). In contrary, our results estimate odds to favor a first motor impairment. But, progression of cognitive decline is gradual and often too slow to detect from a subject's perspective (Stout et al., 2011). This means there is often a long delay before a subject realizes his cognitive impairment is *impacting* his daily life. This delay could sensibly lead to observing a motor impairment first. Alternative measures that do not focus on the impact of cognitive impairments but rather on the effect itself include mild cognitive impairment (Duff et al., 2010): when a subject's cognitive exam score is 1.5 standard deviations below the mean of the cognitive scores for healthy controls. Examining mild cognitive impairment for COHORT is difficult, however, since the study is primarily a retrospective one, and so the cognitive exam scores for subjects are unavailable. Therefore,

future work with prospective studies assessing mild cognitive impairment in addition to ages of first motor symptom would be of interest.

Results from GAMM (Table 3, right, lower half) generally agreed with our method, except that $\widehat{\alpha}(x,t)$ was at times negative and statistically significant, and at other times, positive and statistically significant. Understanding these sign changes is again best illustrated through conditional odds ratios (Table 4, second column). GAMM suggests that for a male with 40 CAG repeats, the conditional odds of experiencing a first motor sign before age 50 is 0.637 times (95% CI: 0.507, 0.766) times the odds of impacting cognitive impairment occurring before age 50. The result reverses at 46 CAG repeats in that the conditional odds ratio is 3.377 (95% CI: 2.454, 4.3). Similar results are observed for females. Thus, according to GAMM, having 40 CAG repeats increases the conditional odds of impacting cognitive impairment occurring, whereas having 46 CAG repeats increases the odds of experiencing a first motor sign. The flip between cognitive and motor signs having increased coditional odds could be an artifact of the discrepancies observed with GAMM (see Section 3) in that it is sensitive to violations of independence between covariates and random effects. Or the flip could be due to an age-effect as subjects with 40 CAG repeats are, on average, 20 years older than those with 46 repeats.

4.3. *Practical impacts on HD research.* In summary, results from our method indicate higher conditional odds of a first motor impairment occuring before impacting cognitive impairment. However, given the nuances of self-reporting measures, our work highlights the need for better cognitive assessments that are objective and that can be measured prospectively. This is important for deciding whether to prioritize cognitive or motor impairments in a clinical trial, as well as deciding how best to intervene with disease impairments (i.e., whether treatments should target cognitive or motor impairments first). Active work in this area is ongoing (Duff et al., 2010; Paulsen and Long, 2014).

In addition, the difference effects for CAG repeats, $\alpha(x,t)$, are nonlinear and vary over x and t (see Figure 3). This result adds a new time-component to the current modeling standard in the clinical literature (Langbehn et al., 2004) which models the effect of CAG repeat-length independent of time. Our graphical results can thus supplement the existing findings from Langbehn et al. (2004) to inform clinicians on the the changes CAG repeats have over time.

5. Discussion. To model the distribution of clustered failure times, we present a novel approach that does not model the intra-class correlation with

a parametric random effect or assume independence between the random effect and covariates. The covariates are modeled using unknown functional forms. The random effect is kept free of any distributional assumptions and is allowed to correlate with some or all covariates. The pseudo-value (Logan et al., 2011) viewpoint allows us to simultaneously handle censoring and derive semiparametric techniques to bypass estimation that directly involves the random effects.

Our estimation procedure is computationally simple, and does not require estimating nor positing a working model for the unknown random effect distribution. Our approach thus circumvents the challenges of modeling dependencies between covariates and random effects which can be detected with a Hausman chi-squared test or graphically, but cannot be precisely defined. Standard methods (e.g., GAMM) assume independence between covariates and random effects, but they can be severely biased (Section 3) or lead to inconclusive results in real applications (e.g., whether motor or cognitive signs in HD have higher odds of occurring first).

Our estimation procedure avoids the problems of modeling the random effect distribution because the estimating equation in Proposition 3 does not have any terms that include the random effect. Our model is an example of a generalized linear mixed effect model which does not include a dispersion parameter nor random slopes. In the general case where the dispersion parameter is unknown and/or the model has a random slope, the ensuing estimating equation would have terms involving the random effect. Forming the estimating equation in this case would then require a working model for the random effect distribution. The computation becomes more involved, but Ma and Genton (2010) contain details for potential working models that can simplify the calculations.

The estimating equations in Step 2(d) of our algorithm is solved separately for each t_0. That is, we do not simultaneously solve L-many sets of estimating equations formed at $t_{01}, \ldots, t_{0L}$. This is because we do not assume smoothness of the parameters $\alpha(x, t)$ and $\boldsymbol{\beta}(t)$ as functions of t. Hence, theoretically speaking, there is no further information that could be gained by considering all time points simultaneously.

A potential extension worth pursuing is developing an estimation procedure when we assume the functional parameters are smooth over time t. Two potential solutions are kernel method and spline method. With the former, we could combine the current estimation equations or combine the current estimators at each t via a weighted average, where the weights are formed by kernels centered at t_0. This would allow us to borrow information around t_0 in estimating the parameter values at t_0. With the latter, improvement could

be achieved by using splines to express the parameter functions into linear combination of spline bases, and then estimating the common parameters across all the different t values.

When implementing these solutions, computational complications may arise. For example, we would need to consider how best to choose the bandwidth, how many time points should be considered in the estimation, and how should the distribution of time points be considered to guarantee a gain instead of a loss. We would also need to make careful decisions on how to choose the bases spline functions, as well as how best we could take into account the correlation of the estimators across the different t's. These issues require future investigation.

SUPPLEMENTARY MATERIAL

Technical Proofs and Empirical Results
(doi: COMPLETED BY THE TYPESETTER). The supplementary material contains theoretical derivations, additional simulation study results, and additional results for the Huntington's disease application.

TABLE 1

Average results for clustered failure times with single event types. 40% censoring, 1000 simulations. Average absolute bias, empirical variance, 95% coverage probabilities and mean squared error (MSE) when the true random intercept is as specified. $\widehat{\beta}(\cdot)$ *denotes results averaged over* t*;* $\widehat{\alpha}(0.50,\cdot)$ *is results at* $x = 0.50$ *averaged over* t*, and* $\widehat{\alpha}(\cdot,46)$ *is results at* $t = 46$ *averaged over* x*.*

	Proposed Method			GAMM Method		
	$\widehat{\beta}(\cdot)$	$\widehat{\alpha}(0.50,\cdot)$	$\widehat{\alpha}(\cdot,46)$	$\widehat{\beta}(\cdot)$	$\widehat{\alpha}(0.50,\cdot)$	$\widehat{\alpha}(\cdot,46)$
X, Z, R independent						
			$R \sim \text{Normal}(0,1)$			
abs bias	0.029	0.040	0.034	0.021	0.032	0.014
emp var	0.075	0.354	0.317	0.015	0.043	0.042
95% cov	0.950	0.944	0.948	0.951	0.948	0.951
MSE	0.075	0.356	0.319	0.016	0.044	0.042
			$R \sim 0.5\text{Normal}(-1,1) + 0.5\text{Normal}(1, 0.25^2)$			
abs bias	0.009	0.044	0.021	0.012	0.028	0.005
emp var	0.089	0.391	0.358	0.016	0.046	0.044
95% cov	0.946	0.946	0.948	0.946	0.947	0.951
MSE	0.089	0.393	0.358	0.016	0.047	0.044
			$R \sim 0.5\text{Normal}(-1,1) + 0.5\text{Beta}(4,2)$			
abs bias	0.029	0.037	0.032	0.051	0.050	0.041
emp var	0.082	0.290	0.273	0.014	0.041	0.039
95% cov	0.944	0.948	0.947	0.932	0.945	0.947
MSE	0.083	0.292	0.274	0.017	0.044	0.041
			$R \sim \text{Uniform}[-2.5, 2.5]$			
abs bias	0.015	0.062	0.055	0.011	0.017	0.005
emp var	0.096	0.426	0.400	0.013	0.036	0.035
95% cov	0.952	0.950	0.950	0.950	0.947	0.950
MSE	0.097	0.430	0.403	0.013	0.037	0.035
(X, Z) and R dependent						
			$R \sim \text{Normal}(0,1)$			
abs bias	0.026	0.031	0.012	1.260	1.672	1.714
emp var	1.018	1.257	1.161	0.106	0.224	0.230
95% cov	0.949	0.947	0.949	0.010	0.037	0.024
MSE	1.019	1.258	1.162	1.693	3.023	3.377
			$R \sim 0.5\text{Normal}(-1,1) + 0.5\text{Normal}(1, 0.25^2)$			
abs bias	0.090	0.038	0.032	2.002	2.498	2.534
emp var	1.046	1.337	1.264	0.164	0.292	0.304
95% cov	0.944	0.944	0.945	0.000	0.001	0.001
MSE	1.055	1.338	1.265	4.174	6.537	7.179
			$R \sim 0.5\text{Normal}(-1,1) + 0.5\text{Beta}(4,2)$			
abs bias	0.027	0.034	0.020	1.252	1.741	1.800
emp var	0.880	1.201	1.066	0.088	0.207	0.223
95% cov	0.947	0.948	0.946	0.001	0.015	0.014
MSE	0.881	1.203	1.066	1.658	3.251	3.692
			$R \sim \text{Uniform}[-2.5, 2.5]$			
abs bias	0.028	0.018	0.014	2.104	2.802	2.869
emp var	0.947	1.278	1.134	0.177	0.384	0.411
95% cov	0.948	0.945	0.947	0.000	0.000	0.001
MSE	0.948	1.278	1.134	4.608	8.252	9.225

TABLE 2

Pointwise results for clustered failure times with single event types. 40% censoring, 1000 simulations. Pointwise bias, empirical variance, estimated variance, 95% coverage probabilities and mean squared error (MSE) for $\widehat{\beta}(t)$ and $\widehat{\alpha}(x,t)$ at $t = 46$ and $x = 0.50$ when the true random intercept is as specified.

	X, Z, R independent				(X, Z) and R dependent			
	Proposed Method		GAMM Method		Proposed Method		GAMM Method	
	$\widehat{\beta}(46)$	$\widehat{\alpha}(0.50, 46)$	$\widehat{\beta}(46)$	$\widehat{\alpha}(0.50, 46)$	$\widehat{\beta}(46)$	$\widehat{\alpha}(0.50, 46)$	$\widehat{\beta}(46)$	$\widehat{\alpha}(0.50, 46)$
				$R \sim \text{Normal}(0, 1)$				
bias	-0.034	-0.051	0.017	0.016	-0.010	-0.026	1.277	-1.710
emp var	0.078	0.355	0.014	0.040	1.040	1.571	0.107	0.227
est var	0.083	0.367	0.013	0.036	1.037	1.153	0.041	0.092
95% cov	0.953	0.947	0.928	0.942	0.952	0.898	0.000	0.003
MSE	0.084	0.369	0.013	0.036	1.038	1.153	1.672	3.017
				$R \sim 0.5\text{Normal}(-1, 1) + 0.5\text{Normal}(1, 0.25^2)$				
bias	-0.018	-0.048	-0.004	0.002	0.057	0.018	2.014	-2.534
emp var	0.089	0.393	0.015	0.044	1.094	1.659	0.166	0.301
est var	0.096	0.421	0.014	0.039	1.108	1.254	0.047	0.090
95% cov	0.948	0.959	0.935	0.936	0.936	0.901	0.000	0.000
MSE	0.096	0.424	0.014	0.039	1.111	1.255	4.101	6.511
				$R \sim 0.5\text{Normal}(-1, 1) + 0.5\text{Beta}(4, 2)$				
bias	-0.021	-0.049	0.045	0.042	-0.013	-0.071	1.279	-1.795
emp var	0.080	0.297	0.013	0.037	0.955	1.477	0.091	0.211
est var	0.083	0.369	0.013	0.036	0.925	1.157	0.035	0.091
95% cov	0.953	0.969	0.926	0.936	0.943	0.910	0.000	0.000
MSE	0.083	0.372	0.015	0.038	0.925	1.162	1.671	3.312
				$R \sim \text{Uniform}[-2.5, 2.5]$				
bias	-0.016	-0.098	0.010	0.003	0.030	-0.063	2.137	-2.874
emp var	0.095	0.418	0.012	0.034	1.070	1.509	0.190	0.408
est var	0.102	0.449	0.012	0.034	1.051	1.296	0.040	0.093
95% cov	0.958	0.959	0.956	0.946	0.949	0.923	0.000	0.000
MSE	0.102	0.458	0.012	0.034	1.052	1.300	4.607	8.350

TABLE 3

Parameter estimates for COHORT study when comparing distributions for age of first motor sign to age when cognitive impairment first impacts daily life. Estimated $\widehat{\beta}(t)$ and $\widehat{\alpha}(x,t)$ and 95% confidence intervals (in parentheses) for different CAG repeats and ages t in years.

CAG	Age	Proposed Method	GAMM Method
		$\widehat{\beta}(t)$	
	35	0.147 (-1.729, 1.308)	0.399 (0.079, 0.72)
	40	-0.173 (-1.003, 0.453)	0.225 (-0.066, 0.515)
	45	-0.249 (-1.445, 0.424)	0.016 (-0.235, 0.268)
	50	0.388 (-0.468, 1.199)	-0.01 (-0.232, 0.212)
	55	0.135 (-0.552, 0.766)	0.019 (-0.189, 0.226)
	60	0.077 (-0.406, 0.604)	0.102 (-0.097, 0.302)
		$\widehat{\alpha}(x,t)$	
40	35	1.527 (-0.771, 4.012)	-0.604 (-0.817, -0.391)
40	40	0.684 (-0.294, 1.869)	-0.784 (-0.977, -0.591)
40	45	0.785 (0.022, 1.892)	-0.764 (-0.921, -0.607)
40	50	1.063 (0.362, 1.802)	-0.441 (-0.57, -0.312)
40	55	1.37 (0.806, 1.899)	-0.156 (-0.271, -0.041)
40	60	0.719 (0.168, 1.087)	0.043 (-0.065, 0.151)
46	35	1.229 (0.455, 2.09)	1.075 (0.773, 1.378)
46	40	1.424 (0.586, 1.964)	1.127 (0.852, 1.402)
46	45	1.315 (0.54, 2.078)	0.99 (0.731, 1.249)
46	50	2.264 (1.303, 3.301)	1.227 (0.972, 1.482)
46	55	2.601 (1.846, 3.344)	1.441 (1.183, 1.699)
46	60	2.012 (1.403, 2.693)	1.697 (1.433, 1.96)

TABLE 4

Conditional odds ratio estimates for COHORT study when comparing distributions for age of first motor sign (T_1) to age when cognitive impairment first impacts daily life (T_2). Estimated odds ratio (OR) for $T_1 < t$ compared to $T_2 < t$ conditional on gender, fixed CAG, and fixed random effect. Estimates shown along with 95% confidence intervals (in parentheses) for different CAG repeats and ages t in years.

CAG	Age	Proposed Method	GAMM Method
		OR for $T_1 < t$ compared to $T_2 < t$ for females	
40	40	1.982 (0.745, 6.484)	0.456 (0.368, 0.544)
40	45	2.193 (1.023, 6.631)	0.466 (0.393, 0.539)
40	50	2.894 (1.436, 6.064)	0.643 (0.56, 0.726)
40	55	3.935 (2.24, 6.676)	0.856 (0.757, 0.954)
46	40	4.152 (1.797, 7.127)	3.086 (2.237, 3.935)
46	45	3.724 (1.716, 7.99)	2.692 (1.995, 3.388)
46	50	9.618 (3.68, 27.145)	3.411 (2.54, 4.283)
46	55	13.471 (6.335, 28.341)	4.227 (3.136, 5.318)
		OR for $T_1 < t$ compared to $T_2 < t$ for males	
40	40	1.667 (0.779, 4.587)	0.571 (0.414, 0.729)
40	45	1.709 (0.754, 4.959)	0.474 (0.362, 0.585)
40	50	4.264 (1.671, 12.013)	0.637 (0.507, 0.766)
40	55	4.504 (2.653, 10.813)	0.872 (0.711, 1.033)
46	40	3.492 (1.657, 6.928)	3.863 (2.713, 5.013)
46	45	2.902 (1.358, 6.947)	2.736 (1.967, 3.505)
46	50	14.171 (4.61, 53.88)	3.377 (2.454, 4.3)
46	55	15.422 (7.23, 36.989)	4.307 (3.128, 5.486)

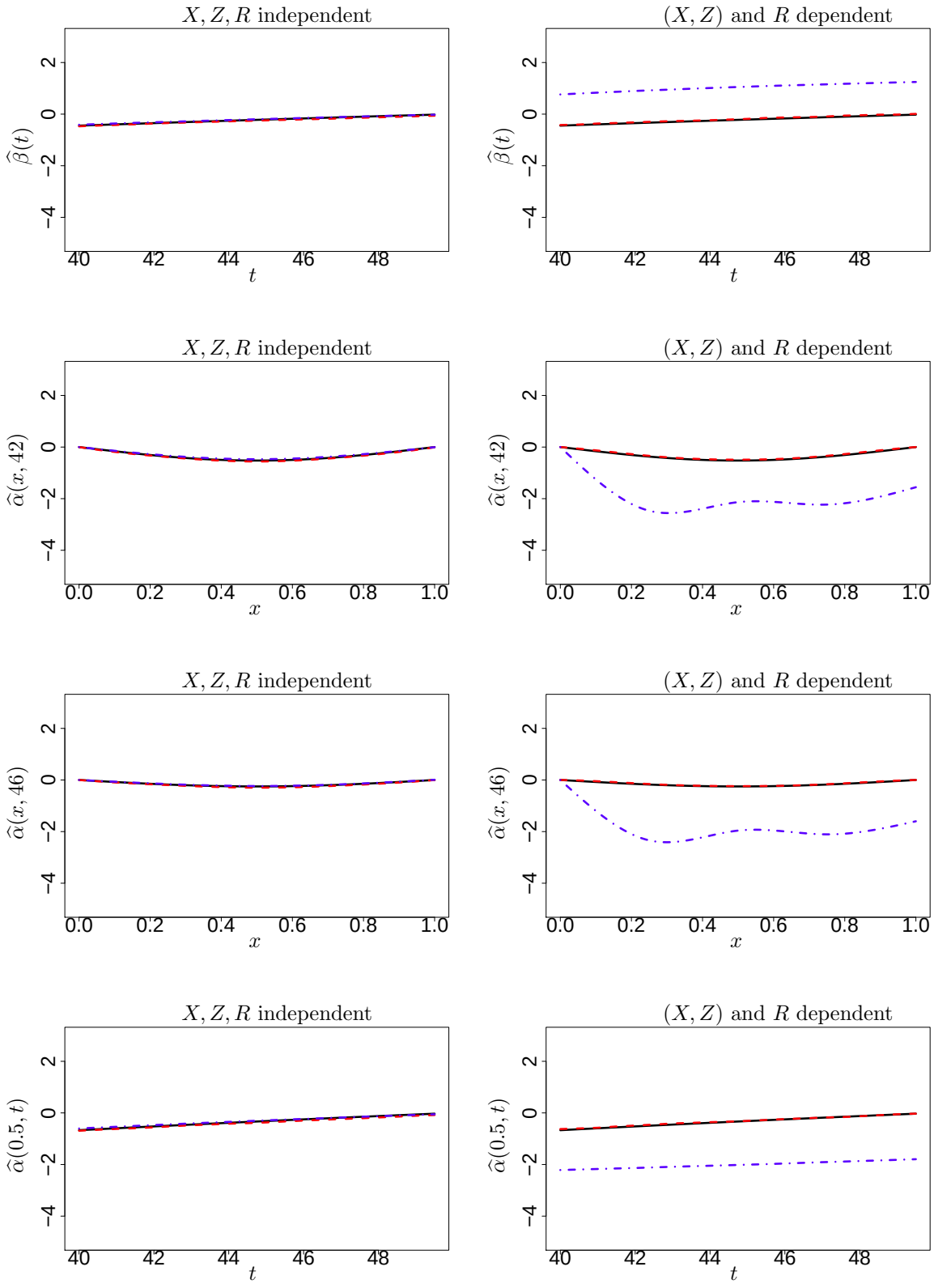

FIG 1. *Clustered failure times with single event types and* $R \sim$ *Normal*$(0, 1)$. *40% censoring, 1000 simulations. True parameter functions (black solid curve), mean of 1000 simulation estimates from our proposed method (red dashed line) and from GAMM (blue dashed-dotted line).*

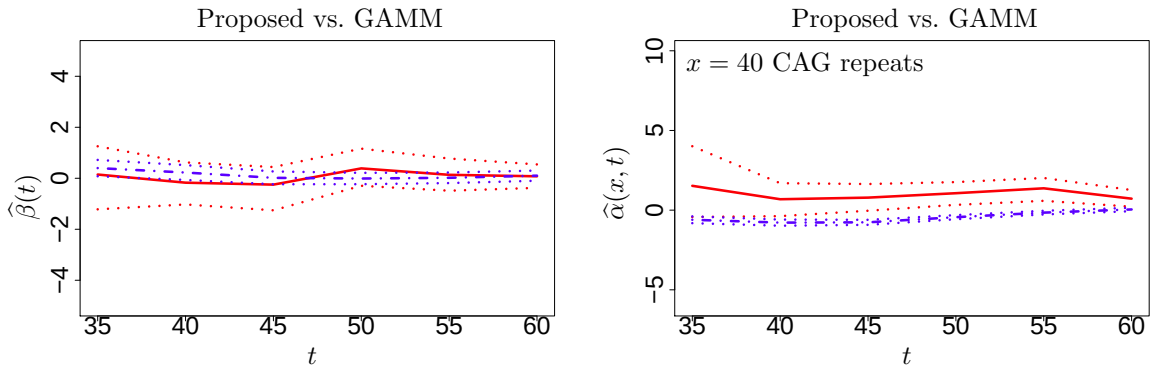

FIG 2. *Comparison of* $\widehat{\alpha}(x,t), \widehat{\beta}(t)$ *from our proposed method (red solid curve) and GAMM (blue dashed-dotted curve) for COHORT study. 95% confidence bands (dotted lines) overlap for* $\widehat{\alpha}(x,t)$ *when* $t \geq 40$ *which, from the Hausman test, indicates that the covariates and random effects are dependent.*

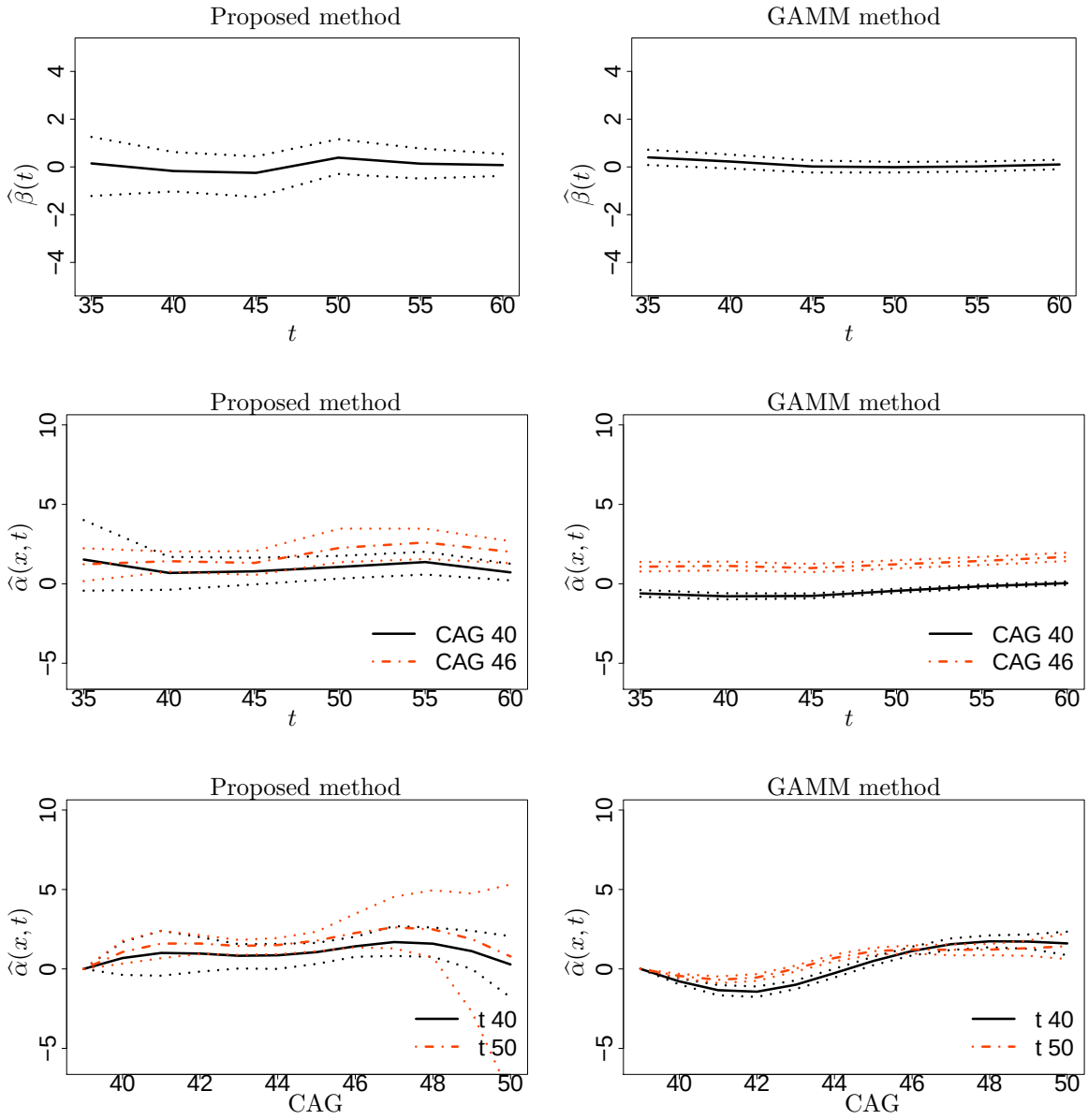

FIG 3. *Estimated difference effects,* $\widehat{\alpha}(x,t), \widehat{\beta}(t)$*, for age of first motor sign vs. age when cognitive impairment first impacts daily life in COHORT study.*

References.

Andersen, P. K. and Perme, M. P. (2010). Pseudo-observations in survival analysis. *Statistical Methods in Medical Research*, **19**, 71-99.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine*, 2, 273-277.

Chen, D. G and Lio, Y.L (2008). Comparative Studies on Frailties in Survival Analysis. *Communications in Statistics - Simulation and Computation*, **37**, 1631-1646.

Chen, M. C. and, Bandeen-Roche, K. (2005). A diagnostic for association in bivariate survival models. *Lifetime Data Analysis*, **11**, 245-264.

Chen, Y., Chen, K. and Ying, Z. (2010). Analysis of multivariate failure time data using marginal proportional hazards model. *Statistica Sinica*, **20**, 1025-1041.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65**, 141-151.

Conne, D., Ronchetti, E. and Victoria-Feser, M. P. (2010). Goodness of fit for generalized linear latent variable models. *Journal of the American Statistical Association*, **105**, 1126-1134.

Congdon, P. (1994). Analyzing mortality in London: life-tables with frailty. *Statistician*, **43**, 277-308.

de Boor, C. (2001). A practical guide to splines. New York: Springer-Verlag.

Duff, K., Paulsen, J., Mills, J., Beglinger, L. J., Moser, D. J., Smith, M. M., Langbehn, D., Stout, J., Queller, S., Harrington, D. L. and the PREDICT-HD Investigators and Coordinators of the Huntington Study Group. (2010). Mild cognitive impairment in prediagnosed Huntington disease. *Neurology*, **75**, 500-507.

Efron, B. (1988). Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve. *Journal of the American Statistical Association*, **83**, 414-425.

Garcia, T. P. and Ma, Y. (2015). Optimal estimator for logistic model with distribution-free random intercept. *Scandinavian Journal of Statistics*, **43**, 156-171.

Geerdens, C., Claeskens, G. and Janssen, P. (2013). Goodness-of-fit tests for the frailty distribution in proportional hazards models with shared frailty. *Biostatistics*, **14**, 433-446.

Glidden, D. V. and Vittinghoff, E. (2004). Modeling clustered survival data from multicenter clinical trials. *Statistics in Medicine*, **23**, 369-388.

Gorfine, M., De-Picciotto, L. and Hsu, L. (2012). Conditional and marginal estimates in case-control family data—extensions and sensitivity analysis. *Journal of Statistical Computation and Simulations*, **29**, 997-1003.

Govindarajulu, U. S., Glickman, M. E. and D'Agostino, R. B. (2007). Modeling frailty as a function of observed covariates. *Journal of Statistical Theory and Practice*, **1**, 117-135.

Harper, P. S. (1996). Huntington's disease. London: W.B. Saunders, 2nd ed.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, **46**, 1251-1271.

Heagerty, P. J. and Kurland, B. F. (2001). Misspecied maximum likelihood estimates and generalized linear mixed models. *Biometrika*, **88**, 973-985.

Henderson, R. and Oman, P. (1999). Effect of frailty on marginal regression estimates in survival analysis. *JRSSB*, **61**, 367-379.

Hsu, L., Gorfine, M. and Malone, K. (2007). On robustness of marginal regression coefficient estimates and hazard functions in multivariate survival analysis of family data when the frailty distribution is mis-specified. *Statistics in Medicine*, **26**, 4657-4678.

Huang, S. S., Yokoe, D. S., Stelling, J., Placzek, H., Kulldorff, M. and Kleinman, K., OBrien, T. F., Calderwood, M. S., Vostok, J., Dunn, J. and Platt, R. (2010) Automated

Detection of Infectious Disease Outbreaks in Hospitals: A Retrospective Cohort Study. *PLoS Med*, **7**, e1000238.

Huber, P., Ronchetti, E. and Victoria-Feser, M. P. (2004). Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society, Series B*, **66**, 893-908.

Huntington's Disease Collaborative Research Group. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, **72**, 971-983.

Johnson, S. G. and Narasimhan, B. (2013). Cubature: Adaptive multivariate integration over hypercubes. R package version 1.1-2. http://CRAN.R-project.org/package=cubature.

Klein, J. P., van Houwelingen, H. C., Ibrahim, J. C. and Scheike, T. H. (eds.) (2014). Handbook of survival analysis. Boca Raton: CRC Press.

Langbehn, D. R., Brinkman, R. R., Falush, D., Paulsen, J. S., Hayden, M. R. (2004). A new model for prediction of the age of onset and penetrance for Huntingtons disease based on CAG length. *Clinical Genetics*, **65**, 267-277.

Lee, K. J. and Thompson, S. G. (2008). Flexible parametric models for random-effects distributions. *Statistics in Medicine*, **27**, 418-434.

Lesaffre, E. and Molenberghs, G. (2001). Multivariate probit analysis: A neglected procedure in medical statistics. *Statistics in Medicine*, **10**, 1391-1403.

Logan, B. R., Nelson, G. O. and Klein, J. P. (2008). Analyzing center specific outcomes in hematopoietic cell transplantation. *Lifetime Data Analysis*, **14**, 389-404.

Logan, B., Zhang, M. and Klein, J. (2011). Marginal models for clustered time to event data with competing risks using pseudo-values. *Biometrics*, **67**, 1-7.

Ma, Y. and Genton, M. G. (2010). Explicit estimating equations for semiparametric generalized linear latent variable models. *Journal of the Royal Statistical Society, Series B*, **72**, 475-495.

Marder, K., Levy, G., Louis, E. D., Mejia-Santana, H., Cote, L., Andrews, H., Harris, J., Waters, C., Ford, B., Frucht, S., Fahn, S. and Ottman, R. (2003). Accuracy of family history data on Parkinson's Disease. *Neurology*, **61**, 18-23.

Marder, K., Zhao, H., Myers, R., Cudkowicz, M., Kayson, E., Kieburtz, K., Orme, C., Paulsen, J., Penney, J., Siemers, E., Shoulson, I., and the Huntington Study Group. (2000). Rate of Functional Decline in Huntington's Disease. *Neurology*, **369**, 452-458.

Murphy, S., Rossini, A., and van der Vaart, A. (1997). Maximum likelihood estimation in the proportional odds model.*Journal of the American Statistical Association,* 92, 968-976.

Paulsen, J. and Long. J. (2014). Onset of Huntington's disease: Can it be purely cognitive? *Movement Disorders*, **29**, 1342-1350.

Piepho, H. P. and McCulloch, C. E. (2004). Transformations in mixed models: Application to risk analysis for a multienvironment trial. *J. Agric. Biol. Environ. Statist.*, **9**, 123-137.

Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, **56**, 1016-1022.

Rizopoulos, D. and Ghosh, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. Statistics in medicine, 30(12), 1366-1380.

Ross, C. A. and Tabrizi, S. J. (2010). Huntington's disease: from molecular pathogenesis to clinical treatment. *The Lancet Neurology*, **10**, 83-98.

Shih, J. H. and Louis,T. A. (1995) Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, **51**, 1384-1399.

Stout, J. C., Paulsen, J. S., Queller, S., Solomon, A. C., Whitlock, K. B., Campbell, J. C., Carlozzi, N., Duff, K., Beglinger, L. J., Langbehn, D. R., Johnson, S. A., Biglan, K.

M., and Aylward, E. H. (2011). Neurocognitive signs in prodromal Huntington disease. *Neuropsychology*, **25**, 1-14.
Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.
Tsiatis, A. A. and Ma, Y. (2004). Locally efficient semiparametric estimators for functional measurement error models. *Biometrika*, **91**, 835-848.
Wood, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society, Series B*, **70**, 495-518.
Ying, Z. and Wei, L. J. (1994). The Kaplan-Meier estimate for dependent failure time observations. *Journal of Multivariate Analysis*, **50**, 17-29.
Zeng, D., Lin, D. and Yin, G. (2005). Maximum likelihood estimation for the proportional odds model with random effects. *Journal of the American Statistical Association*, **100**, 470-483.

APPENDIX A: ASYMPTOTIC PROPERTIES

We describe the asymptotic properties for $\widehat{\alpha}(x,t)$ and $\widehat{\boldsymbol{\beta}}(t)$ at any t; for notational simplicity, we drop the variable t and use $\widehat{\alpha}(x), \widehat{\boldsymbol{\beta}}$. Recall that we approximate $\alpha(x)$ using the B-spline approximation in (2.2) with order r and N internal knots, denoted as $\xi_{r+1}, \ldots, \xi_{N+r}$. We have also assumed that the distance between neighboring knots is $h_k = \xi_{k+1} - \xi_k$ for $r \le k \le N+r$ and $h = \max_{r \le k \le N+r} h_k$.

To derive the asymptotic properties, we make the following regularity conditions.

(C1) The density function $f_X(x)$ of random variable X has a compact support, is bounded away from 0 and satisfies the Lipschitz condition of order 1 on its support. Denote the support $[a,b]$, which corresponds to the knot endpoints in (2.1); i.e., $a = \xi_1$, $b = \xi_{N+2r}$.

(C2) The true $\alpha(x)$ function is $\alpha_0(x) \in C^q[a,b]$ for $q \ge 2$ and the spline order r satisfies $r \ge q$. Here, $C^q[a,b]$ denotes functions on $[a,b]$ that have qth continuous derivative.

(C3) There exists $0 < c_h < \infty$, such that

$$\max_{r \le k \le N+r} |h_{k+1} - h_k| = o(N^{-1}) \text{ and } h / \min_{r \le k \le N+r} h_k < c_h.$$

Furthermore, the number of internal knots satisfies $N \to \infty$, $N^{-4}n \to \infty$ and $Nn^{-1/(2q)} \to \infty$ as $n \to \infty$.

(C4) The expectation $E\{\boldsymbol{S}^{\mathrm{T}}_{\mathrm{eff},\boldsymbol{\beta}}(\boldsymbol{Y}, X, \boldsymbol{Z}, \boldsymbol{\beta}, \alpha), \boldsymbol{S}^{\mathrm{T}}_{\mathrm{eff},\boldsymbol{a}}(\boldsymbol{Y}, X, \boldsymbol{Z}, \boldsymbol{\beta}, \alpha)\}^{\mathrm{T}} = \mathbf{0}$ has a unique zero in the neighborhood of the true parameter value. The derivative of $\boldsymbol{S}_{\mathrm{eff}}(\boldsymbol{Y}, X, \boldsymbol{Z}, \boldsymbol{\theta}) = \{\boldsymbol{S}^{\mathrm{T}}_{\mathrm{eff},\boldsymbol{\beta}}(\boldsymbol{Y}, X, \boldsymbol{Z}, \boldsymbol{\theta}), \boldsymbol{S}^{\mathrm{T}}_{\mathrm{eff},\boldsymbol{a}}(\boldsymbol{Y}, X, \boldsymbol{Z}, \boldsymbol{\theta})\}^{\mathrm{T}}$ with respect to $\boldsymbol{\theta}$ has bounded and nonsingular expectation.

Proofs of the asymptotic properties make use of the estimating equations in Proposition 3 and derivatives of these terms. For ease in notation, we define $\boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}, \boldsymbol{a}) = n^{-1} \sum_{i=1}^{n} \boldsymbol{S}_{\mathrm{eff},\boldsymbol{a}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})$, $\boldsymbol{g}_{n,\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{a}) =$

$n^{-1}\sum_{i=1}^{n} \boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})$, which correspond to the estimating equations for $\boldsymbol{a}$ and $\boldsymbol{\beta}$, respectively. Derivatives of these terms will be denoted by $\boldsymbol{V}_{n,\boldsymbol{aa}}(\boldsymbol{\beta}, \boldsymbol{a}) = -\partial \boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}, \boldsymbol{a})/\partial \boldsymbol{a}^{\mathrm{T}}$, $\boldsymbol{V}_{n,\boldsymbol{a\beta}}(\boldsymbol{\beta}, \boldsymbol{a}) = -\partial \boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}, \boldsymbol{a})/\partial \boldsymbol{\beta}^{\mathrm{T}}$, $\boldsymbol{V}_{n,\boldsymbol{\beta a}}(\boldsymbol{\beta}, \boldsymbol{a}) = -\partial \boldsymbol{g}_{n,\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{a})/\partial \boldsymbol{a}^{\mathrm{T}}$ and $\boldsymbol{V}_{n,\boldsymbol{\beta\beta}}(\boldsymbol{\beta}, \boldsymbol{a}) = -\partial \boldsymbol{g}_{n,\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{a})/\partial \boldsymbol{\beta}^{\mathrm{T}}$. Analogous to these terms, let $\boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}, \alpha)$, $\boldsymbol{g}_{n,\boldsymbol{\beta}}(\boldsymbol{\beta}, \alpha)$, $\boldsymbol{V}_{n,\boldsymbol{aa}}(\boldsymbol{\beta}, \alpha)$, $\boldsymbol{V}_{n,\boldsymbol{a\beta}}(\boldsymbol{\beta}, \alpha)$, $\boldsymbol{V}_{n,\boldsymbol{\beta a}}(\boldsymbol{\beta}, \alpha)$ and $\boldsymbol{V}_{n,\boldsymbol{\beta\beta}}(\boldsymbol{\beta}, \alpha)$ denote the corresponding quantities when $\boldsymbol{B}^{\mathrm{T}}(x)\boldsymbol{a}$ is replaced by $\alpha(x)$ at all x values. Also, define $\boldsymbol{V}_{\boldsymbol{aa}}(\boldsymbol{\beta}, \alpha)$, $\boldsymbol{V}_{\boldsymbol{a\beta}}(\boldsymbol{\beta}, \alpha)$, $\boldsymbol{V}_{\boldsymbol{\beta a}}(\boldsymbol{\beta}, \alpha)$ and $\boldsymbol{V}_{\boldsymbol{\beta\beta}}(\boldsymbol{\beta}, \alpha)$ as replacing the averages across n respectively in $\boldsymbol{V}_{n,\boldsymbol{aa}}(\boldsymbol{\beta}, \alpha)$, $\boldsymbol{V}_{n,\boldsymbol{a\beta}}(\boldsymbol{\beta}, \alpha)$, $\boldsymbol{V}_{n,\boldsymbol{\beta a}}(\boldsymbol{\beta}, \alpha)$ and $\boldsymbol{V}_{n,\boldsymbol{\beta\beta}}(\boldsymbol{\beta}, \alpha)$ by expectations. Let $\boldsymbol{J}_i = (-\mathbf{1}_{m_i-1}, \boldsymbol{I}_{m_i-1})^{\mathrm{T}}$ for $i = 1, \ldots, n$.

We first investigate the asymptotic properties in estimating $\alpha(x)$ and $\boldsymbol{\beta}$ in the situation when no censoring has occurred and the event numbers are identical for each subject (Theorems 1 and 2). We then proceed to study the large sample properties under the assumption of censoring and allowing different event numbers (Theorem 3). Proofs of Theorems 1 and 2 are in Sections S.1.7 and S.1.8 (Supplementary Material).

Theorem 1. *Assume $m_i = m < \infty$ for $i = 1, \ldots, n$, where m_i is the number of events for the ith individual as previously defined. Assume censoring does not occur. Under regularity conditions (C1)-(C4), the estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\alpha}(x)$ satisfy $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \to \mathbf{0}$ and $\|\boldsymbol{B}^{\mathrm{T}}(x)\widehat{\boldsymbol{a}} - \alpha_0(x)\|_\infty \to 0$ as $n \to \infty$. Let $\boldsymbol{\beta}$ be either the true parameter $\boldsymbol{\beta}_0$ or a root-n consistent estimator of $\boldsymbol{\beta}_0$. Then $|\widehat{\alpha}(x, \boldsymbol{\beta}) - \alpha(x)| = O_p\{(nh)^{-1/2} + h^q\}$ uniformly in $x \in [a, b]$ and as $n \to \infty$, $\widehat{\alpha}(x, \boldsymbol{\beta}) - \alpha(x)$ converges to a mean zero normal distribution.*

The result in Theorem 1 establishes the asymptotic consistency, normality and variability of $\widehat{\alpha}(x)$. The proof relies on three Lemmas (see Section S.1.7) that calculate various quantities explicitly and establish necessary bounds on the B-spline approximations. These properties combined with results from de Boor (2001), Taylor expansion and the Central Limit Theorem then yield the asymptotic normality and assess the variability of $\widehat{\alpha}(x)$. Given these consistency results in Theorem 1, we can then apply a Taylor expansion to yield the following asymptotic normality result for $\widehat{\boldsymbol{\beta}}$.

Theorem 2. *Assume $m_i = m < \infty$ for $i = 1, \ldots, n$, where recall that m_i is the number of events for the ith individual. Assume censoring does not occur. Under regularity conditions (C1)-(C4), $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$. When $n \to \infty$, $n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to$ Normal$\{\mathbf{0}_{p_1}, \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}_0, \alpha_0)\}$ in distribution, where*

$$\boldsymbol{\Sigma}_0(\boldsymbol{\beta}_0, \alpha_0) = E\left([\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)m - \Pi\{\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0) \mid \mathcal{S}_\alpha\}]^{\otimes 2}\right).$$

Here $\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)$ *is as defined in Proposition 3 except that* $\boldsymbol{B}^{\mathrm{T}}(x)\boldsymbol{a}$ *is replaced by* $\alpha_0(x)$ *for all* x*. Additionally,* $\mathcal{S}_\alpha$ *is the functional space defined as*

$$\mathcal{S}_\alpha = [\{\boldsymbol{f}(X_{i1}), \ldots, \boldsymbol{f}(X_{i_i})\}\boldsymbol{J}_i diag\{V_{ij} - E(V_{ij} \mid W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0), j = 2, \ldots, m_i\}\mathbf{1}_{m_i - 1}],$$

where $\boldsymbol{f}(x)$ *is any arbitrary* p_1*-component function with each component in* $C^q[a, b]$*, and* $\Pi\left[\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)\right\} \mid \mathcal{S}_\alpha]$ *denotes the orthogonal projection of* $\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)$ *onto* $\mathcal{S}_\alpha$*. In addition,* $\boldsymbol{\Sigma}_0(\boldsymbol{\beta}_0, \alpha_0)$ *can be estimated using*

$$\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}, \widehat{\alpha}) = \boldsymbol{V}_{n,\boldsymbol{\beta}\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}, \widehat{\alpha}) - \boldsymbol{V}_{n,\boldsymbol{\beta}\boldsymbol{a}}(\widehat{\boldsymbol{\beta}}, \widehat{\alpha})\boldsymbol{V}^{-1}_{n,\boldsymbol{a}\boldsymbol{a}}(\widehat{\boldsymbol{\beta}}, \widehat{\alpha})^{\mathrm{T}}\boldsymbol{V}_{n,\boldsymbol{a}\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}, \widehat{\alpha}).$$

Corollary 1. *Under the same conditions as those required in Theorem 2, the estimator* $\widehat{\boldsymbol{\beta}}$ *obtained from solving the estimating equations in Proposition 3 reaches the optimal semiparametric efficiency bound, given as* $\boldsymbol{\Sigma}_0(\boldsymbol{\beta}_0, \alpha_0)$*.*

The efficiency result stated in Corollary 1 is for $\widehat{\boldsymbol{\beta}}$ at each t_0 and without censoring. The result is immediate following the proof of Theorem 2, and by noting that $\mathcal{S}_\alpha$ is the residual of the tangent space with respect to α after projecting it to the tangent space with respect to $f_{\boldsymbol{X},\boldsymbol{Z},R}$. In establishing the results in Theorems 1, 2 and Corollary 1, we have assumed all the subjects experience the same number of events and all the events are observed. When the number of events m_i varies and when some of the events are censored, similar results hold, as we state in Theorem 3. The derivation of the results in Theorem 3 is almost identical to those in the proofs of Theorems 1 and 2, except that we are obliged to retain summation across all the n individuals instead of using a single expectation, hence we omit the details of the proof. We emphasize that the $\boldsymbol{\Sigma}_0$ here is different from that in Theorem 2 not only in the additional $n^{-1}\sum_{i=1}^n$, but also in that the calculation of expectations $E(V_{ij} \mid W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$, which is different under censoring and not censoring.

Theorem 3. *Under the regularity conditions (C1)-(C4), the estimators* $\widehat{\boldsymbol{\beta}}$ *and* $\widehat{\alpha}(x)$ *satisfy* $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \to \mathbf{0}$ *and* $\|\boldsymbol{B}^{\mathrm{T}}(x)\widehat{\boldsymbol{a}} - \alpha_0(x)\|_\infty \to 0$ *as* $n \to \infty$*. For* $\boldsymbol{\beta}$ *that is either the true parameter* $\boldsymbol{\beta}_0$ *or a root-*n *consistent estimator of* $\boldsymbol{\beta}_0$*,* $|\widehat{\alpha}(x, \boldsymbol{\beta}) - \alpha(x)| = O_p\{(nh)^{-1/2} + h^q\}$ *uniformly in* $x \in [a, b]$ *and as* $n \to \infty$*,* $\widehat{\alpha}(x, \boldsymbol{\beta}) - \alpha(x)$ *converges to a mean zero normal distribution. Further,* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$*. When* $n \to \infty$*,* $n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to$ *Normal*$\{\mathbf{0}_{p_1}, \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}_0, \alpha_0)\}$ *in distribution, where*

$$\boldsymbol{\Sigma}_0(\boldsymbol{\beta}_0, \alpha_0) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n E\left([\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0) - \Pi\left\{\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0) \mid \mathcal{S}_{\alpha_i}\right\}]^{\otimes 2}\right).$$

Here $\mathcal{S}_{\alpha_i}$ *is the functional space defined as*

$$\mathcal{S}_{\alpha_i} = [\{\boldsymbol{f}(X_{i1}), \ldots, \boldsymbol{f}(X_{im_i})\}\boldsymbol{J}_i diag\{V_{ij} - E(V_{ij} \mid W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0), j = 2, \ldots, m_i\}\mathbf{1}_{m_i-1}],$$

where $\boldsymbol{f}(x)$ *is any arbitrary* p_1*-component function with each component in* $C^q[a, b]$*, and* $\Pi[\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)\} \mid \mathcal{S}_{\alpha_i}]$ *denotes the orthogonal projection of* $\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)$ *onto* $\mathcal{S}_{\alpha_i}$*. In addition,* $\boldsymbol{\Sigma}_0(\boldsymbol{\beta}_0, \alpha_0)$ *can be estimated using*

$$\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}, \widehat{\alpha}) = \boldsymbol{V}_{n,\boldsymbol{\beta}\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}, \widehat{\alpha}) - \boldsymbol{V}_{n,\boldsymbol{\beta}\boldsymbol{a}}(\widehat{\boldsymbol{\beta}}, \widehat{\alpha})\boldsymbol{V}_{n,\boldsymbol{a}\boldsymbol{a}}^{-1}(\widehat{\boldsymbol{\beta}}, \widehat{\alpha})^{\mathrm{T}}\boldsymbol{V}_{n,\boldsymbol{a}\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}, \widehat{\alpha}).$$

Supplementary Material for *Robust mixed-effects model for clustered failure time data: application to Huntington's disease event measures*

Tanya P. Garcia, Yanyuan Ma, Karen Marder, and Yuanjia Wang

APPENDIX S.1: TECHNICAL PROOFS

S.1.1. Justification of properties (P1) and (P2). We highlight the key results in Logan et al. (2011) that lead to properties (P1) and (P2). First, $\widehat{F}(t_0)$ can be written using an inverse probability of censoring weighting (IPCW) formulation (Scheike et al., 2008) as $\widehat{F}(t_0) = M^{-1}\sum_{i,j} N_{ij}(t_0)/\widehat{G}(L_{ij})$ where $N_{ij}(t_0) = I(L_{ij} \leq t_0)\delta_{ij}$ and $\widehat{G}(L_{ij})$ is the Kaplan-Meier estimator for the censoring survival distribution.

Second, the IPCW formulation and martingale properties leads to the pseudo-values satisfying

$$Y^*_{ij}(t_0) = \frac{N_{ij}(t_0)}{G(L_{ij})} + \int_0^{L_{ij}} \frac{\mathrm{pr}(T \leq t_0, \delta = 1 | T \geq u)}{G(u)} d\mathcal{M}^c_{ij}(u) + O_p(M^{-1/2}) \tag{S.1}$$

where $\mathcal{M}^c_{ij}(t) = I(L_{ij} \leq t)(1-\delta_{ij}) - \int_0^t I(L_{ij} \geq u)\lambda^c(u)du$ is the martingale for the censoring process for observation (i,j) with censoring hazard function $\lambda^c(u) = -d\log\{G(u)\}/du$. See the Appendix of Logan et al. (2011) for exact details.

Property (P1) then follows since $Y^*_{ij}(t_0)$ only depends on L_{ij} and $d\mathcal{M}^c_{ij}$ as M approaches infinity. Property (P2) follows since the second term in (S.1) is a martingale with mean zero, is independent of covariates and because $E\{N_{ij}(t_0)/G(L_{ij})\} = \mathrm{pr}\{T \leq t_0 | X_{ij}, Z_{ij}, R_i(t_0)\}$ (Scheike et al., 2008).

S.1.2. Proof of properties (P1†) and (P2†). Let $N_{ij}(t_0) = I(L_{ij} \leq t_0)\delta_{ij}$, $N_j(t_0) = \sum_i N_{ij}(t_0)$ denote the number of type j events observed up to time t_0, and $\mathcal{Q}_j(t_0) = \sum_i I(L_{ij} \geq t_0)$ denote the risk set for event j at t_0. Write $\widehat{F}_j(t_0)$ using an inverse probability of censoring weighting (IPCW) formulation (Scheike et al., 2008) as $\widehat{F}_j(t_0) = n^{-1}\sum_i N_{ij}(t_0)/\widehat{G}_j(L_{ij})$ where $\widehat{G}_j(t_0)$ is the Kaplan-Meier estimate of the censoring survival distribution for event j. Let $\widehat{G}^{(i)}_j(t_0)$ denote the Kaplan-Meier estimate for the censoring

survival distribution for event j after removing cluster i. Based on the IPCW representation, the pseudo-value $Y_{ij}^{\dagger}(t_0)$ becomes

$$
\begin{aligned}
Y_{ij}^{\dagger}(t_0) \quad &= \sum_i \frac{N_{ij}(t_0)}{\widehat{G}_j(L_{ij})} - \sum_{k\neq i} \frac{N_{kj}(t_0)}{\widehat{G}_j^{(i)}(L_{kj})} \\
&= \sum_i \frac{N_{ij}(t_0)}{G_j(L_{ij})} - \sum_{k\neq i} \frac{N_{kj}(t_0)}{G_j(L_{kj})} \qquad \text{(S.2)} \\
&+ \sum_{k\neq i} N_{kj}(t_0) \left[\left\{ \frac{1}{\widehat{G}_j(L_{kj})} - \frac{1}{G_j(L_{kj})} \right\} - \left\{ \frac{1}{\widehat{G}_j^{(i)}(L_{kj})} - \frac{1}{G_j(L_{kj})} \right\} \right] \qquad \text{(S.3)} \\
&+ N_{ij}(t_0) \left\{ \frac{1}{\widehat{G}_j(L_{ij})} - \frac{1}{G_j(L_{ij})} \right\}. \qquad \text{(S.4)}
\end{aligned}
$$

We now proceed to simplify the terms above. First, the term in (S.2) is $N_{ij}(t_0)/G_j(L_{ij})$. Second, to simplify the term in (S.3), let $\mathcal{M}_{ij}^c(t_0) = I(L_{ij} \leq t_0)(1-\delta_{ij}) - \int_0^t I(L_{ij} \geq u)\lambda_j^c(u)du$ be the martingale corresponding to the censoring process for observation (i,j) with censoring hazard function $\lambda_j^c(u) = -d\log\{G_j(u)\}/du$. Also, let $\mathcal{M}_j^c(u) = \sum_i \mathcal{M}_{ij}^c(u)$ and $\mathcal{M}_j^{c(i)}(u) = \sum_{k\neq i} \mathcal{M}_{kj}^c(u)$. We will also make use the following useful fact (Bang and Tsiatis, 2000; Robins and Rotnitzky, 1992):

$$
\frac{\widehat{G}_j(t) - G_j(t)}{G_j(t)} = -\int_0^t \frac{\widehat{G}_j(u-)}{G_j(u)} \frac{d\mathcal{M}_j^c(u)}{\mathcal{Q}_j(u)}.
$$

Based on the introduced notation, this useful fact, and that $\mathcal{M}_j^c(u) = \mathcal{M}_j^{c(i)}(u) + \mathcal{M}_{ij}^c(u)$, we can re-write the term in (S.3) as

$$
\begin{aligned}
&\sum_{k\neq i} \frac{N_{kj}(t_0)}{\widehat{G}_j(L_{kj})\widehat{G}_j^{(i)}(L_{kj})} \qquad \text{(S.5)} \\
&\times \int_0^{L_{kj}} \left\{ \frac{\widehat{G}_j^{(i)}(L_{kj})\widehat{G}_j(u-)}{G_j(u)\mathcal{Q}_j(u)} - \frac{\widehat{G}_j(L_{kj})\widehat{G}_j^{(i)}(u-)}{G_j(u)\mathcal{Q}_j^{(i)}(u)} \right\} d\mathcal{M}_j^{c(i)}(u) \\
&+ \sum_{k\neq i} \frac{N_{kj}(t_0)}{\widehat{G}_j(L_{kj})} \int_0^{L_{kj}} \frac{\widehat{G}_j(u-)}{G_j(u)} \frac{d\mathcal{M}_{ij}^c(u)}{\mathcal{Q}_j(u)}. \qquad \text{(S.6)}
\end{aligned}
$$

Let $r_j(u)$ be such that $\mathcal{Q}_j(u)/n$ converges in probability to $r_j(u)$. It thus follows that

$$
\int_0^{L_{kj}} \left\{ \frac{(n-1)\widehat{G}_j^{(i)}(L_{kj})\widehat{G}_j(u-)}{G_j(u)\mathcal{Q}_j(u)/n} - \frac{n\widehat{G}_j(L_{kj})\widehat{G}_j^{(i)}(u-)}{G_j(u)\mathcal{Q}_j^{(i)}(u)/(n-1)} \right\} \frac{d\mathcal{M}_j^{c(i)}(u)}{\sqrt{n}}
$$

converges in distribution to $\mathcal{W}_{kj}$, say. Therefore the first term in (S.5) is asymptotically equivalent to

$$\frac{1}{\sqrt{n}}\sum_{k\neq i}\frac{N_{kj}(t_0)\mathcal{W}_{kj}}{(n-1)\widehat{G}_j(L_{kj})\widehat{G}_j^{(i)}(L_{kj})} = O_p(n^{-1/2}).$$

The term in (S.6) is asymptotically equivalent to

$$\sum_{k\neq i}\frac{N_{kj}(t_0)}{G_j(L_{kj})}\int_0^{L_{kj}}\frac{d\mathcal{M}_{ij}^c(u)}{\mathcal{Q}_j(u)}$$

which is equivalent to

$$\int_0^{L_{ij}}\frac{d\mathcal{M}_{ij}^c(u)}{\mathcal{Q}_j(u)/n}\sum_{k\neq i}\frac{dN_{kj}(t_0)I(L_{kj}\geq u)}{nG_j(L_{kj})} \tag{S.7}$$

because $\int_0^{L_{kj}} d\mathcal{M}_{ij}^c(u)/\mathcal{Q}_j(u) = \int_0^{L_{ij}} I(u\leq L_{kj})/\mathcal{Q}_j(u)d\mathcal{M}_{ij}^c(u)$. Following the IPCW representation, the term in (S.7) is asymptotically equivalent to

$$\int_0^{L_{ij}}\frac{\text{pr}(T_j\leq t_0,\delta=1|T_j\geq u)}{G_j(u)}d\mathcal{M}_{ij}^c(u).$$

Combining the above results, we thus have that

$$Y_{ij}^\dagger(t_0) = \frac{N_{ij}(t_0)}{G_j(L_{ij})} + \int_0^{L_{ij}}\frac{\text{pr}(T_j\leq t_0,\delta=1|T_j\geq u)}{G_j(u)}d\mathcal{M}_{ij}^c(u) + O_p(n^{1/2}).$$

It therefore follows that (P1†) holds because $Y_{ij}^\dagger(t_0)$ only depends on L_{ij} and $d\mathcal{M}_{ij}^c$ as n approaches infinity. Also, (P2†) holds because the second term above is a martingale with mean zero, is independent of covariates and because $E\{N_{ij}(t_0)/G_j(L_{ij})\} = \text{pr}\{T_j\leq t_0|X_{ij},\boldsymbol{Z}_{ij},R_i(t_0)\}$ (Scheike et al., 2008).

S.1.3. Proof of Proposition 1. Consider the Hilbert space $\mathcal{H} = \{\boldsymbol{h}(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z}) : E(\boldsymbol{h}) = \boldsymbol{0}, \text{var}(\boldsymbol{h}) < \infty\}$ consisting of mean zero, finite variance q-dimensional functions $\boldsymbol{h}(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z})$. Two important subspaces of $\mathcal{H}$ are the nuisance tangent space and its orthogonal complement. The nuisance tangent space is the mean-squared closure of the space of elements $\boldsymbol{BS}$ where $\boldsymbol{S}$ is an arbitrary nuisance score vector and $\boldsymbol{B}$ is a conformable matrix. Viewing the unknown density $f_{\boldsymbol{X},\boldsymbol{Z},R}$ as a nuisance parameter, the nuisance tangent space is

$$\Lambda \;=\; [E\{\boldsymbol{h}(\boldsymbol{X},\boldsymbol{Z},R)|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z}\} : E(\boldsymbol{h}) = \boldsymbol{0}, E(\boldsymbol{h}^{\text{T}}\boldsymbol{h}) < \infty]$$

where $\boldsymbol{h}$ is a function of dimension the same length as $\boldsymbol{\theta}$. The space orthogonal to the nuisance tangent space is

$$\Lambda^{\perp} = [\boldsymbol{g}(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z}) : E\{\boldsymbol{g}(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z})|\boldsymbol{X},\boldsymbol{Z},R\} = \boldsymbol{0}, E(\boldsymbol{g}^{\mathrm{T}}\boldsymbol{g})\} < \infty],$$

where $\boldsymbol{g}$ is a q-dimensional function.

By results in semiparametric theory (Tsiatis, 2006), constructing unbiased estimating equations for $\boldsymbol{\theta}$ is based on the result of projecting the score function with respect to $\boldsymbol{\theta}$ onto $\Lambda^{\perp}$. The result of this projection is the so-called efficient score vector which is

$$\begin{aligned}\boldsymbol{S}_{\text{eff}}(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta}) &= \Pi\{\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta})|\Lambda^{\perp}\} \\ &= \boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta}) - \Pi\{\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta})|\Lambda\},\end{aligned}$$

where $\Pi(\cdot|\cdot)$ denotes projection. Hence, based on our form of Λ and $\Lambda^{\perp}$, we have

$$\boldsymbol{S}_{\text{eff}}(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta}) = \boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z}) - E\{\boldsymbol{h}(\boldsymbol{X},\boldsymbol{Z},R)|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z}\},$$

where $\boldsymbol{h}$ satisfies the condition in (2.4). It is important to note that solving for $\widehat{\boldsymbol{\theta}}$ from $\sum_{i=1}^{n}\boldsymbol{S}_{\text{eff}}\{\boldsymbol{Y}_i,\boldsymbol{X}_i,\boldsymbol{Z}_i;\boldsymbol{\theta}\} = \boldsymbol{0}$ indeed yields a consistent estimator since $E\{\boldsymbol{S}_{\text{eff}}(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta})\} = E[E\{\boldsymbol{S}_{\text{eff}}(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta})|\boldsymbol{X},\boldsymbol{Z},R\}] = \boldsymbol{0}$ because the terms in $\boldsymbol{S}_{\text{eff}}$ satisfy (2.4). This result holds even when $\boldsymbol{Y}$ consists of pseudo-values.

S.1.4. Proof of Proposition 2. Based on the form of $f_{\boldsymbol{Y}|X,\boldsymbol{Z},R}$ in Proposition 1 and using a change of variables, we have that

$$\begin{aligned}& f_{W,V|X,\boldsymbol{Z},R}(w_i,\boldsymbol{v}_i|\boldsymbol{x}_i,\boldsymbol{z}_i,r_i) \\ = \;& f_{\boldsymbol{Y}|X,\boldsymbol{Z},R}\left\{\boldsymbol{A}_i^{-1}(w_i,\boldsymbol{v}_i)^{\mathrm{T}}\middle|\boldsymbol{x}_i,\boldsymbol{z}_i,r_i\right\}\det(\boldsymbol{A}_i^{-1}) \\ = \;& \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta})\boldsymbol{A}_i^{-1}(w_i,\boldsymbol{v}_i)^{\mathrm{T}}\right\} \\ & \times\exp\left(r_i w_i - \sum_{j=1}^{m_i}\log[1+\exp\{\eta(x_{ij},\boldsymbol{z}_{ij};\boldsymbol{\theta}) + r_i\}]\right).\end{aligned}$$

Under this form of the joint conditional density $f_{W,\boldsymbol{V}|\boldsymbol{X},\boldsymbol{Z},R}$, we can now show the first property that R and $\boldsymbol{V}$ are conditionally independent given

$(W, \boldsymbol{X}, \boldsymbol{Z})$. That is,

$$f_{\boldsymbol{V}|W,\boldsymbol{X},\boldsymbol{Z},R} = f_{\boldsymbol{V}|W,\boldsymbol{X},\boldsymbol{Z}}, \quad f_{R|W,\boldsymbol{X},\boldsymbol{Z},\boldsymbol{V}} = f_{R|W,\boldsymbol{X},\boldsymbol{Z}}$$

This property holds since

$$\begin{aligned} f_{\boldsymbol{V}|W,\boldsymbol{X},\boldsymbol{Z},R} &= \frac{f_{W,\boldsymbol{V}|\boldsymbol{X},\boldsymbol{Z},R}(w_i, \boldsymbol{v}_i|\boldsymbol{x}_i, \boldsymbol{z}_i, r_i)}{\int f_{W,\boldsymbol{V}|\boldsymbol{X},\boldsymbol{Z},R}(w_i, \boldsymbol{v}_i|\boldsymbol{x}_i, \boldsymbol{z}_i, r_i) d\mu(\boldsymbol{v}_i)} \\ &= \frac{\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta}) \boldsymbol{A}_i^{-1}(w_i, \boldsymbol{v}_i^{\mathrm{T}})^{\mathrm{T}}\right\}}{\int \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta}) \boldsymbol{A}_i^{-1}(w_i, \boldsymbol{v}_i^{\mathrm{T}})^{\mathrm{T}}\right\} d\mu(\boldsymbol{v}_i)}. \end{aligned} \tag{S.8}$$

The last equality is completely independent of R suggesting that given $(W, \boldsymbol{X}, \boldsymbol{Z})$, $\boldsymbol{V}$ and R are independent. Therefore, $f_{\boldsymbol{V}|W,\boldsymbol{X},\boldsymbol{Z},R} = f_{\boldsymbol{V}|W,\boldsymbol{X},\boldsymbol{Z}}$ and similarly, $f_{R|W,\boldsymbol{X},\boldsymbol{Z},\boldsymbol{V}} = f_{R|W,\boldsymbol{X},\boldsymbol{Z}}$.

We can also show the second property in that

$$E\{\boldsymbol{g}(W, \boldsymbol{X}, \boldsymbol{Z})|\boldsymbol{x}, \boldsymbol{z}, r\} = \boldsymbol{0} \implies \boldsymbol{g}(W, \boldsymbol{X}, \boldsymbol{Z}) = \boldsymbol{0}.$$

Let $k(w_i, \boldsymbol{x}_i, \boldsymbol{z}_i) = \int \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta}) \boldsymbol{A}_i^{-1}(w_i, \boldsymbol{v}_i^{\mathrm{T}})^{\mathrm{T}}\right\} d\mu(\boldsymbol{v}_i)$ which is positive for all $(w_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$. Then,

$$\begin{aligned} \boldsymbol{0} &= E\{\boldsymbol{g}(W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)|\boldsymbol{x}_i, \boldsymbol{z}_i, r_i\} \\ &= \int \boldsymbol{g}(w_i, \boldsymbol{x}_i, \boldsymbol{z}_i) f_{W_i,\boldsymbol{V}_i|\boldsymbol{X},\boldsymbol{Z},R}(w_i, \boldsymbol{v}_i|\boldsymbol{x}_i, \boldsymbol{z}_i, r_i) d\mu(w_i) d\mu(\boldsymbol{v}_i) \\ &= \int k(w_i, \boldsymbol{x}_i, \boldsymbol{z}_i) \boldsymbol{g}(w_i, \boldsymbol{x}_i, \boldsymbol{z}_i) \exp(r_i w_i) d\mu(w_i) \\ &\quad \times \exp\left(-\sum_{j=1}^{m_i} \log[1 + \exp\{\eta(x_{ij}, \boldsymbol{z}_{ij}; \boldsymbol{\theta}) + r_i\}]\right). \end{aligned}$$

This implies $\boldsymbol{0} = \int k(w_i, \boldsymbol{x}_i, \boldsymbol{z}_i) \boldsymbol{g}(w_i, \boldsymbol{x}_i, \boldsymbol{z}_i) \exp(r_i w_i) d\mu(w_i)$, which means $k(w_i, \boldsymbol{x}_i, \boldsymbol{z}_i) \boldsymbol{g}(w_i, \boldsymbol{x}_i, \boldsymbol{z}_i) = \boldsymbol{0}$. But because $k(w_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ is positive for all $(w_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$, we have that $\boldsymbol{g}(w_i, \boldsymbol{x}_i, \boldsymbol{z}_i) = \boldsymbol{0}$. Therefore, the second property holds.

Now based on these two properties, we can simplify the requirement in (2.4) for the estimating equation. First the requirement can be rewritten as

$$E[E\{\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})|W, \boldsymbol{X}, \boldsymbol{Z}, R\}|\boldsymbol{X}, \boldsymbol{Z}, R\}] = E[E\{\boldsymbol{h}(\boldsymbol{X}, \boldsymbol{Z}, R)|W, \boldsymbol{V}, \boldsymbol{X}, \boldsymbol{Z})\}|\boldsymbol{X}, \boldsymbol{Z}, R].$$

We now simplify the above display. Because R and $\boldsymbol{V}$ are independent given $(W, \boldsymbol{X}, \boldsymbol{Z})$, we have that

$$E\{\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})|W, \boldsymbol{X}, \boldsymbol{Z}, R\} = E\left[\boldsymbol{S}_{\boldsymbol{\theta}}\left\{\boldsymbol{A}^{-1}(W, \boldsymbol{V}^{\mathrm{T}})^{\mathrm{T}}, \boldsymbol{X}, \boldsymbol{Z}\right\} \middle| W, \boldsymbol{X}, \boldsymbol{Z}, R\right]$$
$$= E\{\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})|W, \boldsymbol{X}, \boldsymbol{Z}\}$$

and $E\{\boldsymbol{h}(\boldsymbol{X}, \boldsymbol{Z}, R)|W, \boldsymbol{V}, \boldsymbol{X}, \boldsymbol{Z})\} = E\{\boldsymbol{h}(\boldsymbol{X}, \boldsymbol{Z}, R)|W, \boldsymbol{X}, \boldsymbol{Z})\}$. Therefore, combining these two results, the requirement in (2.4) becomes

$$E[E\{\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})|W, \boldsymbol{X}, \boldsymbol{Z}\}|\boldsymbol{X}, \boldsymbol{Z}, R\}] = E[E\{\boldsymbol{h}(\boldsymbol{X}, \boldsymbol{Z}, R)|W, \boldsymbol{X}, \boldsymbol{Z})\}|\boldsymbol{X}, \boldsymbol{Z}, R].$$

However, this is of the form $E\{\boldsymbol{g}(W, \boldsymbol{X}, \boldsymbol{Z})|\boldsymbol{X}, \boldsymbol{Z}, R\} = \boldsymbol{0}$ with $\boldsymbol{g}(W, \boldsymbol{X}, \boldsymbol{Z}) = E\{\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}) - \boldsymbol{h}(\boldsymbol{X}, \boldsymbol{Z}, R)|W, \boldsymbol{X}, \boldsymbol{Z}\}$. Hence, from our second property derived above, we have that

$$E\{\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})|W, \boldsymbol{X}, \boldsymbol{Z}\} = E\{\boldsymbol{h}(\boldsymbol{X}, \boldsymbol{Z}, R)|W, \boldsymbol{X}, \boldsymbol{Z})\}.$$

We can thus write

$$E\{\boldsymbol{h}(\boldsymbol{X}, \boldsymbol{Z}, R)|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}\} = E\{\boldsymbol{h}(\boldsymbol{X}, \boldsymbol{Z}, R)|W, \boldsymbol{V}, \boldsymbol{X}, \boldsymbol{Z}\}$$
$$= E\{\boldsymbol{h}(\boldsymbol{X}, \boldsymbol{Z}, R)|W, \boldsymbol{X}, \boldsymbol{Z}\} = E\{\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})|W, \boldsymbol{X}, \boldsymbol{Z}\}.$$

Therefore, the efficient score vector is

$$\boldsymbol{S}_{\text{eff}} = \boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}) - E\{\boldsymbol{h}(\boldsymbol{X}, \boldsymbol{Z}, R)|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}\}$$
$$= \boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}) - E\{\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})|W, \boldsymbol{X}, \boldsymbol{Z}\}$$

which is a closed form solution. The above simplifies to the form in Proposition 2 because $\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}) = E\{\boldsymbol{U}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, R)|W, \boldsymbol{V}, \boldsymbol{X}, \boldsymbol{Z}\}$ and because $E\{\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})|W, \boldsymbol{X}, \boldsymbol{Z}\} = E\{\boldsymbol{U}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, R)|W, \boldsymbol{X}, \boldsymbol{Z}\}$.

S.1.5. Proof of Proposition 3. Direct calculation shows that for $f_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},R}$ given in (2.3),

$$\boldsymbol{U}_{\boldsymbol{\theta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, R_i) = \frac{\partial \log f_{Y|X,\boldsymbol{Z},R}(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, R_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$
$$= \frac{\partial \boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \boldsymbol{Y}_i - \boldsymbol{k}(\boldsymbol{X}_i, \boldsymbol{Z}_i, R_i),$$

where $\boldsymbol{k}(\boldsymbol{X}_i, \boldsymbol{Z}_i, R_i) = \sum_{j=1}^{m_i} \exp\{\eta(X_{ij}, \boldsymbol{Z}_{ij}) + R_i\}/[1 + \exp\{\eta(X_{ij}, \boldsymbol{Z}_{ij}; \boldsymbol{\theta}) + R_i\}]\partial\eta(X_{ij}, \boldsymbol{Z}_{ij}; \boldsymbol{\theta})/\partial\boldsymbol{\theta}$. To form the estimating equation based on $\boldsymbol{U}_{\boldsymbol{\theta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, R_i)$ as in Proposition 2, we utilize the special properties of W_i and $\boldsymbol{V}_i$ derived in Appendix S.1.4. Specifically, because R and $\boldsymbol{V}$ are conditionally independent given $(W, \boldsymbol{X}, \boldsymbol{Z})$ we have that for any $\boldsymbol{k}(\boldsymbol{X}_i, \boldsymbol{Z}_i, R_i)$,

$$E\{\boldsymbol{k}(\boldsymbol{X}_i, \boldsymbol{Z}_i, R_i)|W_i, V_i, \boldsymbol{X}_i, \boldsymbol{Z}_i\} - E\{\boldsymbol{k}(\boldsymbol{X}_i, \boldsymbol{Z}_i, R_i)|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i\} = \boldsymbol{0}.$$

Therefore, the term $\boldsymbol{k}(\boldsymbol{X}_i, \boldsymbol{Z}_i, R_i)$ does not contribute in computing $\boldsymbol{S}_{\text{eff}}$.

Next, because $\boldsymbol{Y}_i = \boldsymbol{A}_i^{-1}(W_i, \boldsymbol{V}_i^{\mathrm{T}})^{\mathrm{T}}$, we have that

$$\begin{aligned} & E(\boldsymbol{Y}_i|W_i, \boldsymbol{V}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) - E(\boldsymbol{Y}_i|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) \\ = \; & \boldsymbol{A}_i^{-1}(W_i, \boldsymbol{V}_i^{\mathrm{T}})^{\mathrm{T}} - E\{\boldsymbol{A}_i^{-1}(W_i, \boldsymbol{V}_i^{\mathrm{T}})^{\mathrm{T}}|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i\} \\ = \; & \boldsymbol{A}_i^{-1}[0, \boldsymbol{V}_i^{\mathrm{T}} - E(\boldsymbol{V}_i^{\mathrm{T}}|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})]^{\mathrm{T}}. \end{aligned}$$

Putting these two pieces together,

$$\boldsymbol{S}_{\text{eff}} \;=\; \frac{\partial\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\boldsymbol{A}_i^{-1}[0, \boldsymbol{V}_i^{\mathrm{T}} - E(\boldsymbol{V}_i^{\mathrm{T}}|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})]^{\mathrm{T}}.$$

Applying this to our model, we have $\partial\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})/\partial\boldsymbol{a} = \{\boldsymbol{B}(X_{i1}), \ldots, \boldsymbol{B}(X_{im_i})\}$ and $\partial\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})/\partial\boldsymbol{\beta} = \boldsymbol{Z}_i$. Hence, $\boldsymbol{S}_{\text{eff},\boldsymbol{a}}$ and $\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}$ are as specified.

S.1.6. Computation of $\boldsymbol{E(V_{ij}|W_i, X_i, Z_i)}$. Recall that $\boldsymbol{V}_i = (Y_{i2}, \ldots, Y_{im_i})^{\mathrm{T}}$, so below we interchangeably refer to Y_{ij}'s and V_{ij}'s.

The computation of $E(V_{ij}|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ involves the density $f_{V|W,\boldsymbol{X},\boldsymbol{Z}}(v_{ij}|w_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$. However, in Section S.1.4, we showed that $f_{V|W,\boldsymbol{X},\boldsymbol{Z}} = f_{\boldsymbol{V}|W,\boldsymbol{X},\boldsymbol{Z},R}$ and that

$$f_{V|W,\boldsymbol{X},\boldsymbol{Z}}(v_{ij}|w_i, \boldsymbol{x}_i, \boldsymbol{z}_i) = \frac{\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta})\boldsymbol{A}_i^{-1}(w_i, \boldsymbol{v}_i^{\mathrm{T}})^{\mathrm{T}}\right\}}{\int_{\mathcal{R}(\boldsymbol{v}_i)} \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta})\boldsymbol{A}_i^{-1}(w_i, \boldsymbol{v}_i^{\mathrm{T}})^{\mathrm{T}}\right\} d\mu(\boldsymbol{v}_i)}$$

as in equation (S.8).

Computing $E(V_{ij}|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ requires specifying the range $\mathcal{R}(\boldsymbol{v}_i)$. Specifying $\mathcal{R}(\boldsymbol{v}_i)$ involves using the fact that $W_i = \sum_{j=1}^{m_i} Y_{ij}$ and that Y_{ij} can either take values in $\{0, 1\}$ or in the interval $[0, 1]$ depending on censoring and whether or not $C_{ij} < t_0$. We therefore separate the computation based on the Y_{ij} ranges. Let $K_{ij} = 1$ if $Y_{ij} = I(T_{ij} \leq t_0)$, and let $K_{ij} = 0$ if $Y_{ij} = \text{pr}(T_{ij} \leq t_0)$. Also let $\mathcal{S}_m$ be a $2^m \times m$ matrix containing all 2^m combinations of 0's and 1's.

In the description below, we will make use of different components of $\mathcal{S}_m$. For ease in readability, we introduce the following notation. For an arbitrary matrix $\mathcal{S}$, let $s_{\ell j}$ denote its (ℓ, j) element, let $(\mathcal{S})_\ell$ denote its ℓth row, and $\mathcal{S}^{[-1]}$ denote the result of $\mathcal{S}$ after removing its first column.

Case 1: $W_i = 0$. Because $Y_{ij} \geq 0$, we must have $\mathcal{R}(\boldsymbol{v}_i) = (0, \ldots, 0)^{\mathrm{T}}$. In this case, the integrals in $E(V_{ij}|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ are both discrete sums.

Case 2: $W_i = m_i$. Because $Y_{ij} \leq 1$, we must have $\mathcal{R}(\boldsymbol{v}_i) = (1, \ldots, 1)^{\mathrm{T}}$. In this case, the integrals in $E(V_{ij}|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ are both discrete sums.

Case 3: $0 < W_i < m_i$.

1. $\sum_{j=1}^{m_i} K_{ij} = m_i$. Then $\mathcal{R}(\boldsymbol{v}_i)$ includes those rows of $\mathcal{S}_{m_i}^{[-1]}$ corresponding to the rows of $\mathcal{S}_{m_i}$ that sum to W_i. In this case, the integrals in $E(V_{ij}|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ are both discrete sums.
2. $\sum_{j=1}^{m_i} K_{ij} = m_i - 1$. Then let j^* denote the index associated with $K_{ij^*} = 0$. Define $\mathcal{S}^*_{m_i} = \mathcal{S}_{m_i}$, except with the entry $s_{\ell j^*}$ as $s_{\ell j^*} = W_i - \sum_{j \neq j^*} s_{\ell j}$ for $\ell = 1, \ldots, 2^{m_i}$. The range $\mathcal{R}(\boldsymbol{v}_i)$ includes those rows ℓ of $\mathcal{S}^{*[-1]}_{m_i}$ corresponding to the rows of $\mathcal{S}^*_{m_i}$ where $s_{\ell j^*} \in [0, 1]$. In this case, the integrals in $E(V_{ij}|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ are both discrete sums.
3. $\sum_{j=1}^{m_i} K_{ij} = m_i^*$ and $0 \leq m_i^* < m_i - 1$. Without loss of generality, suppose $K_{i1} = \cdots = K_{im_i^*} = 1$ so that $Y_{ij} \in \{0, 1\}$ for $j \leq m_i^*$, and $K_{i,m_i*+1} = \cdots = K_{im_i} = 0$ so that $Y_{ij} \in [0, 1]$ for $j > m_i^*$. Then, the range $\mathcal{R}(\boldsymbol{v}_i)$ is the union of each row of $\mathcal{S}^{[-1]}_{m_i^*}$ with $([0, 1], \ldots, [0, 1])$ for those combinations that satisfy $\sum_{j=1}^{m_i} Y_{ij} = W_i$. To write this more explicitly, note that the requirement of $W_i = \sum_{j=1}^{m_i} Y_{ij}$ is equivalent to $Y_{i1} = W_i - \sum_{j=2}^{m_i} Y_{ij}$. Also, because we require that $Y_{i1} \in [0, 1]$, we thus have that the range $\mathcal{R}(\boldsymbol{v}_i)$ is

$$\begin{aligned} \mathcal{R}(\boldsymbol{v}_i) \quad = \quad & \Bigg\{ \boldsymbol{v}_i : (v_{i2}, \ldots, v_{im_i^*}) \in \bigcup_{\ell=1}^{2^{m_i^*}} \left(\mathcal{S}^{[-1]}_{m_i^*} \right)_\ell , \\ & v_{i,m_i^*+1}, \ldots, v_{im_i} \in [0, 1], 0 \leq w_i - \sum_{j=2}^{m_i} v_{ij} \leq 1 \Bigg\}. \end{aligned}$$

In this case then, the integrals in $E(V_{ij}|W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ are a combination of sums over $(v_{i2}, \ldots, v_{im_i^*})$ and integrals over $(v_{i,m_i^*+1}, \ldots, v_{im_i})$.

Specifically, it is

$$\frac{\sum_{(v_{i2},\ldots,v_{im_i^*})\in\bigcup_{\ell=1}^{2^{m_i^*}}\left(\mathcal{S}_{m_i^*}^{[-1]}\right)_\ell}\int_{v_{i,m_i^*+1}\in[0,1]}\cdots\int_{v_{im_i}\in[0,1]} v_{ij}g(w_i,\boldsymbol{v}_i,\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta})dv_{i,m_i^*+1}\cdots dv_{im_i}}{\sum_{(v_{i2},\ldots,v_{im_i^*})\in\bigcup_{\ell=1}^{2^{m_i^*}}\left(\mathcal{S}_{m_i^*}^{[-1]}\right)_\ell}\int_{v_{i,m_i^*+1}\in[0,1]}\cdots\int_{v_{im_i}\in[0,1]} g(w_i,\boldsymbol{v}_i,\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta})dv_{i,m_i^*+1}\cdots dv_{im_i}}$$

where $g(w_i,\boldsymbol{v}_i,\boldsymbol{x}_i,\boldsymbol{z}_i)=\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta})\boldsymbol{A}_i^{-1}(w_i,\boldsymbol{v}_i^{\mathrm{T}})^{\mathrm{T}}\right\}$ and $(v_{i2},\ldots,v_{im_i})$ satisfies $0\le w_i-\sum_{j=1}^{m_i}v_{ij}\le 1$.

The expectation thus involves two multidimensional integrals over hypercubes. This can easily be carried out using the adaptIntegrate function in R as follows:

(a) Set $\ell=1$, $f_{\text{num}}=0$, $f_{\text{den}}=0$.

(b) Set $v_{i2},\ldots,v_{im_i^*}$ as the ℓth row of $\mathcal{S}_{m_i^*}^{[-1]}$.

(c) Define

$$f_1(v_{i,m_i*+1},\ldots,v_{i,m_i})=\begin{cases}v_{ij}g(w_i,\boldsymbol{v}_i,\boldsymbol{x}_i,\boldsymbol{z}_i), & \text{if } 0\le w_i-\sum_{j=2}^{m_i}v_{ij}\le 1,\\ 0, & \text{otherwise,}\end{cases}$$

and

$$f_2(v_{i,m_i*+1},\ldots,v_{i,m_i})=\begin{cases}g(w_i,\boldsymbol{v}_i,\boldsymbol{x}_i,\boldsymbol{z}_i), & \text{if } 0\le w_i-\sum_{j=2}^{m_i}v_{ij}\le 1,\\ 0, & \text{otherwise.}\end{cases}$$

The arguments of f_1, f_2 are to emphasize that the integrations will be performed on the variables $v_{i,m_i*+1},\ldots,v_{i,m_i}$. Recall, that $v_{i2},\ldots,v_{im_i^*}$ are fixed.

(d) Compute the integral on f_1, f_2 with the adaptIntegrate function as

$$\begin{aligned}f_{1,\text{int}} &= \text{adaptIntegrate}(f_1, \text{lowerLimit}=\mathbf{0}_{m_i-m_i^*},\\ &\quad \text{upperLimit}=\mathbf{1}_{m_i-m_i^*}, \text{fDim}=m_i-m_i^*),\\ f_{2,\text{int}} &= \text{adaptIntegrate}(f_2, \text{lowerLimit}=\mathbf{0}_{m_i-m_i^*},\\ &\quad \text{upperLimit}=\mathbf{1}_{m_i-m_i^*}, \text{fDim}=1),\end{aligned}$$

where $\mathbf{0}_{m_i-m_i^*}$ and $\mathbf{1}_{m_i-m_i^*}$ are $(m_i-m_i^*)$-dimensional vectors of zeros and ones, respectively.

(e) Update $f_{\text{num}}\leftarrow f_{\text{num}}+f_{1,\text{int}}$, and $f_{\text{den}}\leftarrow f_{\text{den}}+f_{2,\text{int}}$. Update $\ell\leftarrow\ell+1$. Go to Step (b).

S.1.7. Proof of Theorem 1. For simplicity and the convenience of the derivation, we slightly modify the estimation procedure in the proofs of Theorems 1 and 2. Instead of solving the two sets of equations in Proposition 3 simultaneously, we consider solving the first set of equations in Proposition 3 to obtain $\widehat{\boldsymbol{a}}$ as a function of $\boldsymbol{\beta}$, i.e. $\widehat{\boldsymbol{a}}(\boldsymbol{\beta})$, then plugging the resulting $\widehat{\boldsymbol{a}}(\boldsymbol{\beta})$ into the second set of equations to obtain $\widehat{\boldsymbol{\beta}}$. We finally update the estimator to obtain $\widehat{\boldsymbol{a}}(\widehat{\boldsymbol{\beta}})$. That is, we implement the profiling procedure. Note that the resulting estimator solves all the equations in Proposition 3 hence is equivalent to solving the estimating equations in Proposition 3 directly to obtain the estimator $\widehat{\boldsymbol{\theta}}$ simultaneously. We first establish several lemmas.

Lemma 1. *For $j, k = 1, \ldots, m_i$, let*

$$c_{ijk}(\boldsymbol{\beta}, \boldsymbol{a}) = \frac{\int_{\mathcal{R}(\boldsymbol{v}_i)} v_{ij} v_{ik} \exp\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) \boldsymbol{J}_i \boldsymbol{v}_i\} d\mu(\boldsymbol{v}_i)}{\int_{\mathcal{R}(\boldsymbol{v}_i)} \exp\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) \boldsymbol{J}_i \boldsymbol{v}_i\} d\mu(\boldsymbol{v}_i)}$$
$$- \frac{\left[\int_{\mathcal{R}(\boldsymbol{v}_i)} v_{ij} \exp\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) \boldsymbol{J}_i \boldsymbol{v}_i\} d\mu(\boldsymbol{v}_i)\right] \left[\int_{\mathcal{R}(\boldsymbol{v}_i)} v_{ik} \exp\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) \boldsymbol{J}_i \boldsymbol{v}_i\} d\mu(\boldsymbol{v}_i)\right]}{\left[\int_{\mathcal{R}(\boldsymbol{v}_i)} \exp\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) \boldsymbol{J}_i \boldsymbol{v}_i\} d\mu(\boldsymbol{v}_i)\right]^2},$$

and define $\boldsymbol{C}_i(\boldsymbol{\beta}, \boldsymbol{a}) = (c_{ijk}, j, k = 2, \ldots, m_i)$. Let $\boldsymbol{C}_i(\boldsymbol{\beta}, \alpha)$ be $\boldsymbol{C}_i(\boldsymbol{\beta}, \boldsymbol{a})$ with $\mathbf{B}(x)^{\mathrm{T}}\boldsymbol{a}$ replaced with $\alpha(x)$ inside $\boldsymbol{\eta}$. Then both $\boldsymbol{C}_i(\boldsymbol{\beta}, \boldsymbol{a})$ and $\boldsymbol{C}_i(\boldsymbol{\beta}, \alpha)$ are symmetric positive-definite matrices of size $(m_i - 1) \times (m_i - 1)$. In addition,

$$\begin{aligned}
\boldsymbol{V}_{n,\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}, \alpha) &= \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{Z}_{i1}, \ldots, \boldsymbol{Z}_{im_i}) \boldsymbol{J}_i \boldsymbol{C}_i(\boldsymbol{\beta}, \alpha) \boldsymbol{J}_i^{\mathrm{T}} (\boldsymbol{Z}_{i1}, \ldots, \boldsymbol{Z}_{im_i})^{\mathrm{T}}, \\
\boldsymbol{V}_{n,\boldsymbol{\beta}\boldsymbol{a}}(\boldsymbol{\beta}, \alpha) &= \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{Z}_{i1}, \ldots, \boldsymbol{Z}_{im_i}) \boldsymbol{J}_i \boldsymbol{C}_i(\boldsymbol{\beta}, \alpha) \boldsymbol{J}_i^{\mathrm{T}} \{\boldsymbol{B}(X_{i1}), \ldots, \boldsymbol{B}(X_{im_i})\}^{\mathrm{T}}, \\
\boldsymbol{V}_{n,\boldsymbol{a}\boldsymbol{\beta}}(\boldsymbol{\beta}, \alpha) &= \frac{1}{n} \sum_{i=1}^{n} \{\boldsymbol{B}(X_{i1}), \ldots, \boldsymbol{B}(X_{im_i})\} \boldsymbol{J}_i \boldsymbol{C}_i(\boldsymbol{\beta}, \alpha) \boldsymbol{J}_i^{\mathrm{T}} (\boldsymbol{Z}_{i1}, \ldots, \boldsymbol{Z}_{im_i})^{\mathrm{T}}, \\
\boldsymbol{V}_{n,\boldsymbol{a}\boldsymbol{a}}(\boldsymbol{\beta}, \alpha) &= \frac{1}{n} \sum_{i=1}^{n} \{\boldsymbol{B}(X_{i1}), \ldots, \boldsymbol{B}(X_{im_i})\} \boldsymbol{J}_i \boldsymbol{C}_i(\boldsymbol{\beta}, \alpha) \boldsymbol{J}_i^{\mathrm{T}} \{\boldsymbol{B}(X_{i1}), \ldots, \boldsymbol{B}(X_{im_i})\}^{\mathrm{T}},
\end{aligned}$$

and they equal the sample averages $n^{-1} \sum_{i=1}^{n} E\{\boldsymbol{S}_{\mathrm{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})^{\otimes 2} \mid \boldsymbol{X}_i, \boldsymbol{Z}_i\}$, $n^{-1} \sum_{i=1}^{n} E\{\boldsymbol{S}_{\mathrm{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) \boldsymbol{S}_{\mathrm{eff},\boldsymbol{a}}^{\mathrm{T}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) \mid \boldsymbol{X}_i, \boldsymbol{Z}_i\}$, $n^{-1} \sum_{i=1}^{n} E\{\boldsymbol{S}_{\mathrm{eff},\boldsymbol{a}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) \boldsymbol{S}_{\mathrm{eff},\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) \mid \boldsymbol{X}_i, \boldsymbol{Z}_i\}$ and $n^{-1} \sum_{i=1}^{n} E\{\boldsymbol{S}_{\mathrm{eff},\boldsymbol{a}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\theta})^{\otimes 2} \mid \boldsymbol{X}_i, \boldsymbol{Z}_i\}$

respectively, with $\boldsymbol{B}^{\mathrm{T}}(x)\boldsymbol{a}$ *replaced by* $\alpha(x)$*. Note that* $\boldsymbol{V}_{n,\boldsymbol{\beta a}}(\boldsymbol{\beta},\alpha)=\boldsymbol{V}^{\mathrm{T}}_{n,\boldsymbol{a\beta}}(\boldsymbol{\beta},\alpha)$ *and* $\boldsymbol{V}_{n,\boldsymbol{\beta\beta}}(\boldsymbol{\beta},\alpha)$ *and* $\boldsymbol{V}_{n,\boldsymbol{aa}}(\boldsymbol{\beta},\alpha)$ *are symmetric positive definite matrices.*

Proof of Lemma 1:
Through careful calculation, noting that the multiplier that contains only w_i without $\boldsymbol{v}_i$ can be canceled in the ratio of the integrations, we can therefore write

$$E(V_{ij}|W_i,\boldsymbol{X}_i,\boldsymbol{Z}_i;\boldsymbol{\theta}) \;=\; \frac{\int_{\mathcal{R}(\boldsymbol{v}_i)} v_{ij}\exp\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i,\boldsymbol{Z}_i;\boldsymbol{\theta})\boldsymbol{J}_i\boldsymbol{v}_i\}d\mu(\boldsymbol{v}_i)}{\int_{\mathcal{R}(\boldsymbol{v}_i)}\exp\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i,\boldsymbol{Z}_i;\boldsymbol{\theta})\boldsymbol{J}_i\boldsymbol{v}_i\}d\mu(\boldsymbol{v}_i)}.$$

We can further verify that

$$\begin{aligned}
&\frac{\partial\int_{\mathcal{R}(\boldsymbol{v}_i)}\exp\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i,\boldsymbol{Z}_i;\boldsymbol{\theta})\boldsymbol{J}_i\boldsymbol{v}_i\}d\mu(\boldsymbol{v}_i)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\\
=&\sum_{k=1}^{m_i}\int_{\mathcal{R}(\boldsymbol{v}_i)} v_{ik}\exp\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i,\boldsymbol{Z}_i;\boldsymbol{\theta})\boldsymbol{J}_i\boldsymbol{v}_i\}d\mu(\boldsymbol{v}_i)(\boldsymbol{Z}_{ik}-\boldsymbol{Z}_{i1})^{\mathrm{T}},\\
&\frac{\partial\int_{\mathcal{R}(\boldsymbol{v}_i)} v_{ij}\exp\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i,\boldsymbol{Z}_i;\boldsymbol{\theta})\boldsymbol{J}_i\boldsymbol{v}_i\}d\mu(\boldsymbol{v}_i)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\\
=&\sum_{k=1}^{m_i}\int_{\mathcal{R}(\boldsymbol{v}_i)} v_{ij}v_{ik}\exp\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i,\boldsymbol{Z}_i;\boldsymbol{\theta})\boldsymbol{J}_i\boldsymbol{v}_i\}d\mu(\boldsymbol{v}_i)(\boldsymbol{Z}_{ik}-\boldsymbol{Z}_{i1})^{\mathrm{T}},\\
&\frac{\partial\int_{\mathcal{R}(\boldsymbol{v}_i)}\exp\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i,\boldsymbol{Z}_i;\boldsymbol{\theta})\boldsymbol{J}_i\boldsymbol{v}_i\}d\mu(\boldsymbol{v}_i)}{\partial\boldsymbol{a}^{\mathrm{T}}}\\
=&\sum_{k=1}^{m_i}\int_{\mathcal{R}(\boldsymbol{v}_i)} v_{ik}\exp\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i,\boldsymbol{Z}_i;\boldsymbol{\theta})\boldsymbol{J}_i\boldsymbol{v}_i\}d\mu(\boldsymbol{v}_i)\{\boldsymbol{B}(X_{ik})-\boldsymbol{B}(X_{i1})\}^{\mathrm{T}},\\
&\frac{\partial\int_{\mathcal{R}(\boldsymbol{v}_i)} v_{ij}\exp\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i,\boldsymbol{Z}_i;\boldsymbol{\theta})\boldsymbol{J}_i\boldsymbol{v}_i\}d\mu(\boldsymbol{v}_i)}{\partial\boldsymbol{a}^{\mathrm{T}}}\\
=&\sum_{k=1}^{m_i}\int_{\mathcal{R}(\boldsymbol{v}_i)} v_{ij}v_{ik}\exp\{\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{X}_i,\boldsymbol{Z}_i;\boldsymbol{\theta})\boldsymbol{J}_i\boldsymbol{v}_i\}d\mu(\boldsymbol{v}_i)\{\boldsymbol{B}(X_{ik})-\boldsymbol{B}(X_{i1})\}^{\mathrm{T}}.
\end{aligned}$$

Taking into account the definition of $c_{ijk}(\boldsymbol{\beta},\alpha)$, we can then easily verify the displayed results in the Lemma. It is also easy to verify that $\mathrm{cov}[\{V_{ij}-E(V_{ij}|W_i,\boldsymbol{X}_i,\boldsymbol{Z}_i;\boldsymbol{\theta})\}\{V_{ik}-E(V_{ik}|W_i,\boldsymbol{X}_i,\boldsymbol{Z}_i;\boldsymbol{\theta})\}]=c_{ijk}(\boldsymbol{\beta},\alpha)$. Thus the displayed matrices are indeed the corresponding variance-covariance matrices. □

Lemma 2. *For any vector* $\boldsymbol{a} = (a_1, \ldots, a_{p_2})^{\mathrm{T}}$, *there exist constants* $0 < c_a \leq C_a < \infty$, *such that for sufficiently large* n,

$$c_a \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}h \leq \boldsymbol{a}^{\mathrm{T}}\boldsymbol{V}_{\boldsymbol{aa}}(\boldsymbol{\beta}_0, \alpha_0)\boldsymbol{a} \leq C_a \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}h, \tag{S.9}$$

$$c_a h \leq \|\boldsymbol{V}_{\boldsymbol{aa}}(\boldsymbol{\beta}_0, \alpha_0)\|_2 \leq C_a h, \tag{S.10}$$

$$\max_{1\leq k_1,k_2\leq p_2} |\boldsymbol{V}_{n,\boldsymbol{aa}}(\boldsymbol{\beta}_0, \alpha_0) - \boldsymbol{V}_{\boldsymbol{aa}}(\boldsymbol{\beta}_0, \alpha_0)| = O_p\{\sqrt{hn^{-1}log(n)}\}. \tag{S.11}$$

Proof of Lemma 2:

Note that $\boldsymbol{C}_i(\boldsymbol{\beta}_0, \alpha_0)$ is a positive-defnite matrix with positive eigenvalue. Assume the eigenvalues are $0 < \rho_1 < \cdots < \rho_{m_i-1}$. Thus

$$\begin{aligned}
& \boldsymbol{a}^{\mathrm{T}}\boldsymbol{V}_{\boldsymbol{aa}}(\boldsymbol{\beta}, \alpha_0)\boldsymbol{a} \\
= \; & E\left[\boldsymbol{a}^{\mathrm{T}}\{\boldsymbol{B}(X_{i1}), \ldots, \boldsymbol{B}(X_{im_i})\}\boldsymbol{J}_i\boldsymbol{C}_i(\boldsymbol{\beta}, \alpha_0)\boldsymbol{J}_i^{\mathrm{T}}\{\boldsymbol{B}(X_{i1}), \ldots, \boldsymbol{B}(X_{im_i})\}^{\mathrm{T}}\boldsymbol{a}\right] \\
= \; & E\left[\boldsymbol{a}^{\mathrm{T}}\{\boldsymbol{B}(X_{i2}) - \boldsymbol{B}(X_{i1}), \ldots, \boldsymbol{B}(X_{im_i}) - \boldsymbol{B}(X_{i1})\}\boldsymbol{C}_i(\boldsymbol{\beta}, \alpha_0)\right. \\
& \left.\{\boldsymbol{B}(X_{i2}) - \boldsymbol{B}(X_{i1}), \ldots, \boldsymbol{B}(X_{im_i}) - \boldsymbol{B}(X_{i1})\}^{\mathrm{T}}\boldsymbol{a}\right] \\
= \; & \boldsymbol{a}^{\mathrm{T}}E\left[\rho^{(i)}\{\boldsymbol{B}(X_{i2}) - \boldsymbol{B}(X_{i1}), \ldots, \boldsymbol{B}(X_{im_i}) - \boldsymbol{B}(X_{i1})\}^{\otimes 2}\right]\boldsymbol{a} \\
= \; & \boldsymbol{a}^{\mathrm{T}}E\left[\rho^{(i)}\sum_{j=2}^{m_i}\{\boldsymbol{B}(X_{ij}) - \boldsymbol{B}(X_{i1})\}\{\boldsymbol{B}(X_{ij}) - \boldsymbol{B}(X_{i1})\}^{\mathrm{T}}\right]\boldsymbol{a} \\
= \; & \sum_{j=2}^{m}\boldsymbol{a}^{\mathrm{T}}E\left[\rho^{(i)}\{\boldsymbol{B}(X_{ij}) - \boldsymbol{B}(X_{i1})\}\{\boldsymbol{B}(X_{ij}) - \boldsymbol{B}(X_{i1})\}^{\mathrm{T}}\right]\boldsymbol{a} \\
= \; & \sum_{j=2}^{m} E\left\{\rho^{(i)}\|\boldsymbol{B}(X_{ij})^{\mathrm{T}}\boldsymbol{a} - \boldsymbol{B}(X_{i1})^{\mathrm{T}}\boldsymbol{a}\|^2\right\}, \qquad \text{(S.12)}
\end{aligned}$$

where $\rho_1 < \rho^{(i)} < \rho_{m_i-1}$. From Theorem 5.4.2 on page 145 of DeVore and Lorentz (1993), (S.12) further leads to $c_1\|\boldsymbol{a}\|^2 h \leq \|\boldsymbol{B}(X_{ij})^{\mathrm{T}}\boldsymbol{a} - \boldsymbol{B}(X_{i1})^{\mathrm{T}}\boldsymbol{a}\|^2 \leq c_2\|\boldsymbol{a}\|^2 h$ for $0 < c_1 < c_2 < \infty$. This leads to (S.9). Further (S.9) implies that $\boldsymbol{V}_{\boldsymbol{aa}}(\boldsymbol{\beta}_0, \alpha_0)$ has eigenvalues between $c_a h$ and $C_a h$. Since $\boldsymbol{V}_{\boldsymbol{aa}}(\boldsymbol{\beta}_0, \alpha_0)$ is symmetric, hence it also has singular values between $c_a h$ and $C_a h$, i.e. $\|\boldsymbol{V}_{\boldsymbol{aa}}(\boldsymbol{\beta}_0, \alpha_0)\|_2$ is between $c_a h$ and $C_a h$, which leads to (S.10). Finally, (S.11) is a direct result of Bernstein's inequality in Bosq (1998). □

Lemma 3. *Let* c_a, C_a *be defined in Lemma 2. Let* C_S *be a constant s.t.*

$0 < C_S < \infty$. *Then*

$$C_a^{-1}h^{-1} \leq \|\boldsymbol{V}_{\boldsymbol{aa}}^{-1}(\boldsymbol{\beta}_0, \alpha_0)\|_2 \leq c_a^{-1}h^{-1}, \tag{S.13}$$

$$\|\boldsymbol{V}_{\boldsymbol{aa}}^{-1}(\boldsymbol{\beta}_0, \alpha_0)\|_\infty \leq C_S h^{-1} \tag{S.14}$$

Proof of Lemma 3: Result (S.13) follows directly from (S.10). (S.14) follows from Lemma 2, (S.13) and Theorem 13.4.3 in DeVore and Lorentz (1993). We omit the details here. □

Since $\alpha_0(x) \in C^q[a,b]$, following de Boor (2001), there exists $\boldsymbol{a} \in R^{p_2}$, such that

$$\sup_{x\in[a,b]} \left|\alpha_0(x) - \boldsymbol{B}^{\mathrm{T}}(x)\boldsymbol{a}\right| = O(h^q). \tag{S.15}$$

Let $\widetilde{\alpha}(x) = \boldsymbol{B}^{\mathrm{T}}(x)\boldsymbol{a}$. Also, let $\boldsymbol{g}_{\boldsymbol{a}}(\boldsymbol{\beta}, \boldsymbol{a}) = E\{\boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}, \boldsymbol{a})\}$, $\boldsymbol{g}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{a}) = E\{\boldsymbol{g}_{n,\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{a})\}$, $\boldsymbol{g}_{\boldsymbol{a}}(\boldsymbol{\beta}, \alpha) = E\{\boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}, \alpha)\}$, $\boldsymbol{g}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \alpha) = E\{\boldsymbol{g}_{n,\boldsymbol{\beta}}(\boldsymbol{\beta}, \alpha)\}$, $\boldsymbol{g}_n(\boldsymbol{\beta}, \boldsymbol{a}) = \{\boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}, \boldsymbol{a})^{\mathrm{T}}, \boldsymbol{g}_{n,\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{a})^{\mathrm{T}}\}^{\mathrm{T}}$, and $\boldsymbol{g}(\boldsymbol{\beta}, \boldsymbol{a}) = \{\boldsymbol{g}_{\boldsymbol{a}}(\boldsymbol{\beta}, \boldsymbol{a})^{\mathrm{T}}, \boldsymbol{g}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{a})^{\mathrm{T}}\}^{\mathrm{T}}$. We have that at the true parameter values $\boldsymbol{\beta}_0, \alpha_0$, $\boldsymbol{g}(\boldsymbol{\beta}_0, \alpha_0) = \boldsymbol{0}$. Because of (S.15), $\|\boldsymbol{g}(\boldsymbol{\beta}_0, \boldsymbol{a})\|_2 = \|\boldsymbol{g}(\boldsymbol{\beta}_0, \widetilde{\alpha})\|_2 = o_p(1)$. From condition (C4) and the law of large numbers, $\|\boldsymbol{g}_n(\boldsymbol{\beta}_0, \boldsymbol{a})\|_2 = o_p(1)$. However, $\boldsymbol{g}_n(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{a}}) = \boldsymbol{0}$, hence the uniqueness of zero and the continuous differentiable property leads to $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = o_p(1)$ and $\|\widehat{\boldsymbol{a}} - \boldsymbol{a}\|_2 = o_p(1)$. Hence $\|\boldsymbol{B}^{\mathrm{T}}(x)\widehat{\boldsymbol{a}} - \alpha_0(x)\|_\infty \leq \|\boldsymbol{B}^{\mathrm{T}}(x)\widehat{\boldsymbol{a}} - \widetilde{\alpha}(x)\|_\infty + O(h^q) = \|\boldsymbol{B}^{\mathrm{T}}(x)(\widehat{\boldsymbol{a}} - \boldsymbol{a})\|_\infty + O(h^q) \leq \sup_{x\in[a,b]} \|\boldsymbol{B}(x)\|_2 \|\widehat{\boldsymbol{a}} - \boldsymbol{a}\|_2 + O(h^q) \to 0$. Note that $\widehat{\alpha}(x) = \boldsymbol{B}^{\mathrm{T}}(x)\widehat{\boldsymbol{a}}$ and

$$\begin{aligned}
& \|\boldsymbol{V}_{n,\boldsymbol{aa}}(\boldsymbol{\beta}_0, \boldsymbol{a}) - \boldsymbol{V}_{n,\boldsymbol{aa}}(\boldsymbol{\beta}_0, \alpha_0)\|_\infty \\
= & \|\frac{1}{n}\sum_{i=1}^{n}\{\boldsymbol{B}(X_{i1}), \ldots, \boldsymbol{B}(X_{im_i})\}\boldsymbol{J}_i\{\boldsymbol{C}_i(\boldsymbol{\beta}_0, \boldsymbol{a}) - \boldsymbol{C}_i(\boldsymbol{\beta}_0, \alpha_0)\}\boldsymbol{J}_i^{\mathrm{T}}\{\boldsymbol{B}(X_{i1}), \ldots, \boldsymbol{B}(X_{im_i})\}^{\mathrm{T}}\|_\infty \\
= & \|O(h^q)n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{m_i}\boldsymbol{B}(X_{ij})\boldsymbol{B}^{\mathrm{T}}(X_{ik})\|_\infty \\
= & O(h^q)\sup_l\left\{n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{m_i}\mathbf{B}_l(X_{ij})\sum_{l=1}^{p_2}\mathbf{B}_l(X_{ik})\right\} \\
= & O(h^q)\sup_l\left\{n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m_i} m_i\mathbf{B}_l(X_{ij})\right\} \\
= & O(h^{q+1}), \qquad \text{(S.16)}
\end{aligned}$$

where in the last two steps, we used the fact that each B-spline basis function takes values in $[0,1]$, are supported on only $2r-1$ intervals $[\xi_k, \xi_{k+1}]$, and the summation of all B-spline basis functions at any value x is 1. From (S.14) and (S.11), $\|\boldsymbol{V}_{n,\boldsymbol{aa}}(\boldsymbol{\beta}_0,\alpha_0)^{-1}\|_\infty = O_p(h^{-1})$. Further from (S.15), $\|\boldsymbol{V}_{n,\boldsymbol{aa}}(\boldsymbol{\beta}_0,\boldsymbol{a})^{-1}\|_\infty = O_p(h^{-1})$. Thus, using (S.16), we have

$$\begin{aligned} & \|\boldsymbol{V}_{n,\boldsymbol{aa}}^{-1}(\boldsymbol{\beta}_0,\boldsymbol{a}) - \boldsymbol{V}_{n,\boldsymbol{aa}}^{-1}(\boldsymbol{\beta}_0,\alpha_0)\|_\infty \\ \leq\ & \|\boldsymbol{V}_{n,\boldsymbol{aa}}^{-1}(\boldsymbol{\beta}_0,\boldsymbol{a})\|_\infty \|\boldsymbol{V}_{n,\boldsymbol{aa}}(\boldsymbol{\beta}_0,\boldsymbol{a}) - \boldsymbol{V}_{n,\boldsymbol{aa}}(\boldsymbol{\beta}_0,\alpha_0)\|_\infty \|\boldsymbol{V}_{n,\boldsymbol{aa}}^{-1}(\boldsymbol{\beta}_0,\alpha_0)\|_\infty \\ =\ & O_p(h^{q-1}). \end{aligned}$$

From (S.13) and (S.11), there are constants $0 < c_v' < C_v' < \infty$, such that with probability 1, $c_v' h^{-1} \leq \|\boldsymbol{V}_{n,\boldsymbol{aa}}^{-1}(\boldsymbol{\beta}_0,\alpha_0)\|_2 \leq C_v' h^{-1}$ and

$$\begin{aligned} & \|\boldsymbol{V}_{n,\boldsymbol{aa}}^{-1}(\boldsymbol{\beta}_0,\alpha_0) - \boldsymbol{V}_{\boldsymbol{aa}}^{-1}(\boldsymbol{\beta}_0,\alpha_0)\|_2 \\ \leq\ & \|\boldsymbol{V}_{n,\boldsymbol{aa}}^{-1}(\boldsymbol{\beta}_0,\alpha_0)\|_2 \|\boldsymbol{V}_{n,\boldsymbol{aa}}(\boldsymbol{\beta}_0,\alpha_0) - \boldsymbol{V}_{\boldsymbol{aa}}(\boldsymbol{\beta}_0,\alpha_0)\|_2 \|\boldsymbol{V}_{\boldsymbol{aa}}^{-1}(\boldsymbol{\beta}_0,\alpha_0)\|_2 \\ =\ & O_p\{h^{-2}\sqrt{hn^{-1}\log(n)}\}. \end{aligned} \tag{S.17}$$

By (S.15), following the same derivation as in (S.16),

$$\|\boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}_0,\boldsymbol{a}) - \boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}_0,\alpha_0)\|_\infty = O(h^q)\|n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\boldsymbol{B}(X_{ij})\|_\infty = O_p\left(h^{q+1}\right).$$

By Bernstein's inequality in Bosq (1998), $\|\boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}_0,\alpha_0)\|_\infty = O_p\left\{\sqrt{hn^{-1}\log(n)}\right\}$. Thus for any $\boldsymbol{c} \in R^{p_2}$ with $\|\boldsymbol{c}\|_2 = 1$,

$$\begin{aligned} & |\boldsymbol{c}^{\mathrm{T}}\left\{\boldsymbol{V}_{n,\boldsymbol{aa}}^{-1}(\boldsymbol{\beta}_0,\boldsymbol{a})\boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}_0,\boldsymbol{a}) - \boldsymbol{V}_{n,\boldsymbol{aa}}^{-1}(\boldsymbol{\beta}_0,\alpha_0)\boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}_0,\alpha_0)\right\}| \\ \leq\ & \|\boldsymbol{c}\|_\infty \left\|\boldsymbol{V}_{n,\boldsymbol{aa}}^{-1}(\boldsymbol{\beta}_0,\boldsymbol{a})\right\|_\infty \left\|\boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}_0,\boldsymbol{a}) - \boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}_0,\alpha_0)\right\|_\infty \\ & + \|\boldsymbol{c}\|_\infty \left\|\boldsymbol{V}_{n,\boldsymbol{aa}}^{-1}(\boldsymbol{\beta}_0,\boldsymbol{a}) - \boldsymbol{V}_{n,\boldsymbol{aa}}^{-1}(\boldsymbol{\beta}_0,\alpha_0)\right\|_\infty \left\|\boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}_0,\alpha_0)\right\|_\infty \\ =\ & O_p\left(h^q\right) + O_p\left(h^{q-1}\right) O_p\left\{\sqrt{hn^{-1}\log(n)}\right\}. \end{aligned} \tag{S.18}$$

Define $\widehat{\boldsymbol{e}} = \boldsymbol{V}_{n,\boldsymbol{aa}}^{-1}(\boldsymbol{\beta}_0,\alpha_0)\boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}_0,\alpha_0)$ and $\widehat{\sigma}^2(x,\boldsymbol{\beta}_0) = \boldsymbol{B}^{\mathrm{T}}(x)\mathrm{var}\left(\widehat{\boldsymbol{e}}\,|\boldsymbol{X},\boldsymbol{Z}\right)\boldsymbol{B}(x)$, where $\boldsymbol{X},\boldsymbol{Z}$ denote the collection of all the covariates. Then from the Central Limit Theorem, we obtain $\{\boldsymbol{B}^{\mathrm{T}}(x)\mathrm{var}(\widehat{\boldsymbol{e}}|\boldsymbol{X},\boldsymbol{Z})\boldsymbol{B}(x)\}^{-1/2}\boldsymbol{B}^{\mathrm{T}}(x)\widehat{\boldsymbol{e}} \to \mathrm{Normal}(0,1)$. Note that $\mathrm{var}\left(\widehat{\boldsymbol{e}}\,|\boldsymbol{X},\boldsymbol{Z}\right) = n^{-1}\boldsymbol{V}_{n,\boldsymbol{aa}}^{-1}(\boldsymbol{\beta}_0,\alpha_0)$. Hence

$$\begin{aligned} \widehat{\sigma}^2(x,\boldsymbol{\beta}_0) &= \frac{1}{n}\boldsymbol{B}^{\mathrm{T}}(x)\boldsymbol{V}_{n,\boldsymbol{aa}}^{-1}(\boldsymbol{\beta}_0,\alpha_0)\boldsymbol{B}(x) \\ &= \frac{1}{n}\boldsymbol{B}^{\mathrm{T}}(x)\boldsymbol{V}_{\boldsymbol{aa}}^{-1}(\boldsymbol{\beta}_0,\alpha_0)\boldsymbol{B}(x)\{1+o_p(1)\}. \end{aligned}$$

Note that for any $x \in [a,b]$, only $2r-1$ terms in $B_1(x),\ldots,B_{p_2}(x)$ can be nonzero, in addition, $0 \le B_k(x) \le 1$ for all $x \in [a,b]$ and all $k = 1,\ldots,p_2$, hence Lemma (3) implies that $\boldsymbol{B}^{\mathrm{T}}(x)\boldsymbol{V}^{-1}_{\boldsymbol{aa}}(\boldsymbol{\beta}_0,\alpha_0)\boldsymbol{B}(x)$ is an order h^{-1} quantity. Thus, there exist constants $0 < c_\sigma \le C_\sigma < \infty$ such that with probability 1 and for sufficiently large n,

$$c_\sigma(nh)^{-1/2} \le \inf_{x\in[a,b]} \widehat{\sigma}(x,\boldsymbol{\beta}_0) \le \sup_{x\in[a,b]} \widehat{\sigma}(x,\boldsymbol{\beta}_0) \le C_\sigma(nh)^{-1/2}. \tag{S.19}$$

Thus $\boldsymbol{B}^{\mathrm{T}}(x)\widehat{\boldsymbol{e}} = O_p\{(nh)^{-1/2}\}$ uniformly in $x \in [a,b]$, and

$$\boldsymbol{B}^{\mathrm{T}}(x)\boldsymbol{V}^{-1}_{n,\boldsymbol{aa}}(\boldsymbol{\beta}_0,\boldsymbol{a})\boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}_0,\boldsymbol{a}) = O_p\left\{(nh)^{-1/2} + h^q\right\}$$

uniformly in $x \in [a,b]$. By the consistency of $\widehat{\boldsymbol{a}}$ and the Taylor expansion,

$$\widehat{\boldsymbol{a}}(\boldsymbol{\beta}_0) - \boldsymbol{a} = \boldsymbol{V}^{-1}_{n,\boldsymbol{aa}}(\boldsymbol{\beta}_0,\boldsymbol{a})\boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}_0,\boldsymbol{a})\left\{1+o_p(1)\right\}. \tag{S.20}$$

Thus by (S.18), (S.19), and Condition (C3),

$$\begin{aligned}
&\sup_{x\in[a,b]} \left|\widehat{\sigma}(x,\boldsymbol{\beta}_0)^{-1}\left[\boldsymbol{B}^{\mathrm{T}}(x)\left\{\widehat{\boldsymbol{a}}(\boldsymbol{\beta}_0)-\boldsymbol{a}\right\} - \boldsymbol{B}^{\mathrm{T}}(x)\widehat{\boldsymbol{e}}\right]\right| \\
= &\sup_{x\in[a,b]} \left|\widehat{\sigma}(x,\boldsymbol{\beta}_0)^{-1}\left[\boldsymbol{B}^{\mathrm{T}}(x)\boldsymbol{V}^{-1}_{n,\boldsymbol{aa}}(\boldsymbol{\beta}_0,\boldsymbol{a})\boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}_0,\boldsymbol{a})\left\{1+o_p(1)\right\}\right.\right. \\
&\left.\left. -\boldsymbol{B}^{\mathrm{T}}(x)\boldsymbol{V}^{-1}_{n,\boldsymbol{aa}}(\boldsymbol{\beta}_0,\alpha_0)\boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}_0,\alpha_0)\right]\right| \\
= &\; O_p\left\{(nh)^{1/2}\right\}\left\{O_p(h^q) + O_p\left(h^{q-1}\right)O_p\left(\sqrt{hn^{-1}\log(n)}\right)\right\} \\
&+O_p\left\{(nh)^{1/2}\right\} o_p\{(nh)^{-1/2}+h^q\} \\
= &\; o_p(1).
\end{aligned}$$

By Slutsky's theorem $\widehat{\sigma}^{-1}(x,\boldsymbol{\beta}_0)\left\{\widehat{\alpha}(x,\boldsymbol{\beta}_0) - \widetilde{\alpha}(x)\right\} \to \mathrm{Normal}(0,1)$ and $\widehat{\alpha}(x,\boldsymbol{\beta}_0) - \widetilde{\alpha}(x) = O_p\left\{(nh)^{-1/2}\right\}$ uniformly in $x \in [a,b]$. From $\sup_{x\in[a,b]}|\alpha_0(x) - \widetilde{\alpha}(x)| = o(h^q)$, we obtain $\sup_{x\in[a,b]}|\widehat{\alpha}(x,\boldsymbol{\beta}_0) - \alpha_0(x)| = O_p\{(nh)^{-1/2}+h^q\}$. Further from Slutsky's theorem,

$$\widehat{\sigma}^{-1}(x,\boldsymbol{\beta}_0)\left\{\widehat{\alpha}(x,\boldsymbol{\beta}_0) - \alpha_0(x)\right\} \to \mathrm{Normal}(0,1).$$

□

S.1.8. Proof of Theorem 2. Following the consistency results established in Theorem 1, we perform a Taylor expansion

$$\begin{aligned}\mathbf{0} &= n^{1/2}\boldsymbol{g}_{n,\boldsymbol{\beta}}\{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{a}}(\widehat{\boldsymbol{\beta}})\} \\ &= n^{1/2}\boldsymbol{g}_{n,\boldsymbol{\beta}}\{\boldsymbol{\beta}_0, \widehat{\boldsymbol{a}}(\boldsymbol{\beta}_0)\} + \left\{\frac{\partial \boldsymbol{g}_{n,\boldsymbol{\beta}}\{\boldsymbol{\beta}, \widehat{\boldsymbol{a}}(\boldsymbol{\beta})\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}} + \frac{\partial \boldsymbol{g}_{n,\boldsymbol{\beta}}\{\boldsymbol{\beta}, \widehat{\boldsymbol{a}}(\boldsymbol{\beta})\}}{\partial \widehat{\boldsymbol{a}}(\boldsymbol{\beta})^{\mathrm{T}}}\frac{\partial \widehat{\boldsymbol{a}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right\}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),\end{aligned}$$

where $\boldsymbol{\beta}^*$ lies on the line connecting $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$. The profiling procedures imply that $\boldsymbol{g}_{n,\boldsymbol{a}}\{\boldsymbol{\beta}, \widehat{\boldsymbol{a}}(\boldsymbol{\beta})\} = \mathbf{0}$ for all $\boldsymbol{\beta}$, hence

$$\frac{\partial \boldsymbol{g}_{n,\boldsymbol{a}}\{\boldsymbol{\beta}, \widehat{\boldsymbol{a}}(\boldsymbol{\beta})\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}} + \frac{\partial \boldsymbol{g}_{n,\boldsymbol{a}}\{\boldsymbol{\beta}, \widehat{\boldsymbol{a}}(\boldsymbol{\beta})\}}{\partial \widehat{\boldsymbol{a}}(\boldsymbol{\beta})^{\mathrm{T}}}\frac{\partial \widehat{\boldsymbol{a}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\mathrm{T}}} = \mathbf{0}$$

for all $\boldsymbol{\beta}$, which leads to

$$\frac{\partial \widehat{\boldsymbol{a}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\mathrm{T}}} = -\left[\frac{\partial \boldsymbol{g}_{n,\boldsymbol{a}}\{\boldsymbol{\beta}, \widehat{\boldsymbol{a}}(\boldsymbol{\beta})\}}{\partial \widehat{\boldsymbol{a}}(\boldsymbol{\beta})^{\mathrm{T}}}\right]^{-1}\frac{\partial \boldsymbol{g}_{n,\boldsymbol{a}}\{\boldsymbol{\beta}, \widehat{\boldsymbol{a}}(\boldsymbol{\beta})\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}}.$$

Hence,

$$\begin{aligned}&\left\{\frac{\partial \boldsymbol{g}_{n,\boldsymbol{\beta}}\{\boldsymbol{\beta}, \widehat{\boldsymbol{a}}(\boldsymbol{\beta})\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}} + \frac{\partial \boldsymbol{g}_{n,\boldsymbol{\beta}}\{\boldsymbol{\beta}, \widehat{\boldsymbol{a}}(\boldsymbol{\beta})\}}{\partial \widehat{\boldsymbol{a}}(\boldsymbol{\beta})^{\mathrm{T}}}\frac{\partial \widehat{\boldsymbol{a}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right\}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \\ = &\left[\frac{\partial \boldsymbol{g}_{n,\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \boldsymbol{a})}{\partial \boldsymbol{\beta}_0^{\mathrm{T}}} - \frac{\partial \boldsymbol{g}_{n,\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \boldsymbol{a})}{\partial \boldsymbol{a}^{\mathrm{T}}}\left\{\frac{\partial \boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}_0, \boldsymbol{a})}{\partial \boldsymbol{a}^{\mathrm{T}}}\right\}^{-1}\frac{\partial \boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}_0, \boldsymbol{a})}{\partial \boldsymbol{\beta}_0^{\mathrm{T}}}\right]\{1 + o_p(1)\} \\ = &\left[-\boldsymbol{V}_{n,\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \boldsymbol{a}) + \boldsymbol{V}_{n,\boldsymbol{\beta}\boldsymbol{a}}(\boldsymbol{\beta}_0, \boldsymbol{a})\boldsymbol{V}_{n,\boldsymbol{a}\boldsymbol{a}}^{-1}(\boldsymbol{\beta}_0, \boldsymbol{a})\boldsymbol{V}_{n,\boldsymbol{a}\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \boldsymbol{a})\right]\{1 + o_p(1)\} \\ = &\left[-\boldsymbol{V}_{n,\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \alpha_0) + \boldsymbol{V}_{n,\boldsymbol{\beta}\boldsymbol{a}}(\boldsymbol{\beta}_0, \alpha_0)\boldsymbol{V}_{n,\boldsymbol{a}\boldsymbol{a}}^{-1}(\boldsymbol{\beta}_0, \alpha_0)\boldsymbol{V}_{n,\boldsymbol{a}\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \alpha_0)\right]\{1 + o_p(1)\} \\ = &\left[-\boldsymbol{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \alpha_0) + \boldsymbol{V}_{\boldsymbol{\beta}\boldsymbol{a}}(\boldsymbol{\beta}_0, \alpha_0)\boldsymbol{V}_{\boldsymbol{a}\boldsymbol{a}}^{-1}(\boldsymbol{\beta}_0, \alpha_0)\boldsymbol{V}_{\boldsymbol{a}\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \alpha_0)\right]\{1 + o_p(1)\}.\end{aligned}$$

Here, the first equality used the consistency of $\widehat{\boldsymbol{\beta}}$ and (S.15), the second equality used the definition in Lemma 1, the third equality used the result in (S.15) again, and the last equality follows from the law of large numbers. Now let

$$\boldsymbol{\Sigma}(\boldsymbol{b}, \alpha) = \boldsymbol{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}, \alpha) - \boldsymbol{V}_{\boldsymbol{\beta}\boldsymbol{a}}(\boldsymbol{\beta}, \alpha)\boldsymbol{V}_{\boldsymbol{a}\boldsymbol{a}}^{-1}(\boldsymbol{\beta}, \alpha)\boldsymbol{V}_{\boldsymbol{a}\boldsymbol{\beta}}(\boldsymbol{\beta}, \alpha),$$

then

$$\left\{\frac{\partial \boldsymbol{g}_{n,\boldsymbol{\beta}}\{\boldsymbol{\beta}, \widehat{\boldsymbol{a}}(\boldsymbol{\beta})\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}} + \frac{\partial \boldsymbol{g}_{n,\boldsymbol{\beta}}\{\boldsymbol{\beta}, \widehat{\boldsymbol{a}}(\boldsymbol{\beta})\}}{\partial \widehat{\boldsymbol{a}}(\boldsymbol{\beta})^{\mathrm{T}}}\frac{\partial \widehat{\boldsymbol{a}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right\}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = -\boldsymbol{\Sigma}(\boldsymbol{\beta}_0, \alpha_0)\{1 + o_p(1)\}.$$

On the other hand, under Condition (C3),

$$\begin{aligned} n^{1/2}\boldsymbol{g}_{n,\boldsymbol{\beta}}\{\boldsymbol{\beta}_0, \widehat{\boldsymbol{a}}(\boldsymbol{\beta}_0)\} &= n^{1/2}\boldsymbol{g}_{n,\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \alpha_0) + n^{1/2}[\boldsymbol{g}_{n,\boldsymbol{\beta}}\{\boldsymbol{\beta}_0, \widehat{\boldsymbol{a}}(\boldsymbol{\beta}_0)\} - \boldsymbol{g}_{n,\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \boldsymbol{a})] \\ &\quad + n^{1/2}\{\boldsymbol{g}_{n,\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \boldsymbol{a}) - \boldsymbol{g}_{n,\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \alpha_0)\} \\ &= n^{1/2}\boldsymbol{g}_{n,\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \alpha_0) + n^{1/2}[\boldsymbol{g}_{n,\boldsymbol{\beta}}\{\boldsymbol{\beta}_0, \widehat{\boldsymbol{a}}(\boldsymbol{\beta}_0)\} - \boldsymbol{g}_{n,\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \boldsymbol{a})] + o_p(1). \end{aligned}$$

Now using the Taylor expansion, (S.20), (S.15) and the law of large numbers respectively,

$$\begin{aligned} & n^{1/2}[\boldsymbol{g}_{n,\boldsymbol{\beta}}\{\boldsymbol{\beta}_0, \widehat{\boldsymbol{a}}(\boldsymbol{\beta}_0)\} - \boldsymbol{g}_{n,\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \boldsymbol{a})] \\ = & \frac{\partial \boldsymbol{g}_{n,\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \boldsymbol{a})}{\partial \boldsymbol{a}^{\mathrm{T}}}\sqrt{n}\{\widehat{\boldsymbol{a}}(\boldsymbol{\beta}_0) - \boldsymbol{a}\}\{1 + o_p(1)\} \\ = & -\boldsymbol{V}_{n,\boldsymbol{\beta}\boldsymbol{a}}(\boldsymbol{\beta}_0, \boldsymbol{a})\sqrt{n}\{\widehat{\boldsymbol{a}}(\boldsymbol{\beta}_0) - \boldsymbol{a}\}\{1 + o_p(1)\} \\ = & -\sqrt{n}\boldsymbol{V}_{n,\boldsymbol{\beta}\boldsymbol{a}}(\boldsymbol{\beta}_0, \boldsymbol{a})\boldsymbol{V}_{n,\boldsymbol{a}\boldsymbol{a}}^{-1}(\boldsymbol{\beta}_0, \boldsymbol{a})\boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}_0, \boldsymbol{a})\{1 + o_p(1)\} \\ = & -\sqrt{n}\boldsymbol{V}_{\boldsymbol{\beta}\boldsymbol{a}}(\boldsymbol{\beta}_0, \alpha_0)\boldsymbol{V}_{\boldsymbol{a}\boldsymbol{a}}^{-1}(\boldsymbol{\beta}_0, \alpha_0)\boldsymbol{g}_{n,\boldsymbol{a}}(\boldsymbol{\beta}_0, \alpha_0)\{1 + o_p(1)\}. \end{aligned}$$

This leads to the expansion

$$\begin{aligned} n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}_0, \alpha_0)n^{-1/2}\sum_{i=1}^{n}[\boldsymbol{S}_{\mathrm{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0) - \boldsymbol{V}_{\boldsymbol{\beta}\boldsymbol{a}}(\boldsymbol{\beta}_0, \alpha_0) \\ &\quad \times \boldsymbol{V}_{\boldsymbol{a}\boldsymbol{a}}^{-1}(\boldsymbol{\beta}_0, \alpha_0)\boldsymbol{S}_{\mathrm{eff},\boldsymbol{a}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)]\{1 + o_p(1)\}. \end{aligned} \tag{S.21}$$

Now using Lemma 1,

$$\begin{aligned} & \mathrm{var}\left\{\boldsymbol{S}_{\mathrm{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0) - \boldsymbol{V}_{\boldsymbol{\beta}\boldsymbol{a}}(\boldsymbol{\beta}_0, \alpha_0)\boldsymbol{V}_{\boldsymbol{a}\boldsymbol{a}}^{-1}(\boldsymbol{\beta}_0, \alpha_0)\boldsymbol{S}_{\mathrm{eff},\boldsymbol{a}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)\right\} \\ = & E\left\{\boldsymbol{S}_{\mathrm{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)^{\otimes 2}\right\} - E\left\{\boldsymbol{S}_{\mathrm{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)\boldsymbol{S}_{\mathrm{eff},\boldsymbol{a}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)^{\mathrm{T}}\right\} \\ & \times \boldsymbol{V}_{\boldsymbol{a}\boldsymbol{a}}^{-1}(\boldsymbol{\beta}_0, \alpha_0)^{\mathrm{T}}\boldsymbol{V}_{\boldsymbol{\beta}\boldsymbol{a}}(\boldsymbol{\beta}_0, \alpha_0)^{\mathrm{T}} + \boldsymbol{V}_{\boldsymbol{\beta}\boldsymbol{a}}(\boldsymbol{\beta}_0, \alpha_0)\boldsymbol{V}_{\boldsymbol{a}\boldsymbol{a}}^{-1}(\boldsymbol{\beta}_0, \alpha_0)E\left\{\boldsymbol{S}_{\mathrm{eff},\boldsymbol{a}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)^{\otimes 2}\right\} \\ & \times \boldsymbol{V}_{\boldsymbol{a}\boldsymbol{a}}^{-1}(\boldsymbol{\beta}_0, \alpha_0)^{\mathrm{T}}\boldsymbol{V}_{\boldsymbol{\beta}\boldsymbol{a}}(\boldsymbol{\beta}_0, \alpha_0)^{\mathrm{T}} - \boldsymbol{V}_{\boldsymbol{\beta}\boldsymbol{a}}(\boldsymbol{\beta}_0, \alpha_0)\boldsymbol{V}_{\boldsymbol{a}\boldsymbol{a}}^{-1}(\boldsymbol{\beta}_0, \alpha_0) \\ & \times E\left\{\boldsymbol{S}_{\mathrm{eff},\boldsymbol{a}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)\boldsymbol{S}_{\mathrm{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)^{\mathrm{T}}\right\} \\ = & \boldsymbol{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \alpha_0) - \boldsymbol{V}_{\boldsymbol{\beta}\boldsymbol{a}}(\boldsymbol{\beta}_0, \alpha_0)\boldsymbol{V}_{\boldsymbol{a}\boldsymbol{a}}^{-1}(\boldsymbol{\beta}_0, \alpha_0)^{\mathrm{T}}\boldsymbol{V}_{\boldsymbol{a}\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \alpha_0) \\ = & \boldsymbol{\Sigma}(\boldsymbol{\beta}_0, \alpha_0). \end{aligned}$$

We now inspecting the $p_1 \times p_1$ matrix $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0, \alpha_0)$. Lemma 1 implies that

$$\begin{aligned}
& \boldsymbol{\Sigma}(\boldsymbol{\beta}_0, \alpha_0) \\
= \; & E\{\boldsymbol{S}^{\otimes 2}_{\text{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)\} - E\{\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)\boldsymbol{S}_{\text{eff},\boldsymbol{a}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)^{\mathrm{T}}\} \\
& \times \left[E\{\boldsymbol{S}^{\otimes 2}_{\text{eff},\boldsymbol{a}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)\}\right]^{-1} E\{\boldsymbol{S}_{\text{eff},\boldsymbol{a}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)^{\mathrm{T}}\} \\
= \; & E\left([\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)\} - \Pi\{\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0)\} \mid \mathcal{S}_{\boldsymbol{a}}]^{\otimes 2}\right).
\end{aligned}$$

Here, we use $\mathcal{S}_{\boldsymbol{a}}$ to denote the p_1-component functional space spanned by $\boldsymbol{S}_{\text{eff},\boldsymbol{a}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \alpha_0)$. A closer inspection of $\mathcal{S}_{\boldsymbol{a}}$ reveals that

$$\mathcal{S}_{\boldsymbol{a}} = [\boldsymbol{C}\{\boldsymbol{B}(X_{i1}), \ldots, \boldsymbol{B}(X_{im_i})\}\boldsymbol{J}_i \text{diag}\{V_{ij} - E(V_{ij} \mid W_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0), j = 2, \ldots, m_i\}\mathbf{1}_{m_i - 1}],$$

where $\boldsymbol{C}$ is any $p_1 \times p_2$ matrix. For any $j = 1 \ldots, m_i$ and any function $f_j(X_{ij}) \in C^q(X_{ij})$, (S.15) ensures that when N is large enough, we can find a p_2 (recall $p_2 = N + r$) dimensional vector $\boldsymbol{c}_j$ so that $\boldsymbol{c}_j^{\mathrm{T}}\boldsymbol{B}(X_{ij})$ approaches $f_j(X_{ij})$ uniformly. Thus, $\boldsymbol{C} \equiv (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{q_1})^{\mathrm{T}}\boldsymbol{B}(X_{ij})$ can approach $\boldsymbol{f}(X_{ij}) \equiv \{f_1(X_{ij}), \ldots, f_{q_1}(X_{ij})\}^{\mathrm{T}}$ uniformly. This means $\sup |\boldsymbol{C}\{\boldsymbol{B}(X_{i1}), \ldots, \boldsymbol{B}(X_{im_i})\} - \{\boldsymbol{f}(X_{i1}), \ldots, \boldsymbol{f}(X_{im})\}| \to 0$, where the sup is with respect to all elements in the $p_1 \times m_i$ matrix and all X_{ij}'s in $\alpha_0(x)$'s support space $[a, b]$ for $j = 1, \ldots, m_i$. Comparing the definition of $\mathcal{S}_\alpha$ and $\mathcal{S}_{\boldsymbol{a}}$, this directly leads to $\sup |\Pi\{\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0) \mid \mathcal{S}_{\boldsymbol{a}}\} - \Pi\{\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}_0, \alpha_0) \mid \mathcal{S}_\alpha\}| \to 0$, where the sup is with respect to all p_1 functional components and all X_{ij} values. Hence $\|\boldsymbol{\Sigma}_0(\boldsymbol{\beta}_0, \alpha_0) - \boldsymbol{\Sigma}(\boldsymbol{\beta}_0, \alpha_0)\|_F \to 0$ when $n \to \infty$, where $\|\cdot\|_F$ is the Fronebius norm. Thus, from (S.21) and the Central Limit Theorem for iid data, we have obtained $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to \text{Normal}(\mathbf{0}_{p_1}, \boldsymbol{\Sigma}_0^{-1})$ in distribution. □

APPENDIX S.2: ADDITIONAL SIMULATION RESULTS

We generated data as in Section 3.1 except for clustered failure times with multiple event types (Example 2). Data is generated using the model in (2.5) with $n = 500$ and $m_i = 2$ at 40% censoring, and with $n = 1300$ and $m_i = 2$ at 70% censoring similar to the HD application setting. All remaining aspects of the simulation design are as in Section 3.1, so our focus is on the estimates of $\alpha(x, t), \beta(t)$ in (2.5) which represent the logit differences in covariate effects.

As in Section 3.2, GAMM estimates, though unbiased across non-normally distributed random effects, are largely biased when the random effects and covariates are not independent (Tables S.2, S.3). The biasedness of $\widehat{\alpha}(x,t)$ and $\widehat{\beta}(t)$ lead to mean squared errors (MSE) that are consistently larger than the MSE from the proposed method in all settings. This further exemplifies the sensitivity of GAMM to violations of the assumption that the covariates and random effect are independent.

In comparison, our proposed method shows negligible bias for $\widehat{\beta}(t), \widehat{\alpha}(x,t)$ averaged across x- and t-values (Table S.2) and in pointwise estimates (Table S.3). The unbiasedness is visually evident across all simulations (Figures S.1 and S.2), where estimates from our proposed method generally overlap the true underlying curves of $\alpha(x,t)$ and $\beta(t)$. At 70% censoring, our method and GAMM have more difficulty in unbiasedly estimating the true curve. However, our method better captures the truth than does GAMM particularly when the covariates and random effect are dependent (compare the left and right panel of Figure S.2).

APPENDIX S.3: ADDITIONAL RESULTS FOR HUNTINGTON'S DISEASE APPLICATION

Table S.5 provides the proportion of pseudo-value estimates $\widetilde{Y}$ that fell outside the $[0,1]$ range for the Huntington's disease application at $t = 35, 40, \ldots, 60$. The proportions are similar to those observed for the simulation study (Table S.1).

TABLE S.1

Proportion of jack-knife estimates $\widetilde{Y}(t)$ that fall outside [0,1] at $t = 40, 46, 49$. Results shown for simulation study described in Section 3: clustered failure times with single event types and multiple event types with different true random intercept as specified. 40% censoring. Results averaged over 1000 simulations.

	X, Z, R independent		(X, Z) and R dependent	
	$\widetilde{Y}(t) < 0$	$\widetilde{Y}(t) > 1$	$\widetilde{Y}(t) < 0$	$\widetilde{Y}(t) > 1$
Single event type	$R \sim \text{Normal}(0, 1)$			
$t = 40$	0.001	0	0.001	0
$t = 46$	0.009	0	0.009	0
$t = 49$	0.014	0	0.014	0
	$R \sim 0.5\text{Normal}(-1, 1) + 0.5\text{Normal}(1, 0.25^2)$			
$t = 40$	0.001	0	0.001	0
$t = 46$	0.010	0	0.010	0
$t = 49$	0.014	0	0.014	0
	$R \sim 0.5\text{Normal}(-1, 1) + 0.5\text{Beta}(4, 2)$			
$t = 40$	0.001	0	0.001	0
$t = 46$	0.008	0	0.008	0
$t = 49$	0.014	0	0.014	0
	$R \sim \text{Uniform}[-2.5, 2.5]$			
$t = 40$	0.001	0	0.001	0
$t = 46$	0.009	0	0.009	0
$t = 49$	0.015	0	0.015	0
Multiple event types	$R \sim \text{Normal}(0, 1)$			
$t = 40$	0.010	0	0.010	0
$t = 46$	0.041	0	0.043	0
$t = 49$	0.058	0	0.060	0
	$R \sim 0.5\text{Normal}(-1, 1) + 0.5\text{Normal}(1, 0.25^2)$			
$t = 40$	0.011	0	0.011	0
$t = 46$	0.046	0	0.047	0
$t = 49$	0.064	0	0.065	0
	$R \sim 0.5\text{Normal}(-1, 1) + 0.5\text{Beta}(4, 2)$			
$t = 40$	0.010	0	0.010	0
$t = 46$	0.041	0	0.041	0
$t = 49$	0.057	0	0.057	0
	$R \sim \text{Uniform}[-2.5, 2.5]$			
$t = 40$	0.010	0	0.010	0
$t = 46$	0.043	0	0.042	0
$t = 49$	0.061	0	0.060	0

TABLE S.2

Average results for clustered failure times with multiple event types. 70% censoring, 1000 simulations. Average absolute bias, empirical variance, 95% coverage probabilities and mean squared errors (MSE) when the true random intercept is as specified. $\widehat{\beta}(\cdot)$ *denotes results averaged over* t*;* $\widehat{\alpha}(0.50, \cdot)$ *is results at* $x = 0.50$ *averaged over* t*, and* $\widehat{\alpha}(\cdot, 46)$ *is results at* $t = 46$ *averaged over* x*.*

	Proposed Method			GAMM Method		
	$\widehat{\beta}(\cdot)$	$\widehat{\alpha}(0.50, \cdot)$	$\widehat{\alpha}(\cdot, 46)$	$\widehat{\beta}(\cdot)$	$\widehat{\alpha}(0.50, \cdot)$	$\widehat{\alpha}(\cdot, 46)$
X, Z, R independent						
			$R \sim \text{Normal}(0, 1)$			
abs bias	0.016	0.085	0.111	0.029	0.039	0.007
emp var	0.188	0.594	0.540	0.042	0.124	0.117
95% cov	0.949	0.950	0.946	0.947	0.945	0.947
MSE	0.188	0.603	0.554	0.043	0.126	0.118
			$R \sim 0.5\text{Normal}(-1, 1) + 0.5\text{Normal}(1, 0.25^2)$			
abs bias	0.019	0.098	0.103	0.005	0.062	0.012
emp var	0.212	0.653	0.622	0.050	0.150	0.137
95% cov	0.946	0.950	0.948	0.952	0.946	0.950
MSE	0.212	0.664	0.633	0.050	0.156	0.138
			$R \sim 0.5\text{Normal}(-1, 1) + 0.5\text{Beta}(4, 2)$			
abs bias	0.022	0.093	0.088	0.036	0.067	0.044
emp var	0.186	0.615	0.546	0.042	0.131	0.120
95% cov	0.948	0.946	0.946	0.947	0.942	0.947
MSE	0.186	0.625	0.554	0.043	0.137	0.123
			$R \sim \text{Uniform}[-2.5, 2.5]$			
abs bias	0.025	0.091	0.107	0.022	0.059	0.040
emp var	0.235	0.787	0.684	0.056	0.156	0.149
95% cov	0.944	0.943	0.946	0.946	0.948	0.950
MSE	0.235	0.797	0.696	0.057	0.161	0.151
(X, Z) and R dependent						
			$R \sim \text{Normal}(0, 1)$			
abs bias	0.036	0.073	0.086	1.077	1.432	1.509
emp var	0.477	1.300	1.139	0.155	0.344	0.355
95% cov	0.958	0.955	0.955	0.224	0.326	0.256
MSE	0.478	1.305	1.147	1.316	2.403	2.794
			$R \sim 0.5\text{Normal}(-1, 1) + 0.5\text{Normal}(1, 0.25^2)$			
abs bias	0.036	0.135	0.122	1.679	2.091	2.176
emp var	0.539	1.231	1.119	0.230	0.430	0.430
95% cov	0.950	0.951	0.952	0.041	0.091	0.058
MSE	0.541	1.249	1.135	3.049	4.806	5.500
			$R \sim 0.5\text{Normal}(-1, 1) + 0.5\text{Beta}(4, 2)$			
abs bias	0.015	0.089	0.096	0.966	1.335	1.453
emp var	0.408	1.318	1.144	0.124	0.314	0.322
95% cov	0.951	0.949	0.952	0.233	0.354	0.245
MSE	0.409	1.328	1.154	1.066	2.126	2.581
			$R \sim \text{Uniform}[-2.5, 2.5]$			
abs bias	0.036	0.176	0.159	1.650	2.211	2.311
emp var	0.450	1.272	1.106	0.224	0.490	0.510
95% cov	0.950	0.949	0.945	0.046	0.106	0.067
MSE	0.452	1.304	1.132	2.952	5.397	6.229

TABLE S.3

Pointwise results for clustered failure times with multiple event types. 70% censoring, 1000 simulations. Pointwise bias, empirical variance, estimated variance, 95% coverage probabilities and mean squared error (MSE) for $\widehat{\beta}(t)$ and $\widehat{\alpha}(x,t)$ at $x = 0.50$ and $t = 46$ when the true random intercept is as specified.

	X, Z, R independent				(X, Z) and R dependent			
	Proposed Method		GAMM Method		Proposed Method		GAMM Method	
	$\widehat{\beta}(46)$	$\widehat{\alpha}(0.50, 46)$	$\widehat{\beta}(46)$	$\widehat{\alpha}(0.50, 46)$	$\widehat{\beta}(46)$	$\widehat{\alpha}(0.50, 46)$	$\widehat{\beta}(46)$	$\widehat{\alpha}(0.50, 46)$
				$R \sim \text{Normal}(0, 1)$				
bias	0.021	0.075	0.028	0.013	0.058	0.026	1.112	-1.485
emp var	0.181	0.564	0.039	0.114	0.553	1.470	0.153	0.343
est var	0.189	0.600	0.039	0.113	0.462	1.244	0.099	0.231
95% cov	0.949	0.963	0.945	0.952	0.927	0.933	0.083	0.162
MSE	0.189	0.605	0.040	0.114	0.465	1.245	1.336	2.438
				$R \sim 0.5\text{Normal}(-1, 1) + 0.5\text{Normal}(1, 0.25^2)$				
bias	0.022	0.085	0.003	0.034	0.037	0.074	1.710	-2.141
emp var	0.214	0.642	0.045	0.134	0.549	1.212	0.223	0.412
est var	0.212	0.674	0.042	0.122	0.483	1.108	0.110	0.224
95% cov	0.941	0.961	0.939	0.946	0.939	0.937	0.006	0.020
MSE	0.213	0.682	0.042	0.123	0.484	1.114	3.034	4.808
				$R \sim 0.5\text{Normal}(-1, 1) + 0.5\text{Beta}(4, 2)$				
bias	0.019	0.113	0.039	0.064	0.020	0.096	1.023	-1.416
emp var	0.189	0.613	0.039	0.120	0.445	1.416	0.120	0.301
est var	0.194	0.616	0.039	0.113	0.379	1.236	0.082	0.226
95% cov	0.958	0.939	0.940	0.936	0.938	0.932	0.071	0.195
MSE	0.194	0.629	0.041	0.117	0.379	1.246	1.129	2.232
				$R \sim \text{Uniform}[-2.5, 2.5]$				
bias	0.012	0.111	0.025	0.038	0.017	0.112	1.707	-2.290
emp var	0.225	0.732	0.050	0.143	0.490	1.361	0.224	0.488
est var	0.228	0.732	0.043	0.125	0.411	1.178	0.096	0.234
95% cov	0.947	0.938	0.928	0.931	0.926	0.927	0.002	0.012
MSE	0.228	0.744	0.044	0.127	0.411	1.191	3.011	5.480

TABLE S.4

Average results for clustered failure times with single event types and multiple event types. 0% censoring, 1000 simulations. Average absolute bias, empirical variance, 95% coverage probabilities and mean squared errors (MSE) when the true random intercept is Normal(0,1). $\widehat{\beta}(\cdot)$ *denotes results averaged over* t*;* $\widehat{\alpha}(0.50,\cdot)$ *is results at* $x = 0.50$ *averaged over* t*, and* $\widehat{\alpha}(\cdot,46)$ *is results at* $t = 46$ *averaged over* x*. Results show increased variability for our proposed method compared to GAMM is due to our relaxed assumption about the random effect distribution, and not due to pseudo-values being outside* $[0,1]$*.*

	Proposed Method			GAMM Method		
	$\widehat{\beta}(\cdot)$	$\widehat{\alpha}(0.50,\cdot)$	$\widehat{\alpha}(\cdot,46)$	$\widehat{\beta}(\cdot)$	$\widehat{\alpha}(0.50,\cdot)$	$\widehat{\alpha}(\cdot,46)$
Single event type						
			X, Z, R independent			
abs bias	0.016	0.054	0.042	0.034	0.037	0.021
emp var	0.068	0.278	0.245	0.012	0.036	0.036
est var	0.068	0.278	0.245	0.012	0.036	0.036
95% cov	0.947	0.945	0.942	0.932	0.948	0.951
MSE	0.068	0.281	0.247	0.014	0.038	0.037
			(X, Z) and R dependent			
abs bias	0.035	0.021	0.031	1.390	1.838	1.881
emp var	0.954	1.211	1.098	0.120	0.247	0.258
est var	0.954	1.211	1.098	0.120	0.247	0.258
95% cov	0.948	0.947	0.945	0.005	0.019	0.015
MSE	0.956	1.212	1.099	2.052	3.631	4.048
Multiple event type						
			X, Z, R independent			
abs bias	0.015	0.018	0.010	0.037	0.074	0.033
emp var	0.185	0.597	0.518	0.042	0.127	0.120
est var	0.185	0.597	0.518	0.042	0.127	0.120
95% cov	0.951	0.949	0.951	0.944	0.941	0.945
MSE	0.185	0.598	0.518	0.044	0.134	0.121
			(X, Z) and R dependent			
abs bias	0.062	0.071	0.059	1.250	1.606	1.710
emp var	0.505	1.335	1.167	0.169	0.366	0.376
est var	0.505	1.335	1.167	0.169	0.366	0.376
95% cov	0.953	0.951	0.950	0.133	0.254	0.180
MSE	0.509	1.340	1.170	1.733	2.961	3.508

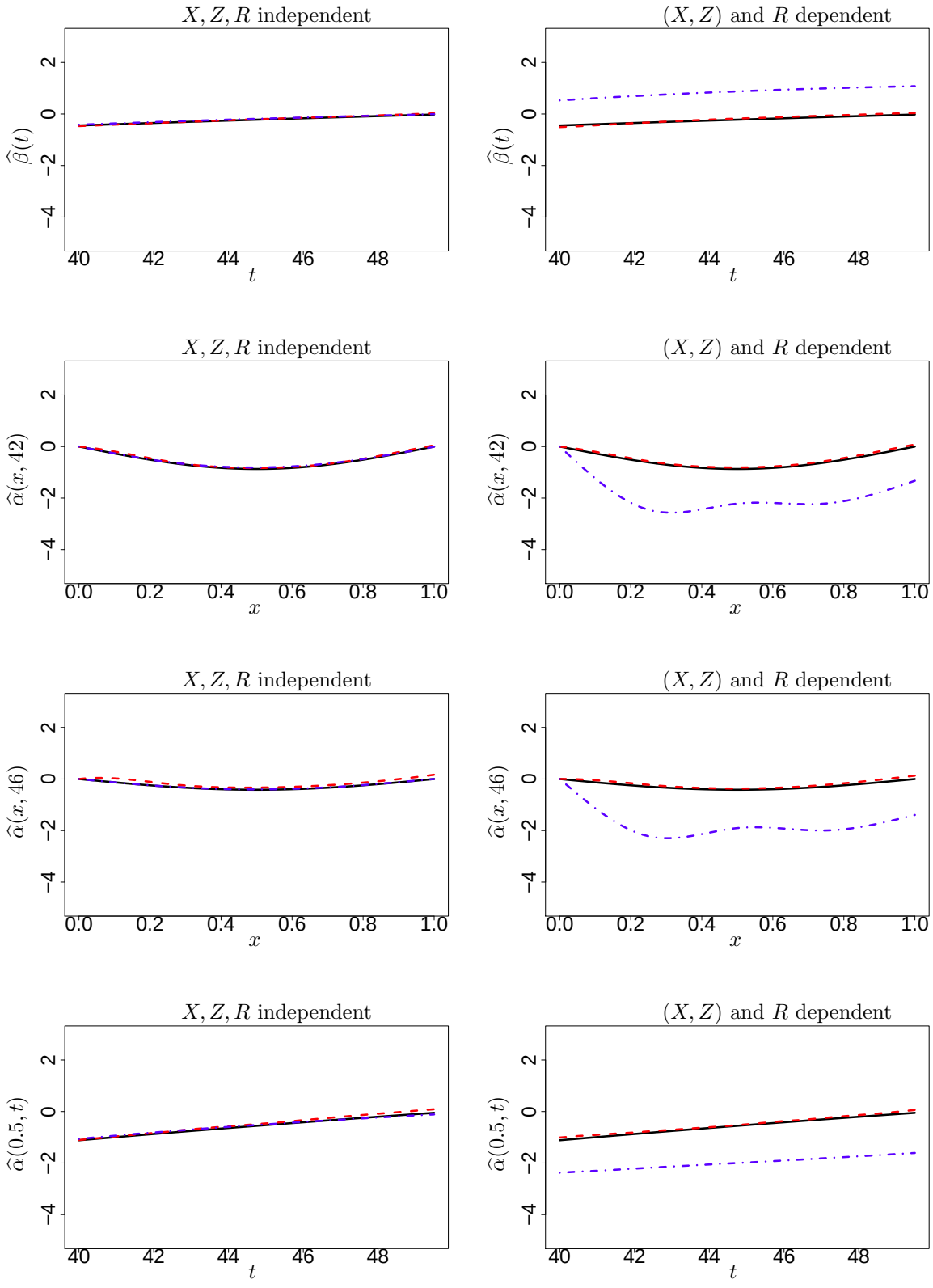

FIG S.1. *Clustered failure times with multiple event types and* $R \sim$ *Normal*$(0, 1)$*; 40% censoring, 1000 simulations. True parameter functions (black solid curve), mean of 1000 simulation estimates from our proposed method (red dashed line) and from GAMM (blue dashed-dotted line).*

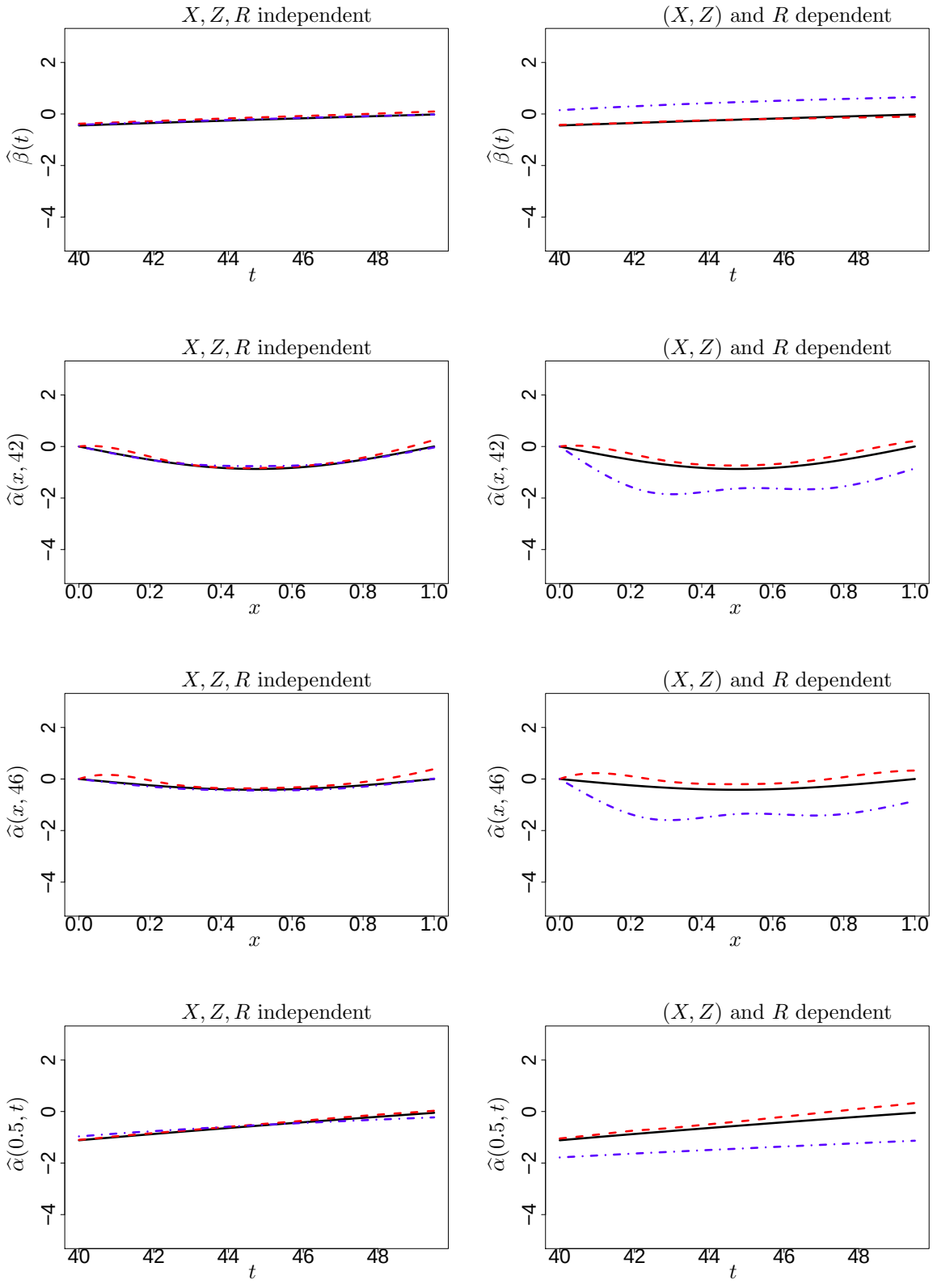

FIG S.2. *Clustered failure times with multiple event types and* $R \sim$ *Normal*$(0, 1)$*; 70% censoring, 1000 simulations. True parameter functions (black solid curve), mean of 1000 simulation estimates from our proposed method (red dashed line) and from GAMM (blue dashed-dotted line).*

TABLE S.5
Proportion of jack-knife estimates $\widetilde{Y}(t)$ *that fall outside* $[0,1]$ *at* $t = 35, 40, \ldots, 60$ *for the Huntington's disease application.*

	$\widetilde{Y}(t) < 0$	$\widetilde{Y}(t) > 1$
$t = 35$	0.000	0
$t = 40$	0.000	0
$t = 45$	0.000	0
$t = 50$	0.028	0
$t = 55$	0.044	0
$t = 60$	0.039	0

REFERENCES

Bang, H. and Tsiatis, A. A. (2000). Estimating Medical Costs with Censored Data. *Biometrika*, **87**, 329-343.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: The Johns Hopkins University Press.

Bosq, D. (1998). Nonparametric statistics for stochastic processes: estimation and prediction. Edited by P. Bickel, P. Diggle, S. Fienberg, K. Krickeber, I. Olkin, N. Wermuth and S. Zeger. New York: Springer-Verlag.

DeVore, R. A. and Lorentz, G. G. (1993). Constructive Approximation. Berlin: Springer-Verlag.

Jiang, F., Ma, Y. and Wang, Y. (2015). Fused kernel-spline smoothing for repeatedly measured outcomes in a generalized partially linear model with functional single index. *Annals of Statistics*, in press.

Logan, B., Zhang, M. and Klein, J. (2011). Marginal models for clustered time to event data with competing risks using pseudo-values. *Biometrics*, **67**, 1-7.

Ma, S., Ma, Y., Wang, Y. and Carroll, R. J. (2015). A Semiparametric Single-Index Risk Score Across Populations. *Preprint.*

Robins, J. and Rotnitzky, A. (1992). Recovery of Information and Adjustment for Dependent Censoring using Surrogate Markers. *AIDS Epidemiology*; Ed. N. Jewell, K. Dietz and V. Farewell, pp. 297-331. Birkhäuser, Boston.

Scheike, T. H., Zhang, M-J and Gerds, T. A. (2008). Predicting cumulative incidence probability by direct binomial regression. *Biometrika*, **95**,205-220.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

TANYA P. GARCIA
DEPARTMENT OF EPIDEMIOLOGY AND BIOSTATISTICS
TEXAS A&M UNIVERSITY
TAMU 1266
COLLEGE STATION, TX 77845
E-MAIL: tpgarcia@sph.tamhsc.edu

KAREN MARDER
DEPARTMENT OF NEUROLOGY AND PSYCHIATRY
SERGIEVSKY CENTER AND TAUB INSTITUTE
COLUMBIA UNIVERSITY MEDICAL CENTER
630 WEST 168TH STREET
NEW YORK, NY 10032
E-MAIL: ksm1@columbia.edu

YANYUAN MA
DEPARTMENT OF STATISTICS
PENNSYLVANIA STATE UNIVERSITY
UNIVERSITY PARK, PA 16802
E-MAIL: yzm63@psu.edu

YUANJIA WANG
DEPARTMENT OF BIOSTATISTICS
MAILMAN SCHOOL OF PUBLIC HEALTH
COLUMBIA UNIVERSITY
NEW YORK, NY 10032
E-MAIL: yuanjia.wang@columbia.edu