Predicting Cumulative Risk of Disease Onset by Re-distributing Weights

Tianle Chen
Biogen Idec, Cambridge, MA
email: tianle.chen@biogenidec.com

Yanyuan Ma
Department of Statistics, Texas A&M University
email: ma@stat.tamu.edu

Yuanjia Wang [1]
Department of Biostatistics, Mailman School of Public Health, Columbia University
email: yuanjia.wang@columbia.edu

Abstract

We propose a simple approach predicting the cumulative risk of disease accommodating predictors with time-varying effects and outcomes subject to censoring. We use a nonparametric function for the coefficient of the time-varying effect and handle censoring through self-consistency equations that redistribute the probability mass of censored outcomes to the right. The computational procedure is extremely convenient and can be implemented by standard software. We prove large sample properties of the proposed estimator and evaluate its finite sample performance through simulation studies. We apply the method to estimate the cumulative risk of developing Huntington's disease (HD) from subjects with huntingtin gene mutation using a large collaborative HD study data and illustrate an inverse relationship between the cumulative risk of HD and the length of cytosine-adenine-guanine (CAG) repeats in the huntingtin gene.

Some Key Words:

Proportional odds model; Self-consistency equation; Varying-coefficient model; Huntington's disease

[1] Correspondence author

1 Introduction

In many biomedical studies, the research goal is to predict the age-specific cumulative risk of onset of a disease from a set of covariates. For example, Huntington's disease (HD) is a progressive neurodegenerative disorder caused by the expansion of cytosine-adenine-guanine (CAG) trinucleotide repeats in the huntingtin gene [1]. The genetic model for HD is dominant [2], and there is an inverse relationship between the age-at-onset of HD and the CAG repeats length: the greater the CAG expansion, the earlier the age-at-onset of the disease. Accurate prediction of a subject's age-at-onset of HD from CAG repeats and other covariates is useful to assess an individual's risk of developing HD based on available genetic mutation testing results when providing genetic counseling. Such estimates are also useful when designing a clinical trial. For instance, estimating the age-at-onset distribution of HD from a subject's CAG repeats and other baseline information can be used to recruit patients who are close to the onset of disease to improve efficiency of a therapeutic trial. Improving existing estimation of HD risk is one of the research goals in the Cooperative Huntington's Observational Research Trial (COHORT) study which includes 42 sites [3, 4].

Age-at-onset of disease information is usually subject to right censoring due to termination of study, patient loss to follow up, or death of a subject. Popular regression models for censored outcomes include Cox-proportional hazards model [5], which relates hazard rate at a particular age to covariates. Although it is possible to obtain cumulative risk function in this model, the interpretation of the regression coefficients of the covariates is reflected through the hazard function. Instead of working through a hazard function, a more appealing approach is to model the cumulative disease risk directly since predicting disease onset is our primary goal and hazard function is not of interest. Another motivation to avoid proportional hazards model and to work with cumulative distribution function directly is that the proportional hazards assumption may not be satisfied in certain applications. For example, Langbehn et al. [6] reported that the proportional hazards assumption does not hold with HD data, and proposed a parametric model for the cumulative risk function involving six parameters through a logistic transformation of $\mathrm{pr}(T_i \leq t|X_i)$, where T_i denotes age-at-onset of HD, and X_i denotes the covariate of interest (i.e., CAG repeats length). Specifically, their model is

$$\mathrm{logit}\{\mathrm{pr}(T_i \leq t|X_i)\} = \{t - \mu(X_i;\alpha)\}/s(X_i;\gamma),$$

where $\mu(X_i;\alpha)$ and $s(X_i;\gamma)$ are exponential functions with parameters α and γ. In Langbehn et al. [6] these functions are $\mu(x;\alpha) = \alpha_1 + \exp(\alpha_2 - \alpha_3 x)$ and $s(x;\gamma) = \sqrt{\gamma_1 + \exp(\gamma_2 - \gamma_3 x)}$. The correlation between the estimated parameters may be high, causing numerical difficulties in practice. We fitted Langbehn's model on COHORT data and the results were reported in an earlier paper [7]. Other ad-hoc parametric models are also proposed for HD data, with no consensus on the best model to use [8].

In this work, we consider a varying-coefficient proportional odds model

$$\mathrm{logit}\{\mathrm{pr}(T_i \leq t|X_i)\} = \beta_0(t) + \beta_1(t)X_i. \tag{1}$$

To provide flexibility and protect against misspecification, $\beta_0(t)$ and $\beta_1(t)$ are left as unknown nonparametric functions. The interpretation of $\beta_1(t)$ is then directly related to the cumulative risk of disease, since $\exp\{\beta_1(t)\}$ is the odds ratio of experiencing disease onset by age t for subjects with one unit difference in X. Since $\mathrm{pr}(T_i \leq t|X_i)$ is a cumulative distribution function, $\beta_0(t)$ and $\beta_0(t) + \beta_1(t)X_i$ are constrained to be non-decreasing functions of t. In applications where X_i's are positive, we require $\beta_1(t)$ to be non-decreasing as well. When $\beta_0(t)$ and $\beta_1(t)$ take some parametric form of t, model (1) reduces to a standard proportional odds model.

An extension to model (1) is a nonparametric varying-coefficient model of the cumulative risk using a logistic link

$$\mathrm{logit}\{\mathrm{pr}(T_i \leq t|X_i)\} = \beta_0(t) + c_0(X_i) + \beta_1(t)c_1(X_i), \tag{2}$$

where $c_0(x)$ and $c_1(x)$ are known parametric functions of covariates. Note that when $c_1(x) = 1/s(x;\gamma)$, $c_0(x) = -\mu(x;\alpha)/s(x;\gamma)$, $\beta_0(t) = 0$, and $\beta_1(t) = t$, model (2) reduces to that in Langbehn et al. [6].

In the literature, Jung [9] directly modeled survival function using regression model at a fixed time point without considering temporal effect. There are a number of other works on extending proportional hazards or proportional odds model to account for temporal covariate effect or time-varying covariates. Peng and Huang [10] proposed an alternative extension of Cox proportional hazards model to account for a nonparametric temporal effect of a covariate. The procedure involves solving a series of estimating equations sequentially. In contrast, our method is proposed for a proportional odds model with a nonparametric time-varying effect. Chen et al. [11] proposed methods to extend transformation models considered in for example, Zeng and Lin [12], to account for external time-varying covariates.

Here, we take a completely different approach that does not involve counting process and with straightforward and simple computational algorithm. When there is no censoring, to estimate the cumulative risk function at a time point t_0 given a covariate, e.g., $\mathrm{pr}(T_i \leq t_0|X_i)$, a straightforward analysis is to fit a logistic regression of $I(T_i \leq t_0)$ on the covariates X_i. When the outcome is subject to censoring, $I(T_i \leq t)$ may not be observed for some of the censored subjects. Let C_i denote the censoring time, Efron [13] proposed a nonparametric estimator of a survival function by re-distributing the conditional masses for the censored subjects, $\mathrm{pr}(T_i > C_i|C_i)$, equally to all the non-censored observations above C_i, where the common weight for these subjects depends on the number of at-risk subjects at C_i. Portnoy [14] and Wang and Wang [15] used similar ideas to fit a quantile regression with covariates X_i, where the conditional point masses $\mathrm{pr}(T_i > C_i|C_i, X_i)$ for censored subjects are re-distributed to the right. For quantile regression, the estimator only depends on the signs of residuals and thus the point masses for censored subjects are re-distributed to $+\infty$. Since there are covariates involved, the conditional masses to be estimated depend on the covariates and the unknown distribution function.

In this work, to estimate $\boldsymbol{\beta}(t_0)$ from (1) or (2), we fit a pseudo-logistic regression of $I(T_i \leq t_0)$ through redistributing weights to the right to account for censoring. We apply the procedure to estimate the coefficient function at distinct uncensored event times, and smooth the coefficient functions across the entire support of event times when necessary. This type of smoothing was found to be equivalent to applying local kernel smoothing directly [16]. The proposed computational procedure is extremely easy to implement and can be handled by standard softwares. We investigate the asymptotic properties of the proposed estimator to show consistency and normality, and conduct simulation studies to examine its finite sample performance. The proposed methods are applied to estimating the cumulative risk of developing HD from subjects with huntingtin gene mutation using the COHORT data and illustrate an positive relationship between the cumulative risk of HD and the length of CAG repeats in the huntingtin gene. We compare the estimates under model (1) with fully nonparametric Kaplan-Meier estimates using subjects with the same CAG values and reveal consistent results.

2 Methods

For the purpose of illustration, we mainly focus on the varying coefficient model (1). Extension to the more general model (2) is discussed in Section 4.

2.1 Uncensored data

First we investigate estimation at a fixed time point t_0 when the outcome is not subject to censoring. Let $\boldsymbol{\beta}(t) = \{\beta_0(t), \beta_1(t)\}^T$, let $\boldsymbol{\beta}_{t_0} = \boldsymbol{\beta}(t_0)$ denote $\boldsymbol{\beta}(\cdot)$ evaluated at t_0, and let $Z_i = (1, X_i)^T$. When there is no censoring, the likelihood for $\{I(T_i \leq t_0), X_i, i = 1, \cdots, n\}$ under a logistic link takes the standard form, $\prod_i \frac{\exp\{I(T_i \leq t_0) Z_i^T \boldsymbol{\beta}_{t_0}\}}{1 + \exp\{Z_i^T \boldsymbol{\beta}_{t_0}\}}$. To estimate $\boldsymbol{\beta}_{t_0}$, we solve the estimating equation

$$\sum_{i=1}^{n} m(X_i, T_i; t_0, \boldsymbol{\beta}_{t_0}) = 0,$$

where $m(X_i, T_i; t_0, \boldsymbol{\beta}_{t_0}) = \left\{I(T_i \leq t_0) - \mu(X_i; \boldsymbol{\beta}_{t_0})\right\} Z_i$, and $\mu\{X_i; \boldsymbol{\beta}_{t_0}\} = \dfrac{\exp\{Z_i^T \boldsymbol{\beta}_{t_0}\}}{1 + \exp\{Z_i^T \boldsymbol{\beta}_{t_0}\}}$. The influence function for the estimate $\widehat{\boldsymbol{\beta}}_{t_0}$ is

$$\phi(X_i, T_i; t_0, \boldsymbol{\beta}_{t_0}) = A(X_i; \boldsymbol{\beta}_{t_0}) \left\{I(T_i \leq t_0) - \mu(X_i; \boldsymbol{\beta}_{t_0})\right\} Z_i,$$

where $A(X_i; \boldsymbol{\beta}_{t_0}) = \Big(E[\mu(X_i; \boldsymbol{\beta}_{t_0})\{1 - \mu(X_i; \boldsymbol{\beta}_{t_0})\} Z_i Z_i^T]\Big)^{-1}$. We fit a logistic regression of $I(T_i \leq t_0)$ on X_i and repeat this process while varying t_0 at all distinct values of observed T_i's. One can then smooth the estimates as a function of t_0 [16] subject to the monotonicity constraint. An alternative is to fit a nonparametric regression (for example using splines) treating $I(T_i \leq t)$ as generalized outcomes. This method was shown to have similar performance as the post-hoc smoothing above [16], but is more difficult to implement under the monotonicity constraint, therefore we do not further explore here.

2.2 Censored data

When a subject is right censored (i.e., $T_i > C_i$) and $C_i \geq t_0$, we still observe $I(T_i \leq t_0) = 0$. Ambiguity occurs when a subject is censored and $C_i < t_0$. One type of estimator for $\boldsymbol{\beta}(\cdot)$ can be obtained by the inverse probability of censoring weighting (IPW) proposed in Bang and Tsiatis [17], which weights subjects having an event by the inverse of their probabilities of not being censored. To be specific, we can obtain the IPW estimator by solving the estimating equation

$$S_n(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} \frac{I(T_i \leq C_i) m(X_i, T_i; t_0, \boldsymbol{\beta})}{G(T_i)} = 0,$$

where $G(\cdot)$ is the survival function for the censoring times C_i. Estimating $G(\cdot)$ by the Kaplan-Meier of the censoring process, the estimating equation for $\boldsymbol{\beta}(\cdot)$ is

$$\sum_{i=1}^{n} \frac{I(T_i \leq C_i) m(X_i, T_i; t_0, \boldsymbol{\beta})}{\widehat{G}(T_i)} = 0. \tag{3}$$

This process is repeated for t_0 on a grid $(u_1, \cdots, u_M)$. Alternatively, one can let the grid points include only uncensored observations, which is equivalent to creating the grid.

Here we propose a new type of estimator that re-distributes weights to the right for ambiguous subjects based on self-consistency equations similar to Efron [13] and Wang and Wang [15]. Let $O_i = \{X_i, T_i \wedge C_i, \Delta_i \equiv I(T_i \leq C_i)\}$ denote the ith observation. We solve the following weighted estimating equation

$$S_n(\boldsymbol{\beta}_{t_0}, \boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} s(O_i; t_0, \boldsymbol{\beta}_{t_0}, \boldsymbol{\beta}) = 0, \tag{4}$$

where

$$s(O_i; t_0, \boldsymbol{\beta}_{t_0}, \boldsymbol{\beta}) \quad = \quad w\{O_i; t_0, \boldsymbol{\beta}(\cdot)\} m(X_i, T_i \wedge C_i; t_0, \boldsymbol{\beta}_{t_0}) + [1 - w\{O_i; t_0, \boldsymbol{\beta}(\cdot)\}] m(X_i, +\infty; t_0, \boldsymbol{\beta}_{t_0}),$$

and

$$w\{O_i; t_0, , \boldsymbol{\beta}(\cdot)\} = \begin{cases} 1, & \Delta_i = 1 \text{ or } (\Delta_i = 0 \text{ and } C_i \geq t_0) \\ \dfrac{F(t_0|X_i) - F(C_i|X_i)}{1 - F(C_i|X_i)}, & \Delta_i = 0 \text{ and } C_i < t_0. \end{cases} \tag{5}$$

Here $F(t|x) = \mu\{x; \boldsymbol{\beta}(t)\}$ is the conditional distribution of T_i given X_i introduced in model (1), and the weight for the ith subject depends on $\boldsymbol{\beta}(\cdot)$ evaluated at t_0 and C_i.

To gain insights on the weights, note that subjects with observed $I(T_i \leq t_0)$ will receive a weight of one for their contributions to the estimating equation. Subjects with missing $I(T_i \leq t_0)$ have conditional probability masses

$$E\{I(T_i \leq t_0) | T_i > C_i, X_i\} \quad = \quad \frac{F(t_0|X_i) - F(C_i|X_i)}{1 - F(C_i|X_i)}.$$

Treating (X_i, C_i) as pseudo-observations for censored subjects with censoring time less than t_0, they receive weights $w\{O_i; t_0, \boldsymbol{\beta}(\cdot)\} = \text{pr}(T_i \le t_0 | T_i > C_i, X_i)$. We re-distribute their complementary weights $1 - w\{O_i; t_0, \boldsymbol{\beta}(\cdot)\} = \text{pr}(T_i > t_0 | T_i > C_i, C_i, X_i)$ to the right. Since the outcomes are binary variables, the complementary masses $1 - w\{O_i; t_0, \boldsymbol{\beta}(\cdot)\}$ for pseudo-observations (X_i, C_i) can be re-distributed to any point that is greater than all observations that is not specific to any observation above C_i (also see Portnoy [14] and Wang and Wang [15]). Thus, any point above C_i contributes the same information to the estimating equation. Without loss of generality, we re-distribute the complementary mass to $+\infty$, and the contribution from these observations to the estimating equation is $m(X_i, +\infty; t_0, \boldsymbol{\beta}_{t_0}) = -\mu(X_i; \boldsymbol{\beta}_{t_0}) Z_i$.

In practice, the weights $w\{O_i; t_0, \boldsymbol{\beta}(\cdot)\}$ in (5) depend on unknown distribution function $F(\cdot|X)$ which needs to be estimated. We substitute $\boldsymbol{\beta}(\cdot)$ with the IPW estimators, denoted as $\widetilde{\boldsymbol{\beta}}(\cdot)$, to obtain the weights $w\{O_i; t_0, \widetilde{\boldsymbol{\beta}}(\cdot)\}$ to be redistributed. The REW estimator $\widehat{\boldsymbol{\beta}}_{t_0}$ then solves the weighted estimating equation

$$S_n(O; t_0, \boldsymbol{\beta}_{t_0}, \widetilde{\boldsymbol{\beta}}(\cdot)) = n^{-1} \sum_{i=1}^{n} s(O_i; t_0, \boldsymbol{\beta}_{t_0}, \widetilde{\boldsymbol{\beta}}(\cdot)) = 0. \tag{6}$$

It is extremely easy to implement this weighting scheme. Without loss of generality, assume the first n_0 subjects have unobserved outcomes $I(T_i \le t_0)$. Create pseudo-observations $\widetilde{O}_1 = (X_1, +\infty, \Delta_1), \cdots, \widetilde{O}_{n_0} = (X_{n_0}, +\infty, \Delta_{n_0})$. Append all pseudo-observations to the original observations to obtain observations $(O_1, \cdots, O_n, \widetilde{O}_1, \cdots, \widetilde{O}_{n_0})$ with weights

$$[w\{O_1; t_0, \widetilde{\boldsymbol{\beta}}(\cdot)\}, \cdots, w\{O_n; t_0, \widetilde{\boldsymbol{\beta}}(\cdot)\}, 1 - w\{O_1; t_0, \widetilde{\boldsymbol{\beta}}(\cdot)\}, \cdots, 1 - w\{O_{n_0}; t_0, \widetilde{\boldsymbol{\beta}}(\cdot)\}].$$

Then $\widehat{\boldsymbol{\beta}}_{t_0}$ is estimated by a weighted logistic regression. The weights $w\{O; t_0, \widetilde{\boldsymbol{\beta}}(\cdot)\}$ extract information at multiple time points simultaneously, and thus pool information across time points to estimate the distribution function at t_0.

2.3 Asymptotic properties

To show consistency and asymptotic normality of $\widehat{\boldsymbol{\beta}}(t)$ at fixed t obtained from (6), we will need the following technical conditions:

A1. Assume that $\boldsymbol{\beta}(t)$ is right continuous with left-hand limits (cadlag) componentwise.

A2. Assume that for $t \in [a, b]$ with $b < \infty$ to be finite, and there exists subjects with $P(\min(C_i, T_i) > b) > 0$. Also assume $\boldsymbol{\beta}(t)$ is uniformly bounded on $[a, b]$ componentwise, that is, $\sup_{t\in[a,b]} |\boldsymbol{\beta}(t)| \leq c < \infty$ componentwise.

A3. Assume that the covariates X_i are not degenerate, i.e., $\mathrm{pr}(X_i = x_0) \neq 1$ and are bounded in probability, i.e., $\mathrm{pr}(|X_i| < c) = 1$.

A4. Assume that the censoring times are bounded, i.e., $\mathrm{pr}(C_i < c) = 1$.

A5. Assume that $E\big(Z_i Z_i^T \exp\{Z_i^T \boldsymbol{\beta}(t)\}/[1 + \exp\{Z_i^T \boldsymbol{\beta}(t)\}]^2\big)$ is positive definite.

The conditions A1-A2 control the size of the parameter space. The condition A2 states that one can only estimate distribution function in the time range where there are still subjects with positive probability of being at risk. The conditions A3-A4 exclude some degenerate cases. The condition A5 ensures a unique solution to the estimating equation. For the simplicity of notation we let $\boldsymbol{\theta} = \boldsymbol{\beta}(t_0)$ denote $\boldsymbol{\beta}(\cdot)$ evaluated at t_0 in this subsection. The following theorem establishes the consistency of the estimator $\widehat{\boldsymbol{\theta}}$.

<u>Theorem</u> 1 *Assume that $\{O_i, i = 1, \cdots, n\}$ are i.i.d. random samples, and T_i and C_i are independent given X_i. Then under model (2) and assumptions A1-A5, $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$ in probability as $n \to \infty$ for any $t_0 \in (a, b)$.*

The proof of this theorem uses the semiparametric asymptotic results developed in Newey [18] and Chen et al. [19].

Since the final estimator involves estimates $\widehat{\boldsymbol{\beta}}(\cdot)$ in the entire range of T_i, uniform consistency of the initial estimator is required. The next theorem establishes the asymptotic normality of $\widehat{\boldsymbol{\theta}}$.

Theorem 2 *Under the assumptions of Theorem 1, as $n \to \infty$,*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to N(0, A^{-1}VA^{-1})$$

in distribution, where $A = E[\mu(X_i; \boldsymbol{\theta})\{1-\mu(X_i; \boldsymbol{\theta})\}Z_iZ_i^T]$, $V = cov\{s(O_i; t_0, \boldsymbol{\theta}, \boldsymbol{\beta}) + \xi(T_i; t_0, \boldsymbol{\theta}, \boldsymbol{\beta})\}$,

$$\begin{aligned}\xi(T_i; t_0, \boldsymbol{\theta}, \boldsymbol{\beta}) &= \int_0^{t_0} g(u) \int h(x)zz^T \Big[F(t_0|x)\{1 - F(t_0|x)\}\psi(x, T_i; t_0, \boldsymbol{\theta}) \\ &\quad -F(u|x)\{1 - F(t_0|x)\}\psi\{x, T_i; u, \boldsymbol{\beta}(u)\}\Big] dxdu,\end{aligned}$$

$g(u)$ is the density function for C_i, $h(x)$ is the density function for X_i, $z = (1, x)^T$, and $\psi\{x, T_i; u, \boldsymbol{\beta}(u)\}$ is defined in the appendix.

The proof of this theorem is in the appendix and it also uses the results in Newey [18].

3 Numeric results

In this section, we provide Monte Carlo results on simulation experiments and application of the method to a real world study.

3.1 Simulation studies

To study the finite sample performance of the proposed estimator, we ran two sets of simulation studies. In each set, the true survival times were generated from the model (1) with $\beta_0(t) = \beta_{00} + \beta_{01}\log(t)$, $\beta_1(t) = \beta_{10} + \beta_{11}\log(t)$ and $(\beta_{00}, \beta_{01}, \beta_{10}, \beta_{11})^T = (-80, 21.5, -1.4, 0.7)^T$. The parameters were designed such that the cumulative risk functions resembles the fit from COHORT data in Section 3.2. In simulation setting A (Tables 1 and 2), we simulated X_i from a multinomial distribution with support on integer values between 41 and 50 representing CAG repeats. CAG 41-45 were simulated with an equal probability, which is two times of the equal probability that CAG 46-50 were simulated from. The censoring times were generated from a Beta distribution where the overall censoring rate is about 35%. We simulated two samples sizes $n = 1000$ and $n = 2000$ since the real data has a sample size of 1151.

We compared two types of estimators. The first is the initial inverse probability weighted estimator (IPW) $\widetilde{\beta}(t)$ from (3) and the second is the proposed redistribution to the right weighted estimator (REW) $\widehat{\beta}(t)$ from (6). Since the theoretical variance estimator involves integrations and unknown quantities which are difficult to compute, we used bootstrap to obtain the mean estimated standard errors of the two estimators in each simulation repetition. The empirical standard errors based on the total 400 repetitions and the empirical MSEs were also summarized in tables 1 and 2, where we presented the estimated distribution functions obtained from the two estimators at various ages and CAG values (42 and 46). It can be seen that both IPW and REW estimators have small finite sample biases. The mean estimated standard errors and empirical standard errors are close to each other over most age range with some exceptions at the extreme tail area. The empirical standard error of REW is smaller than that of IPW, especially at older ages. For example, the efficiency gain of REW over IPW is 10% at age 50 for CAG=42 and $n = 1000$. The coverage of the 95% confidence interval is close to the nominal level when age is below 60 for both IPW and REW. At age 60, since censoring is heavier, the coverages of both IPW and REW are lower than the nominal level, with the performance of REW slightly better between the two. The mean estimated standard error of IPW estimator at age 55 and CAG=42 in table 1 is very large because the IPW estimator has unstable weights at that age range due to a turning point in the distribution function. It is a limitation of the IPW estimator and the increased variance when the weights are unstable was reported in the literature [**?**]. This issue is relieved when we increased the sample size as seen in table 2. In addition, the proposed estimator does not suffer from the high variability of the IPW initial estimator.

We presented the true and the mean estimated cumulative distribution functions (CDFs) obtained from the REW estimator and their empirical 95% CI at various CAGs in figure 1. The estimated curves coincide with the true curves in most cases. When CAG=42 and $n = 1000$, there appears to be a small bias at the tail area, for example, at $t = 65$ (bias=0.0051, empirical SE=0.0015). However, this bias is within the variability range, which may be explained by the higher censoring rate within this range for subjects with CAG=42 (about 45%). When we increase the sample size to $n = 2000$, the bias decreased to almost zero.

In simulation setting B (table 3), we basically kept the same setting as in A, but increased the censoring rate to 45% and also increased the number of simulations to 2000. Due to the computation burden we didn't conduct the bootstrap on each simulation repetition to get the mean estimated SEs and MSEs, as well as the coverage probabilities. Only the empirical SEs and MSEs are reported in table 3. The results are similar to those in tables 1 and 2, where the empirical SEs and MSEs of REW are consistently smaller than those of IPW.

In addition to the above estimators, we also investigated a smoothed REW estimator, where $\widehat{\boldsymbol{\beta}}(t)$ were smoothed across the range of t subject to monotone constraint using a Generalized Pooled-Adjacent-Violators Algorithm [20]. The mean estimated cumulative distribution functions and empirical standard errors are almost identical to those of the non-smoothed estimator. The maximum absolute difference in the mean of the two estimators averaged across simulations was very small. Therefore we omit the results of the smoothed estimator here.

3.2 Application to COHORT data

As introduced in Section 1, despite identification of the causative gene for HD, there is currently no effective treatment that delays HD onset or stops disease progression. To improve the care of HD patients and inform the development of effective treatment, a large genetic epidemiological study on HD, the Cooperative Huntington's Observational Research Trial (COHORT), was started in 1996. This is a study organized by 42 Huntington Study Group research centers in North America and Australia [3, 4]. Participants in COHORT underwent a clinical evaluation where blood samples are genotyped for huntingtin gene mutation and their CAG repeats lengths were obtained. Modeling the inverse association between the CAG repeats length and age-at-onset of HD accurately is important.

In this section, we fit the COHORT data by the model (1) where we do not assume a parametric form of $\beta_0(t)$ or $\beta_1(t)$ and the censoring distribution $G(\dot{})$. In our analysis, information on CAG repeats length, age at the time of evaluation, and age at diagnosis of HD onset (if a subject had been diagnosed) were available for 1151 subjects recruited in COHORT. In the study, both HD affected carriers and pre-symptomatic carriers (24%) were included. Their ages-at-first-motor-symptom were also recorded. Among 1151 subjects, 876 (76%) subjects had experienced HD motor sign onset and the average age of the diagnosis was 44 years of age. There were 280 (24%) participants who did not develop HD by the end of study and were treated as censored. All the participants were alive at the baseline in order to participate in the study, and none of them died without HD during the follow up years. Censoring was assumed to be independent of HD diagnosis.

To estimate the distribution of age-at-onset of HD given a subject's CAG repeats length, we fit three estimators: IPW, REW, and the Kaplan-Meier (KM) estimator using only subjects with a particular CAG repeats length at a time. Figure 2 presents the estimated CDFs at various CAG values. The results show a positive correlation between the onset probability and the CAG repeats, that is, the cumulative risk of HD onset by a given age increases with increasing number of CAG repeats. Subjects with longer CAG repeats have a higher probability of developing HD by a certain age, which is consistent with the literature [6]. We summarize numerical results of estimated CDFs at a few CAGs and ages in table 4. As a comparison, we see that IPW and REW provides point estimates of CDFs similar to KM using only subjects with the same CAG values. However, the standard errors of REW at different ages and CAGs are smaller than both KM and IPW, suggesting an efficiency gain. For example, at CAG=42 and age 50, the standard error of the cumulative risk estimated by IPW is 18% larger than REW, and KM is 40% larger than REW. The post-hoc smoothing of $\widehat{\boldsymbol{\beta}}(t)$ leads to a CDF close to the non-smoothed CDF and therefore not reported here. We also modeled the survival function for the censoring times $G(\cdot)$ based on CAG repeats using a Cox model. The estimates are identical to those in table 4 up to the third decimal place and therefore not reported here.

4 Discussion

We propose methods to estimate cumulative disease risk from a known mutation (i.e., also referred as the penetrance function in genetic epidemiology) from a nonparametric varying-coefficient model. For most complex diseases, predicting the age-at-onset of a disease from genetic markers such as single-nucleiotide polymorphisms continues to be a challenging issue [21]. The proposed method explores a pseudo-logistic regression and redistributes the probability mass at the censored outcomes to the right. The procedure has desirable numerical and asymptotic properties and is extremely easy to implement. Although we focused on assessing the effect of CAG repeats on HD onset, it is easy to include other covariates with time-invariant effect through a backfitting procedure for models such as

$$\mathrm{logit}\{\mathrm{pr}(T_i \leq t|X_i)\} = \beta_0(t) + \beta_1(t)X_i + \gamma^T Z_i.$$

or model (2). The proposed methods have computational advantages compared to, for example, Peng and Huang [10]. In addition to the logistic link as discussed here, the developed methods can be adapted to transformation models with a known link function.

Satten and Datta [22] showed an equivalence between IPW-based and self-consistency-equation-based methods for Kaplan-Meier estimator for a pure nonparametric model. It is less clear whether such equivalence still holds for our model (1) which is equivalent to a proportional odds model with nonparametric time-varying coefficients. This may be worth future exploration. In some applications, investigators may be interested in testing the distribution function at more than one time point or building confidence bands. We proposed a procedure to test a distribution function in a sequence of pre-specified time points simultaneously in Ma and Wang [23], which can be adapted here. The construction of simultaneous confidence bands may rely on theoretical properties of supremes of Gaussian processes (e.g., Fine et al. [24]). However, such confidence bands may be conservative and the details are beyond the scope of this work.

Lastly, in practice it may not be easy to correctly specify a biologically meaningful parametric form for $c_0(X_i)$ and $c_1(X_i)$ as in model (2). In these situations, using a two-dimensional nonparametric function of X_i and t may be helpful. To assist determining a parametric (e.g., Langbehn et. al. 2004) versus a semiparametric model or nonparametric model, a goodness-of-fit statistic is useful. For example, one may consider test statistic based on supremem norm such as $\sup_t |\widehat{F}^1(t) - \widehat{F}^0(t)|$, where $\widehat{F}^1(t)$ is fitted under a nonparametric or semiparametric model, and $\widehat{F}^0(t)$ is fitted under a parametric model. The critical value or the p-value of the test will be computed empirically based on simulations under the null hypothesis.

Acknowledgments

Wang and Ma's research is supported by NIH grant NS073671-01, NS082062 and NSF grants DMS-1000354, DMS-1206693.

Appendix

A.1 Proof of Theorem 1

We show consistency by Lemma 5.2 in Newey [18]. We need to show uniform consistency of the initial IPW estimator, i.e., $\sup_{t\in[a,b]} |\widehat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)| = o_p(1)$, and verify assumption 5.4 and 5.5 in Newey [18]. First show uniform consistency of the initial IPW estimator. Wang et al.

[25] showed that the IPW estimator can be expanded as

$$\widehat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t) = \frac{1}{n}\sum_{i=1}^{n} \psi\{X_i, T_i; t, \boldsymbol{\beta}(t)\} + o_p(n^{-1/2}), \tag{7}$$

where

$$\psi\{X_i, T_i; t, \boldsymbol{\beta}(t)\} = \phi\{X_i, T_i; t, \boldsymbol{\beta}(t)\} - \sum_{i=1}^{n} \int \frac{[\phi\{X_i, T_i; t, \boldsymbol{\beta}(t)\} - \mathcal{B}(\phi, u)]dM_i^c(u)}{G(u)},$$

$\mathcal{B}(\phi, u) = E[\phi\{X_i, T_i; t, \boldsymbol{\beta}(t)\}|T_i \geq u, X_i]$, and $dM_i^c(u)$ is the martingale of the censoring process. To show uniform consistency, we need to show that the set $\big\{\psi\{X_i, T_i; t, \boldsymbol{\beta}(t)\} : t \in [a, b]\big\}$ is a Glivenko-Cantelli class. Note that $\phi\{X_i, T_i; t, \boldsymbol{\beta}(t)\} = A\{X_i; \boldsymbol{\beta}(t)\}[I(T_i \leq t) - \mu\{X_i; \boldsymbol{\beta}(t)\}]Z_i$. Indicator functions are cadlag processes which are bounded in total variation and belong to the Vapnic-Červonencis class. Thus they are bounded in uniform entropy integral with square-integrable envelope. It follows that they belongs to a Donsker class, and hence Glivenko-Cantelli. In addition, $\mu\{X_i; \boldsymbol{\beta}(t)\}$ is Lipschitz continuous. By assumption A1, $\boldsymbol{\beta}(t)$ belongs to a cadlag processes therefore are also bounded in uniform entropy integral. Since Lipschitz continuous functions of classes bounded in uniform entropy integral and pointwise measurable are also bounded in uniform entropy integral and pointwise measurable, $\big\{\mu\{X_i; \boldsymbol{\beta}(t)\}, t \in [a, b]\big\}$, is Glivenko-Cantelli. From $A\{X_i; \boldsymbol{\beta}(t)\} = \Big\{E(\mu\{X_i; \boldsymbol{\beta}(t)\}[1 - \mu\{X_i; \boldsymbol{\beta}(t)\}]Z_i Z_i^T)\Big\}^{-1}$, under assumption A5, $A\{X_i; \boldsymbol{\beta}(t)\}$ is bounded from below and above by positive constants component-wise and bounded in uniform entropy integral, therefore is Glivenko-Cantelli. Lastly, since X_i is bounded and products of classes with bounded uniform entropy integral also have bounded uniform entropy integral, we have $\Big\{\phi\{X_i, T_i; t, \boldsymbol{\beta}(t)\} : t \in [a, b]\Big\}$ is Glivenko-Cantelli.

Now we check the second term in $\psi\{X_i, T_i; t, \boldsymbol{\beta}(t)\}$. Note that

$$\begin{aligned}
\mathcal{B}(\phi, u) &= E[\phi\{X_i, T_i; t, \boldsymbol{\beta}(t)\}|X_i, T_i \geq u] \\
&= \frac{E[\phi\{X_i, T_i; t, \boldsymbol{\beta}(t)\}I(T_i \geq u)|X_i]}{E(T_i \geq u|X_i)} \\
&= \frac{\int_u^\infty A\{X_i; \boldsymbol{\beta}(t)\}[I(s \leq t) - \mu\{X_i; \boldsymbol{\beta}(t)\}]Z_i dF(s|X_i)}{1 - F(u|X_i)} \\
&= \frac{A\{X_i; \boldsymbol{\beta}(t)\}[F(t|X_i) - F(u|X_i) - \mu\{X_i; \boldsymbol{\beta}(t)\}\{1 - F(u|X_i)\}]Z_i}{1 - F(u|X_i)}.
\end{aligned}$$

Therefore

$$\begin{aligned}
&\sum_{i=1}^n \int \frac{[\phi\{X_i, T_i; t, \boldsymbol{\beta}(t)\} - \mathcal{B}(\phi, u)]dM_i^c(u)}{G(u)} \\
= &\sum_{i=1}^n \frac{(1-\delta_i)}{G(C_i)}\Big\{\phi\{X_i, T_i; t, \boldsymbol{\beta}(t)\} \\
&- \frac{A\{X_i; \boldsymbol{\beta}(t)\}[F(t|X_i) - F(C_i|X_i) - \mu\{X_i; \boldsymbol{\beta}(t)\}\{1 - F(C_i|X_i)\}]Z_i}{1 - F(C_i|X_i)}\Big\}. \qquad (8)
\end{aligned}$$

Under condition A4, $G(C_i) > 0$. Under model (2) and conditions A1, A2, the above term indexed by t is also Glivenko-Cantelli. This proves that $\Big\{\psi\{X_i, T_i; t, \beta(t)\} : t \in [a, b]\Big\}$ is Glivenko-Cantelli. It follows that

$$\sup_{t\in[a,b]} \left| n^{-1} \sum_{i=1}^n \psi\{X_i, T_i; t, \boldsymbol{\beta}(t)\} - E[\psi\{X_i, T_i; t, \boldsymbol{\beta}(t)\}] \right| \to 0.$$

Since $E[\psi\{X_i, T_i; t, \boldsymbol{\beta}(t)\}] = 0$, we have shown the uniform consistency of the IPW estimator,

$$\sup_{t\in[a,b]} |\widehat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)| = 0.$$

Now we verify assumptions 5.4 and 5.5 in Newey [18]. In what follows, we use $\boldsymbol{\theta}$ and $\boldsymbol{\beta}(\cdot)$ to denote true parameter values and use $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\beta}}(\cdot)$ to denote other values different from the truth. For assumption 5.4 (i), it is straightforward to see that $s(O_i; t_0, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\beta})$ is continuous in

$\widetilde{\boldsymbol{\theta}}$ and is bounded under the assumptions A1, A3, and A4. For the assumption 5.4 (ii), note

$$\begin{array}{rl} & s(O_i; t_0, \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\beta}}) - s(O_i, ; t_0, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\beta}) \\ = & I(T_i > C_i) I(C_i < t_0) \{ w(O_i; t_0, \widetilde{\boldsymbol{\beta}}) - w(O_i; t_0, \boldsymbol{\beta}) \} Z_i \\ = & I(T_i > C_i) I(C_i < t_0) Z_i \left[\dfrac{\mu\{X_i; \widetilde{\boldsymbol{\beta}}(t_0)\} - \mu\{X_i; \widetilde{\boldsymbol{\beta}}(C_i)\}}{1 - \mu\{X_i; \widetilde{\boldsymbol{\beta}}(C_i)\}} - \dfrac{\mu\{X_i; \boldsymbol{\beta}(t_0)\} - \mu\{X_i; \boldsymbol{\beta}(C_i)\}}{1 - \mu\{X_i; \boldsymbol{\beta}(C_i)\}} \right] \\ = & I(T_i > C_i) I(C_i < t_0) Z_i Z_i^T \left(\dfrac{\mu\{X_i; \check{\boldsymbol{\beta}}(t_0)\}[1 - \mu\{X_i; \check{\boldsymbol{\beta}}(t_0)\}]}{1 - \mu\{X_i; \check{\boldsymbol{\beta}}(C_i)\}} \{ \widetilde{\boldsymbol{\beta}}(t_0) - \boldsymbol{\beta}(t_0) \} \right. \\ & \left. - \dfrac{\mu\{X_i; \check{\boldsymbol{\beta}}(C_i)\}[1 - \mu\{X_i; \check{\boldsymbol{\beta}}(t_0)\}]}{1 - \mu\{X_i; \check{\boldsymbol{\beta}}(C_i)\}} \{ \widetilde{\boldsymbol{\beta}}(C_i) - \boldsymbol{\beta}(C_i) \} \right), \end{array} \tag{9}$$

where $\check{\boldsymbol{\beta}}(u)$ is on the line segment between $\widetilde{\boldsymbol{\beta}}(u)$ and $\boldsymbol{\beta}(u)$. Here the last equality is obtained by taking pathwise derivative with respect to $\boldsymbol{\beta}$. See also (10). Since $0 < \mu\{x; \check{\boldsymbol{\beta}}(u)\} < 1$ for $u \in [a, b]$, it follows that there exists $b(O_i)$ such that component-wise we have

$$||s(O_i; t_0, \boldsymbol{\theta}, \widetilde{\boldsymbol{\beta}}) - s(O_i, ; t_0, \boldsymbol{\theta}, \boldsymbol{\beta})|| \leq b(O_i) ||\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}||.$$

By condition A5, the assumption 5.5 in Newey (1994) is satisfied. Finally, by Lemma 5.2 of Newey [18], we have $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta} + o_p(1)$.

A.2 Proof of Theorem 2

We show the asymptotic normality of $\widehat{\boldsymbol{\theta}}$ by Lemma 5.3 of Newey [18]. For assumption 5.1(i), note again

$$s(O_i; t_0, \boldsymbol{\theta}, \widetilde{\boldsymbol{\beta}}) - s(O_i; t_0, \boldsymbol{\theta}, \boldsymbol{\beta}) = I(T_i > C_i) I(C_i < t_0) \{ w(O_i; t_0, \widetilde{\boldsymbol{\beta}}) - w(O_i; t_0, \boldsymbol{\beta}) \} Z_i.$$

We now compute a pathwise derivative of $w(O_i; t_0, \boldsymbol{\beta})$ w.r.t. $\boldsymbol{\beta}$ evaluated at the true $\boldsymbol{\beta}$ in the direction $[\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}]$. Let $\boldsymbol{\beta}_\epsilon(u) = \boldsymbol{\beta}(u) + \epsilon\{\widetilde{\boldsymbol{\beta}}(u) - \boldsymbol{\beta}(u)\}$. We can verify that

$$\begin{aligned}
& \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left\{\frac{F(t_0|X_i;\boldsymbol{\beta}_\epsilon) - F(C_i|X_i;\boldsymbol{\beta}_\epsilon)}{1 - F(C_i|X_i;\boldsymbol{\beta}_\epsilon)} - \frac{F(t_0|X_i) - F(C_i|X_i)}{1 - F(C_i|X_i)}\right\} \\
= \quad & \frac{F(t_0|X_i)\{1 - F(t_0|X_i)\}Z_i^T}{1 - F(C_i|X_i)}\{\widetilde{\boldsymbol{\beta}}(t_0) - \boldsymbol{\beta}(t_0)\} \\
& -\frac{F(C_i|X_i)\{1 - F(t_0|X_i)\}Z_i^T}{1 - F(C_i|X_i)}\{\widetilde{\boldsymbol{\beta}}(C_i) - \boldsymbol{\beta}(C_i)\}. \qquad (10)
\end{aligned}$$

Let

$$\begin{aligned}
D(O_i; \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \quad = \quad & I(T_i > C_i)I(C_i < t_0)Z_iZ_i^T\Big[\frac{F(t_0|X_i)\{1 - F(t_0|X_i)\}}{1 - F(C_i|X_i)}\{\widetilde{\boldsymbol{\beta}}(t_0) - \boldsymbol{\beta}(t_0)\} \\
& -\frac{F(C_i|X_i)\{1 - F(t_0|X_i)\}}{1 - F(C_i|X_i)}\{\widetilde{\boldsymbol{\beta}}(C_i) - \boldsymbol{\beta}(C_i)\}\Big]. \qquad (11)
\end{aligned}$$

From (9), we can verify

$$\begin{aligned}
& s(O_i; t_0, \boldsymbol{\theta}, \widetilde{\boldsymbol{\beta}}) - s(O_i; t_0, \boldsymbol{\theta}, \boldsymbol{\beta}) - D(O_i; \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
= \quad & I(T_i > C_i)I(C_i < t_0)Z_iZ_i^T\Bigg(\frac{\mu\{X_i; \check{\boldsymbol{\beta}}(t_0)\}[1 - \mu\{X_i; \check{\boldsymbol{\beta}}(t_0)\}]}{1 - \mu\{X_i; \check{\boldsymbol{\beta}}(C_i)\}}\{\widetilde{\boldsymbol{\beta}}(t_0) - \boldsymbol{\beta}(t_0)\} \\
& -\frac{\mu\{X_i; \check{\boldsymbol{\beta}}(C_i)\}[1 - \mu\{X_i; \check{\boldsymbol{\beta}}(t_0)\}]}{1 - \mu\{X_i; \check{\boldsymbol{\beta}}(C_i)\}}\{\widetilde{\boldsymbol{\beta}}(C_i) - \boldsymbol{\beta}(C_i)\}\Bigg) - D(O_i; \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
= \quad & I(T_i > C_i)I(C_i < t_0)Z_iZ_i^T \\
& \times\Bigg(\Big[\frac{\mu\{X_i; \check{\boldsymbol{\beta}}(t_0)\}[1 - \mu\{X_i; \check{\boldsymbol{\beta}}(t_0)\}]}{1 - \mu\{X_i; \check{\boldsymbol{\beta}}(C_i)\}} - \frac{F(t_0|X_i)\{1 - F(t_0|X_i)\}}{1 - F(C_i|X_i)}\Big]\{\widetilde{\boldsymbol{\beta}}(t_0) - \boldsymbol{\beta}(t_0)\} \\
& -\Big[\frac{\mu\{X_i; \check{\boldsymbol{\beta}}(C_i)\}[1 - \mu\{X_i; \check{\boldsymbol{\beta}}(t_0)\}]}{1 - \mu\{X_i; \check{\boldsymbol{\beta}}(C_i)\}} - \frac{F(C_i|X_i)\{1 - F(t_0|X_i)\}}{1 - F(C_i|X_i)}\{\widetilde{\boldsymbol{\beta}}(C_i) - \boldsymbol{\beta}(C_i)\}\Big]\Bigg),
\end{aligned}$$

where again $\check{\boldsymbol{\beta}}(u)$ is on the line segment of $\widetilde{\boldsymbol{\beta}}(u)$ and $\boldsymbol{\beta}(u)$. It is now easy to see that

$$||s(O_i; t_0, \boldsymbol{\theta}, \widetilde{\boldsymbol{\beta}}) - s(O_i; t_0, \boldsymbol{\theta}, \boldsymbol{\beta}) - D(O_i; \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})|| \le b(O_i)||\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2.$$

For (ii) in assumption 5.1, we need to show that the convergence rate of the IPW estimator

$\widehat{\boldsymbol{\beta}}$ is at least $n^{1/4}$. Let $\mathcal{F}$ denote all cadlag functions uniformly bounded on $[a,b]$. By adapting the proof in the previous item, we know that $\{\psi\{X_i, T_i; \boldsymbol{\beta}(t)\} : t \in [a,b], \boldsymbol{\beta} \in \mathcal{F}\}$ belongs to a Donsker class. Therefore $\sqrt{n}\{\widehat{\boldsymbol{\beta}}(\cdot) - \boldsymbol{\beta}(\cdot)\}$ converges weakly to a Gaussian process. Therefore this assumption is satisfied.

We now prove assumption 5.2 (stochastic equicontinuity). Note

$$\begin{aligned}
&\int D(o; \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) dG dH \\
= &\int_0^{t_0} g(u) \int h(x) \Big[\frac{F(t_0|x)\{1 - F(t_0|x)\} z z^T \{\widetilde{\boldsymbol{\beta}}(t_0) - \boldsymbol{\beta}(t_0)\}}{1 - F(u|x)} \{1 - F(u|x)\} \\
&- \frac{F(u|x)\{1 - F(t_0|x)\} z z^T \{\widetilde{\boldsymbol{\beta}}(u) - \boldsymbol{\beta}(u)\}}{1 - F(u|x)} \{1 - F(u|x)\} \Big] dx du \\
= &\int_0^{t_0} g(u) \int h(x) z z^T \Big[F(t_0|x)\{1 - F(t_0|x)\}\{\widetilde{\boldsymbol{\beta}}(t_0) - \boldsymbol{\beta}(t_0)\} \\
&- F(u|x)\{1 - F(t_0|x)\}\{\widetilde{\boldsymbol{\beta}}(u) - \boldsymbol{\beta}(u)\} \Big] dx du.
\end{aligned}$$

A sufficient condition for stochastic equicontinuity is provided in Chen et al. [19], Remark 2. To be specific, we need to show for $\delta_n = o_p(1)$,

$$\sup_{||\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}|| \le \delta_n} || \frac{1}{n} \sum_{i=1}^n D(O_i, \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \int D(o, \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) dG dH || = o_p(n^{-1/2}).$$

This can be proved by showing the process $\{D(O_i, \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) : t \in [a,b], \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \in \mathcal{F}\}$ belongs to a Donsker class. Note the form of $D(O_i, \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$ in (11), again by adapting proof in item 4 this holds under the conditions A1-A5.

A sufficient condition for assumption 5.3 in Newey (1994) is

$$\sqrt{n} \int D(o; \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) dG dH - \sum_{i=1}^n \alpha(O_i)/\sqrt{n} \to 0,$$

for some $\alpha(\cdot)$ (p.1366, 18). Using the expansion (7) for $\widehat{\boldsymbol{\beta}}(t)$, we obtain

$$\int D(o; \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) dG dH \quad = \quad \frac{1}{n} \sum_{i=1}^n \xi(T_i; t_0, \boldsymbol{\theta}, \boldsymbol{\beta}) + o_p(n^{-1/2}),$$

where

$$\begin{aligned}\xi(T_i;t_0,\boldsymbol{\theta},\boldsymbol{\beta}) &= \int_0^{t_0} g(u)\int h(x)zz^T\Big[F(t_0|x)\{1-F(t_0|x)\}\psi(x,T_i;t_0,\boldsymbol{\theta}) \\ &\quad -F(u|x)\{1-F(t_0|x)\}\psi\{x,T_i;u,\boldsymbol{\beta}(u)\}\Big]dxdu.\end{aligned}$$

Therefore assumption 5.3 holds.

For assumption 5.6, it is straightforward that (i) and (ii) are satisfied. We have

$$A = E\left\{\frac{\partial s(O_i;t_0,\boldsymbol{\theta},\boldsymbol{\beta})}{\partial\boldsymbol{\theta}}\right\} = E[\mu(X_i;\boldsymbol{\theta})\{1-\mu(X_i;\boldsymbol{\theta})\}Z_iZ_i^T],$$

which is nonsingular under the assumption A5. It is easy to see that (iv) holds. For (v), since $\dfrac{\partial s(O_i;t_0,\boldsymbol{\theta},\boldsymbol{\beta})}{\partial\boldsymbol{\theta}}$ is continuous in $\boldsymbol{\theta}$, assumption 5.4 (i) holds for $\dfrac{\partial s(O_i;t_0,\boldsymbol{\theta},\boldsymbol{\beta})}{\partial\boldsymbol{\theta}}$. The assumption 5.4 (ii) holds for $\dfrac{\partial s(O_i;t_0,\boldsymbol{\theta},\boldsymbol{\beta})}{\partial\boldsymbol{\theta}}$ since it does not depend on $\boldsymbol{\beta}$.

By Lemma 5.3 of Newey [18], we obtain

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}) \to N(0, A^{-1}VA^{-1}).$$

References

1. Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on huntingtons disease chromosomes. *Cell* 1993; **72**:971–983.

2. Lee JM, Ramos EM, Lee JH, Gillis T, Mysore JS, Hayden MR, Warby SC, Morrison P, Nance M, Ross CA, et al. Cag repeat expansion in huntington disease determines age at onset in a fully dominant fashion. *Neurology* 2012; **78**(10):690–695.

3. Kieburtz K, Huntington Study Group. The unified huntington's disease rating scale: reliability and consistency. *Movement Disorders* 1996; **11**:136–142.

4. Dorsey ER, Beck C, Adams M, Huntington Study Group. Trend-hd communicating clinical trial results to research participants. *Archives of Neurology* 2008; **65**(12):1590–1595.

5. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B* 1972; **34**(2):187–220.

6. Langbehn DR, Brinkman RR, Falush D, Paulsen JS, Hayden MR. A new model for prediction of the age of onset and penetrance for huntington's disease based on cag length. *Clinical Genetics* 2004; **65**:267–277.

7. Chen T, Wang Y, Ma Y, Marder K, Langbehn DR. Predicting disease onset from mutation status using proband and family data with applications to huntington's disease. *Journal of Probability and Statistics* 2012; **2012, Article ID 375935**.

8. Langbehn DR, Hayden MR, Paulsen JS, the PREDICT-HD Investigators of the Huntington Study Group. Cag-repeat length and the age of onset in huntington disease (hd): A review and validation study of statistical approaches. *American Journal of Medical Genetics Part B* 2009; **153B**:397–408.

9. Jung SH. Regression analysis for long-term survival rate. *Biometrika* 1996; **83**:227–232.

10. Peng L, Huang Y. Survival analysis with temporal covariate effects. *Biometrika* 2007; **94**:719–733.

11. Chen YQ, Hu N, Cheng SC, Musoke P, Zhao LP. Estimating regression parameters in an extended proportional odds model. *Journal of the American Statistical Association* 2012; **107**:318–330.

12. Zeng D, Lin DY. Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society, Series B* 2007; **69**:1–30.

13. Efron B. The two sample problem with censored data. *Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1967; **4**.

14. Portnoy S. Censored regression quantiles. *Journal of the American Statistical Association* 2003; **98**(464):1001–1012.

15. Wang HJ, Wang L. Locally weighted censored quantile regression. *Journal of the American Statistical Association* 2009; **104**(487):1117–1128.

16. Ma Y, Wei Y. Analysis on censored quantile residual life model via spline smoothing. *Statistica Sinica* 2012; **22**:47–68.

17. Bang H, Tsiatis AA. Estimating medical costs with censored data. *Biometrika* 2000; **87**(2):329–343.

18. Newey WK. The asymptotic variance of semiparametric estimators. *Econometrica* 1994; **62**:1349–1382.

19. Chen X, Linton O, Keilegom IV. Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* 2003; **71**(5):1591–1608.

20. de Leeuw J, Hornik K, Mair P. Isotone optimization in r: Pool-adjacent-violators algorithm (pava) and active set methods. *Journal of Statistical Software* 2009; **32**:1–24.

21. Kang J, Chobc J, Zhao H. Practical issues in building risk-predicting models for complex diseases. *Journal of Biopharmaceutical Statistics* 2010; **20**(2):415–440.

22. Satten GA, Datta S. The kaplancmeier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician* 2001; **55**(3):207–210.

23. Ma Y, Wang Y. Estimating disease distribution functions from censored mixture data. *Journal of the Royal Statistical Society, Series C* 2014; **63**:1–23.

24. Fine J, Yan J, Kosorok MR. Temporal process regression. *Biometrika* 2004; **91**:683–703.

25. Wang Y, Garcia TP, Ma Y. Nonparametric estimation for censored mixture data with application to the cooperative huntingtons observational research trial. *Journal of the American Statistical Association* 2012; **107**(500):1324–1338.

Table 1: Mean estimated CDFs by IPW and REW estimators, their mean estimated SEs, empirical SEs, empirical MSEs, and 95% coverages (all in 100 fold scale). $n = 1000$, 400 simulations, CAG=42 and 46.

						CAG=42					
						EMP	EMP	EMP	EMP		
Age	TRUE	IPW†	REW‡	SE(IPW)*	SE(REW)	SE(IPW)*	SE(REW)	MSE(IPW)	MSE(REW)	Cov(IPW)$^{\$}$	Cov(REW)
30	0.005	0.010	0.009	0.072	0.065	0.015	0.014	<0.001	<0.001	84.75	84.25
35	0.108	0.117	0.109	0.062	0.058	0.063	0.059	<0.001	<0.001	89.50	88.00
40	1.426	1.407	1.353	0.433	0.414	0.418	0.392	0.002	0.002	93.50	92.00
45	12.447	12.498	12.243	2.248	2.155	2.227	2.133	0.050	0.046	95.25	94.75
50	52.329	52.702	52.149	4.688	3.995	4.309	3.791	0.187	0.147	96.75	96.00
55	87.460	89.048	88.446	32.094	6.614	4.078	3.303	0.192	0.113	99.00	97.25
60	97.418	99.222	98.886	7.796	6.437	1.485	1.378	0.055	0.020	28.00	66.75
						CAG=46					
						EMP	EMP	EMP	EMP		
Age	TRUE	IPW	REW	SE(IPW)	SE(REW)	SE(IPW)	SE(REW)	MSE(IPW)	MSE(REW)	Cov(IPW)	Cov(REW)
30	0.275	0.282	0.269	1.212	1.142	0.224	0.215	0.001	<0.001	95.50	95.00
35	7.790	7.830	7.635	1.586	1.555	1.609	1.575	0.026	0.025	93.75	93.25
40	62.084	62.444	62.206	3.562	3.476	3.801	3.713	0.146	0.138	92.00	92.75
45	95.722	95.674	95.703	1.231	1.163	1.210	1.158	0.015	0.013	95.50	93.00
50	99.571	99.517	99.529	0.294	0.257	0.276	0.244	0.001	0.001	89.00	89.50
55	99.948	99.908	99.920	0.791	0.143	0.129	0.101	<0.001	<0.001	93.00	86.75
60	99.992	99.973	99.979	0.528	0.506	0.089	0.052	<0.001	<0.001	28.00	58.50

†: CDF by inverse probability weighting (IPW) estimator, $\widetilde{\beta}_t$, solving (3).

‡: CDF by re-distributed to right (REW) weighted estimator, $\widehat{\beta}_t$, using IPW as initial estimator for weight to solve (6).

*: Mean estimated SE of CDF by IPW estimator.

*: Empirical SE of CDF by IPW estimator.

$^{\$}$: 95% coverage probability of CDF by IPW estimator.

Table 2: Mean estimated CDFs by IPW and REW estimators, their mean estimated SEs, empirical SEs, empirical MSEs, and 95% coverages (all in 100 fold scale). $n = 2000$, 400 simulations, CAG=42 and 46.

						CAG=42					
						EMP	EMP	EMP	EMP		
Age	TRUE	IPW†	REW‡	SE(IPW)*	SE(REW)	SE(IPW)*	SE(REW)	MSE(IPW)	MSE(REW)	Cov(IPW)$^\$$	Cov(REW)
30	0.005	0.008	0.007	0.012	0.011	0.009	0.008	<0.001	<0.001	81.75	80.50
35	0.108	0.115	0.107	0.043	0.041	0.042	0.039	<0.001	<0.001	94.75	92.25
40	1.426	1.424	1.364	0.307	0.293	0.280	0.272	0.001	0.001	96.50	92.75
45	12.447	12.479	12.249	1.555	1.509	1.544	1.528	0.024	0.024	95.75	94.25
50	52.329	52.434	52.103	3.047	2.723	2.671	2.523	0.071	0.064	97.00	97.25
55	87.460	87.887	87.623	5.384	2.522	2.450	2.113	0.062	0.045	98.00	97.00
60	97.418	98.511	98.276	12.648	6.951	1.459	1.301	0.033	0.024	64.50	88.50
						CAG=46					
						EMP	EMP	EMP	EMP		
Age	TRUE	IPW	REW	SE(IPW)	SE(REW)	SE(IPW)	SE(REW)	MSE(IPW)	MSE(REW)	Cov(IPW)	Cov(REW)
30	0.275	0.271	0.258	0.249	0.238	0.159	0.152	<0.001	<0.001	91.25	90.00
35	7.790	7.835	7.642	1.108	1.087	1.045	1.024	0.011	0.011	95.25	95.00
40	62.084	62.286	62.123	2.475	2.416	2.478	2.405	0.062	0.058	93.50	95.00
45	95.722	95.650	95.662	0.870	0.823	0.868	0.829	0.008	0.007	95.50	96.00
50	99.571	99.536	99.539	0.198	0.177	0.203	0.178	<0.001	<0.001	91.25	92.25
55	99.948	99.927	99.933	0.123	0.063	0.074	0.056	<0.001	<0.001	89.25	89.50
60	99.992	99.971	99.979	0.593	0.362	0.077	0.045	<0.001	<0.001	62.50	75.00

†: CDF by inverse probability weighting (IPW) estimator, $\widetilde{\beta}_t$, solving (3).

‡: CDF by re-distributed to right (REW) weighted estimator, $\widehat{\beta}_t$, using IPW as initial estimator for weight to solve (6).

*: Mean estimated SE of CDF by IPW estimator.

*: Empirical SE of CDF by IPW estimator.

$^\$$: 95% coverage probability of CDF by IPW estimator.

Table 3: Mean estimated CDFs by IPW and REW estimators, their empirical SEs, and empirical MSEs (all in 100 fold scale). n = 1000 or 2000, 2000 simulations, CAG=42 and 46.

n=1000 CAG=42

Age	TRUE	IPW[†]	REW[‡]	EMP SE(IPW)[⋆]	EMP SE(REW)	EMP MSE(IPW)	EMP MSE(REW)
30	0.005	0.011	0.010	0.018	0.016	<0.001	<0.001
35	0.108	0.120	0.109	0.065	0.060	<0.001	<0.001
40	1.426	1.427	1.344	0.457	0.428	0.002	0.002
45	12.447	12.367	12.066	2.531	2.386	0.064	0.058
50	52.329	52.721	52.275	5.504	4.828	0.305	0.233
55	87.460	89.497	89.032	5.837	4.979	0.382	0.273
60	97.418	99.577	99.451	1.633	1.535	0.073	0.065

n=1000 CAG=46

Age	TRUE	IPW	REW	EMP SE(IPW)	EMP SE(REW)	EMP MSE(IPW)	EMP MSE(REW)
30	0.275	0.285	0.265	0.230	0.216	0.001	<0.001
35	7.790	7.817	7.543	1.621	1.581	0.026	0.026
40	62.084	62.298	62.091	3.704	3.589	0.138	0.129
45	95.722	95.683	95.726	1.384	1.282	0.019	0.016
50	99.571	99.476	99.501	0.414	0.336	0.002	0.001
55	99.948	99.854	99.883	0.292	0.198	0.001	<0.001
60	99.992	99.970	99.974	0.171	0.111	<0.001	<0.001

n=2000 CAG=42

Age	TRUE	IPW	REW	EMP SE(IPW)	EMP SE(REW)	EMP MSE(IPW)	EMP MSE(REW)
30	0.005	0.008	0.007	0.010	0.009	<0.001	<0.001
35	0.108	0.111	0.101	0.044	0.040	<0.001	<0.001
40	1.426	1.447	1.359	0.326	0.304	0.001	0.001
45	12.447	12.419	12.104	1.773	1.691	0.031	0.030
50	52.329	52.348	51.981	3.637	3.265	0.132	0.108
55	87.460	88.330	88.061	3.791	3.263	0.151	0.110
60	97.418	99.052	98.933	1.923	1.752	0.064	0.054

n=2000 CAG=46

Age	TRUE	IPW	REW	EMP SE(IPW)	EMP SE(REW)	EMP MSE(IPW)	EMP MSE(REW)
30	0.275	0.273	0.253	0.163	0.153	<0.001	<0.001
35	7.790	7.713	7.432	1.126	1.092	0.013	0.013
40	62.084	62.084	61.891	2.635	2.588	0.069	0.067
45	95.722	95.712	95.747	0.953	0.884	0.009	0.008
50	99.571	99.529	99.539	0.251	0.216	0.001	<0.001
55	99.948	99.903	99.916	0.144	0.105	<0.001	<0.001
60	99.992	99.959	99.969	0.136	0.093	<0.001	<0.001

[†]: CDF by inverse probability weighting (IPW) estimator, $\widetilde{\beta}_t$, solving (3).

[‡]: CDF by re-distributed to right (REW) weighted estimator, $\widehat{\beta}_t$, using IPW as initial estimator for weight to solve (6).

[⋆]: Empirical SE of CDF by IPW estimator.

Table 4: COHORT data: Estimated CDFs by KM, IPW and REW estimators and their estimated SEs at CAG=42, 44, 46 and 48 (all in 100 fold scale).

			CAG=42			
Age	KM	IPW‡	REW*	SE(KM)	SE(IPW)	SE(REW)
30	<0.001	0.424	0.319	NA	0.170	0.139
35	1.005	1.197	1.028	0.707	0.331	0.303
40	4.210	3.916	3.417	1.458	0.700	0.635
45	10.521	9.520	8.754	2.289	1.350	1.232
50	26.769	28.646	24.286	3.442	2.422	1.989
55	53.365	54.668	50.085	4.015	3.576	2.574
60	74.294	78.742	74.661	3.692	30.086	2.343
			CAG=44			
Age	KM	IPW	REW	SE(KM)	SE(IPW)	SE(REW)
30	2.458	1.347	1.053	1.214	0.394	0.339
35	5.663	4.754	3.747	1.834	0.828	0.729
40	15.638	15.912	13.670	2.938	1.480	1.349
45	36.492	37.504	34.121	3.979	2.140	1.960
50	73.225	68.592	64.816	3.773	2.264	2.251
55	91.844	88.395	86.837	2.436	1.743	1.731
60	96.737	96.008	95.439	1.596	6.436	1.079
			CAG=46			
Age	KM	IPW	REW	SE(KM)	SE(IPW)	SE(REW)
30	4.104	4.195	3.419	2.010	0.867	0.776
35	15.220	17.052	12.731	3.753	1.836	1.557
40	42.329	46.766	41.475	5.230	2.860	2.650
45	82.296	77.389	73.656	4.128	2.779	2.837
50	91.829	92.236	91.364	3.102	1.421	1.493
55	93.872	97.964	97.746	2.923	0.628	0.637
60	97.957	99.364	99.332	1.932	1.014	0.296
			CAG=48			
Age	KM	IPW	REW	SE(KM)	SE(IPW)	SE(REW)
30	13.889	12.314	10.535	5.764	2.078	1.876
35	47.826	45.848	35.347	8.407	4.204	3.594
40	83.437	80.309	76.028	6.469	3.064	3.336
45	93.375	95.127	93.786	4.457	1.271	1.487
50	96.687	98.476	98.381	3.233	0.472	0.478
55	100.000	99.672	99.651	NA	0.157	0.153
60	100.000	99.901	99.905	NA	0.155	0.062

†: CDF by inverse probability weighting (IPW) estimator, $\widetilde{\beta}_t$, solving (3).

‡: CDF by re-distributed to right (REW) weighted estimator, $\widehat{\beta}_t$, using IPW as initial estimator for weight to solve (6).

Figure 1: True and REW CDF curves evaluated at CAG=50, 48, 46, 44, 42 (left to right). The true and mean estimated curves are indistinguishable for most cases. $n = 1000$ (top) and $n = 2000$ (bottom), 400 simulations.

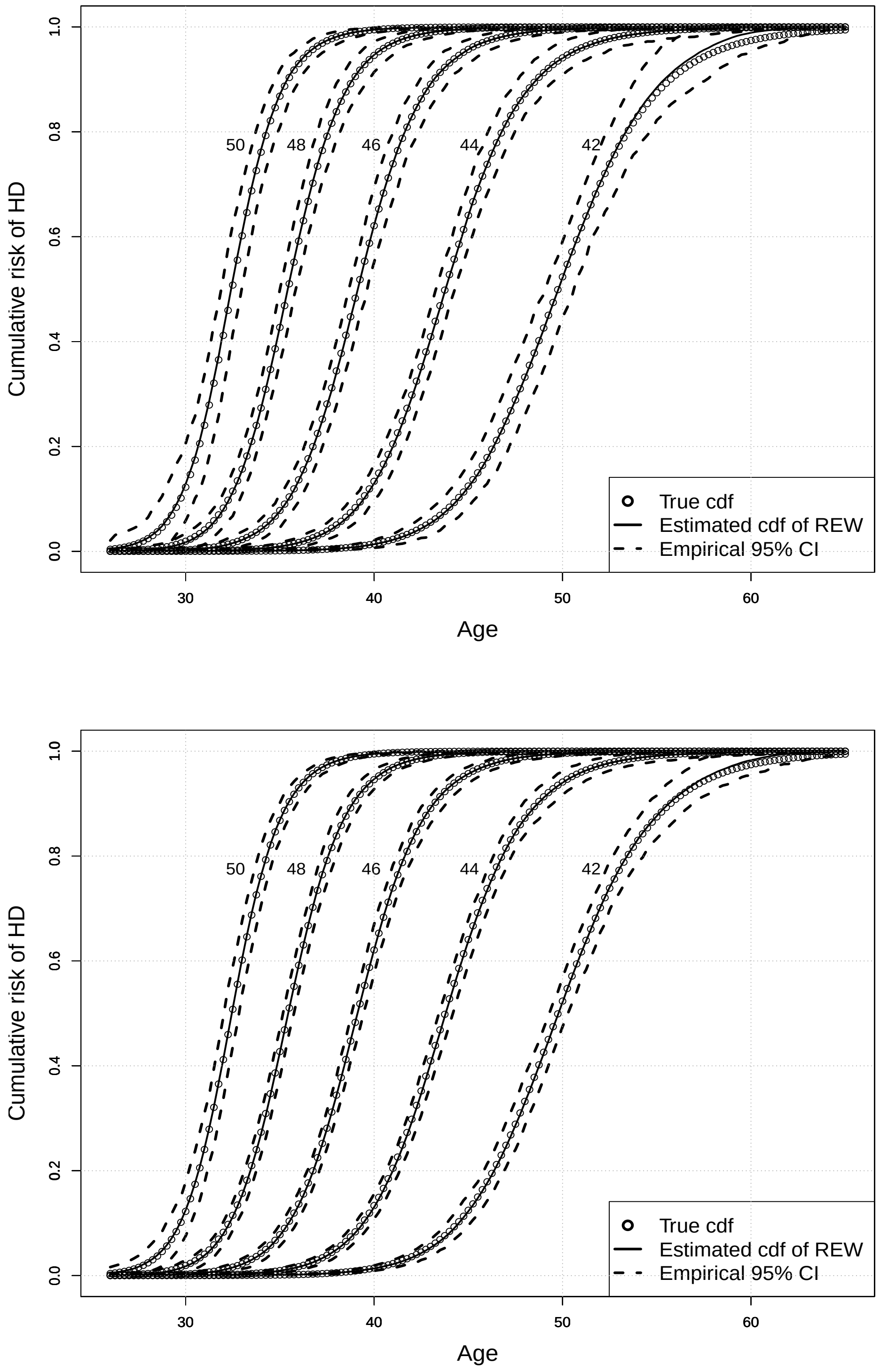
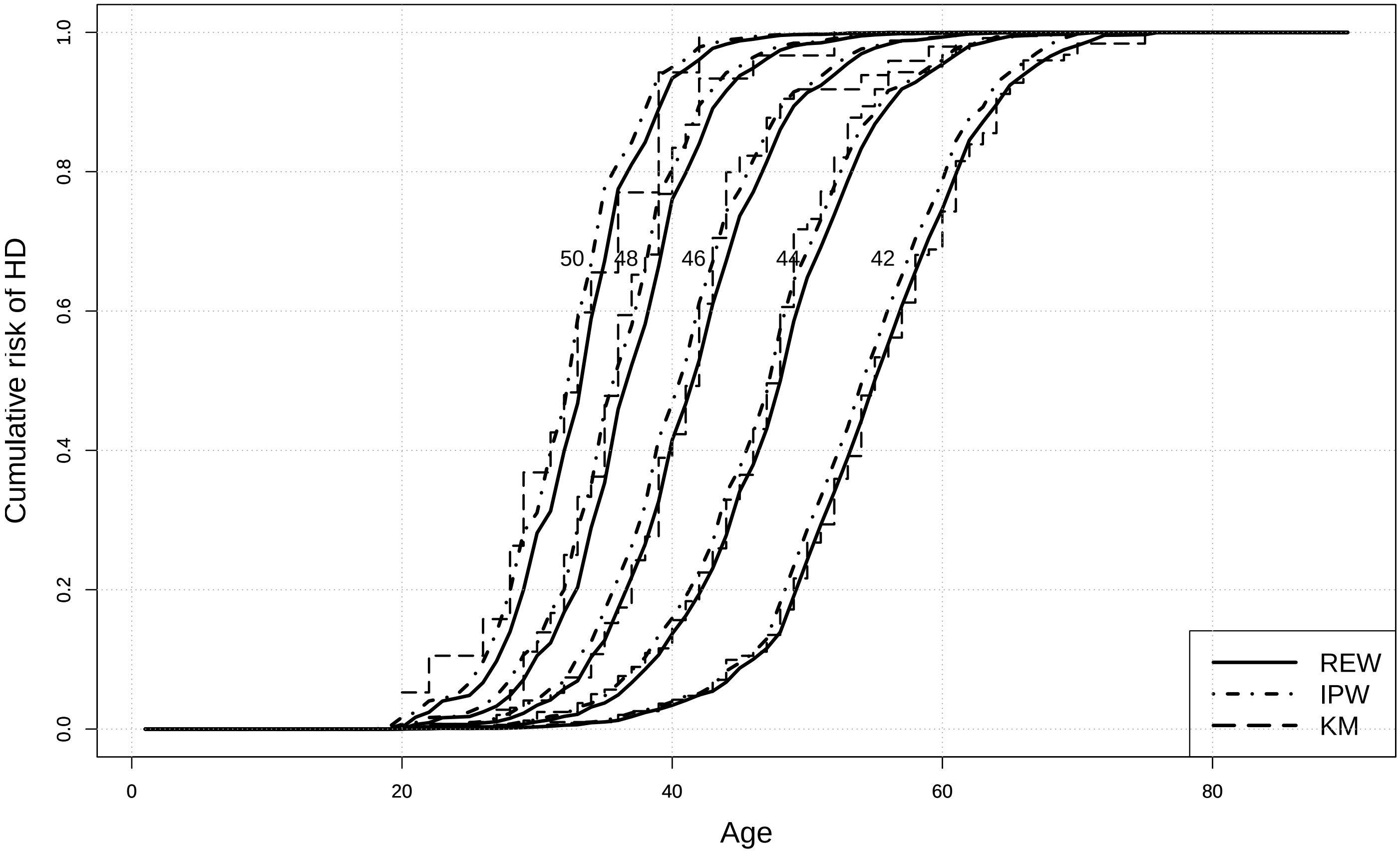


Figure 2: Estimated CDF curves (KM, IPW and REW) on COHORT proband data ($n = 1151$) evaluated at CAG=50, 48, 46, 44, 42 (left to right).


Cumulative risk of HD
Age
REW
IPW
KM