Nonparametric estimation for censored mixture data with application to the Cooperative Huntington's Observational Research Trial

Yuanjia Wang, Tanya P. Garcia and Yanyuan Ma

Abstract

This work presents methods for estimating genotype-specific distributions from genetic epidemiology studies where the event times are subject to right censoring, the genotypes are not directly observed, and the data arise from a mixture of scientifically meaningful subpopulations. Examples of such studies include kin-cohort studies and quantitative trait locus (QTL) studies. Current methods for analyzing censored mixture data include two types of nonparametric maximum likelihood estimators (NPMLEs) which do not make parametric assumptions on the genotype-specific density functions. Although both NPMLEs are commonly used, we show that one is inefficient and the other inconsistent. To overcome these deficiencies, we propose three classes of consistent nonparametric estimators which do not assume parametric density models and are easy to implement. They are based on the inverse probability weighting (IPW), augmented IPW (AIPW), and nonparametric imputation (IMP). The AIPW achieves the efficiency bound without additional modeling assumptions. Extensive simulation experiments demonstrate satisfactory performance of these estimators even when the data are heavily censored. We apply these estimators to the Cooperative Huntington's Observational Research Trial (COHORT), and provide age-specific estimates of the effect of mutation in the Huntington gene on mortality using a sample of family members. The close approximation of the estimated non-carrier survival rates to that of the U.S. population indicates small ascertainment bias in the COHORT family sample. Our analyses underscore an elevated risk of death in Huntington gene mutation carriers compared to non-carriers for a wide age range, and suggest that the mutation equally affects survival rates in both genders. The estimated survival rates are useful in genetic counseling for providing guidelines on interpreting the risk of death associated with a positive genetic testing, and in facilitating future subjects at risk to make informed decisions on whether to undergo genetic mutation testings.

Some Key Words: Censored data; Finite mixture model; Huntington's disease; Kin-cohort design; Quantitative trait locus

Short title: Analysis of Censored Mixture Data

Yuanjia Wang is Assistant Professor and correspondence author (E-mail: yuanjia.wang@columbia.edu), Department of Biostatistics, Columbia University, New York, NY 10032. Tanya P. Garcia is Post-Doctoral Fellow (E-mail: tpgarcia@stat.tamu.edu). Yanyuan Ma is Professor (E-mail: ma@stat.tamu.edu), Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143. This research is supported in part by the National GEM Consortium, P.E.O. Scholarship Award, and grants from the U.S. National Science Foundation (0906341 and 1206693), the U.S. National Institute of Health (NS073671-01, AG031113-01A2), and the National Cancer Institute (R25T-CA090301). Samples and data from the COHORT Study, which receives support from HP Therapeutics, Inc., were used in this study. We thank the Huntington Study Group COHORT investigators and coordinators who collected data and/or samples used in this study, as well as participants and their families who made this work possible.

1 Introduction

In some genetic epidemiology studies, a research goal is to estimate genotype-specific cumulative distributions of an outcome from mixture data of scientifically meaningful subpopulations where genotypes are not directly observed. Examples of such studies include kin-cohort studies (Struewing et al. 1997; Wacholder et al. 1998; Wang et al. 2008; Mai et al. 2009) and quantitative trait locus (QTL) studies (Lander and Botstein 1989; Wu et al. 2007). In kin-cohort studies, scientists sample and genotype an initial cohort of subjects (probands), possibly enriched with mutation carriers. They then collect family history of the disease (phenotype) from family members of the probands through systematic and validated interviews of the probands (Marder et al. 2003). While it is impractical and costly to interview family members in-person to collect their blood samples and obtain genotypes, it is possible to calculate the probability of each relative having a certain genotype based on the relationship with the proband and the proband's genotype. Thus, kin-cohort studies differ from other types of case-control family studies (Li et al. 1998) in that genetic information in family members is not readily available. Distributions of the observed phenotypes in the relatives are therefore a mixture of genotype-specific distributions.

In the interval mapping of quantitative traits (Lander and Botstein 1989), the genotype of a quantitative trait locus (QTL) is not observed, so trait distributions are mixtures of the QTL genotype-specific distributions. The mixing proportions are computed based on the observed flanking marker genotypes and recombination fractions between the marker and the putative QTL. In many controlled QTL experiments such as backcross or intercross, the mixing proportions can be easily obtained, and interest is in estimating the genotype-specific distributions.

The unobserved genotype information in both kin-cohort and QTL studies makes inference of genotype-specific distributions difficult. Inference is further complicated by right censoring as patients in the study may drop out or become lost to follow-up. The focus of the current paper is to develop simple, robust, and efficient estimators to improve upon the available methods in the literature for analyzing such censored mixture data.

Many statistical methods have been developed for modeling and analyzing censored mixture data in QTL mappings and kin-cohort studies. Sometimes the biological underpinning of the development of a disease trait suggests a suitable parametric function which offers meaningful interpretation of the biological structure (Wu et al. 2000). In these cases, it is reasonable to use maximum likelihood based parametric methods (Lander and Botstein 1989; Wu et al. 2002). In some QTL experiments, a semiparametric Cox proportional hazards model may also be suitable (Diao and Lin 2005; Zeng and Lin 2007), but a proportional hazards assumption is not always valid, such as in some applications with Huntington's disease (HD) data (Langbehn et al. 2004). In fact, in many situations, there may not be sufficient biological knowledge to warrant particular parametric or semiparametric models, hence concerns of model mis-specification naturally arise. To alleviate these issues, more flexible nonparametric estimation of the distribution functions become essential (Zhao and Wu 2008; Yu and Lin 2008). Throughout this work, the term "nonparametric" refers to leaving the probability density (or hazard) functions completely unspecified.

For the QTL data, Fine et al. (2004) developed a nonparametric method which exploits the property of independence between the censoring and the event of interest. Wang et al. (2007) proposed a nonparametric method for kin-cohort data when the censoring times are observed for all subjects. When censoring times are random and not all observed, Wacholder et al. (1998) proposed a nonparametric maximum likelihood estimator (type I NPMLE) consisting of a combination of several NPMLEs and a linear transformation. Chatterjee and Wacholder (2001) proposed a direct maximization of the nonparametric likelihood (type II NPMLE) with respect to the conditional distributions and used an Expectation-Maximization (EM) algorithm to find the maximizer. Although in many situations NPMLEs are consistent and even efficient, we demonstrate the surprising result that the type I is highly inefficient and the type II is inconsistent.

To overcome the shortcomings of the aforementioned methods, we provide several consistent and efficient nonparametric estimators by casting this problem in a missing data framework. Given a complete data influence function when there is no censoring (see Appendix and Ma and Wang 2012), we propose an inverse probability weighting (IPW) estimator, and derive an optimal augmentation term to obtain the optimal estimator. We demonstrate that the optimal augmented IPW (AIPW) estimator achieves the efficiency bound without extra modeling assumptions or complicated computational procedures. We also propose an imputation (IMP) estimator which is easy to implement and does not require additional modeling assumptions for the imputation step.

The rest of the paper is organized as follows. Section 1.1 presents a large collaborative study of Huntington's disease to which we apply our proposed estimators. Section 2 describes the inefficiency and inconsistency of the two existing NPMLE methods. To improve upon these methods, we propose several nonparametric estimators in Section 3 which are consistent, efficient, and easy to implement. We demonstrate the asymptotic properties of these estimators and examine their finite sample performance through comprehensive simulation studies in Section 4. The methods are applied to the Huntington's disease study in Section 5, and Section 6 concludes the paper with some discussions. The technical details and additional numerical results are in an Appendix and an online Supplementary Material, with tables, figures and section numbers in the Supplementary Material indicated with "S." proceeding a number.

1.1 The Cooperative Huntington's Observational Research Trial (COHORT)

Huntington's disease is a degenerative, genetic disorder which targets nerve cells in the brain and leads to cognitive decline, involuntary muscle spasms, and psychological problems. Affected individuals typically begin to see neurological and physical symptoms around 30-50 years of age, and eventually die from pneumonia, heart failure or other complications 15-20 years after the disease onset (Foroud et al. 1999). The severity of the disease has prompted the development of several organizations, like the Huntington Study Group (Huntington Study Group 2011), which are devoted to studying the causes, effects, and possible treatments for HD. A particular study organized by roughly 42 Huntington Study Group research centers in North America and Australia is the Cooperative Huntington's Observational Research Trial (COHORT; Dorsey et al. 2008). Since 2005, the principal investigators of COHORT have been collecting ongoing information from affected or at-risk adults and their family members 15 years of age and older.

Huntington's disease is caused by unstable CAG repeats expansion in the HD gene (Huntingtons Disease Collaborative Research Group, 1993). In a genetic counseling setting, CAG repeats $\geqslant 36$ is defined as positive for HD gene mutation, or carrier, and CAG < 36 is defined as negative, or non-carrier (Rubinsztein et al. 1996). Proband participants in COHORT undergo a clinical evaluation where blood samples are genotyped for being a carrier or non-carrier of HD mutation. While the HD mutation status is ascertained in probands, high costs of in-person interviews on family members prevents collection of their blood samples. Family members' morbidity and mortality information such as age-at-death is obtained through a systematic interview of the probands. Although a relative's HD gene mutation status is unavailable, the probability of carrying a mutation can still be obtained based on the relative's relationship with the proband and the proband's mutation status (Khoury et al. 1993, Section 8.4). The distribution of the relative's age-at-death is therefore a mixture of the genotype-specific distributions with known, subject-specific mixing proportions.

Despite the identification of the causative gene, there is currently no effective treatment that modifies HD progression. One of the goals of COHORT is to estimate the risks of adverse events, such as disease onset or death, associated with carrying a mutation, and to use these parameters to design future clinical trials for intervention or treatment of HD. For example, the power calculation of a trial with survival as the primary endpoint will depend on parameters such as risk ratio in carriers and non-carriers. The proposed methods here can aid in estimating these important parameters, and also has benefits in genetic counseling for patients and their family members. The estimated survival function in HD mutation carriers provides guidelines on interpreting the risk of death associated with a positive genetic mutation test, and facilitate subjects at risk to make important life decisions such as marriage or having children. We show some examples of the utilities of the survival estimates in Section 5.

2 Some existing nonparametric estimators for censored mixture data

We consider censored mixture data denoted as triplets $(\boldsymbol{Q}_i = \boldsymbol{q}_i, X_i = x_i, \Delta_i = \delta_i)$, which are independent and identically distributed. For the ith subject, $\boldsymbol{Q}_i$ is a p-dimensional vector of the random mixture proportions computed from available genotype data on the proband in kin-cohort studies or from flanking markers in QTL studies. In a kin-cohort study, $\boldsymbol{Q}_i$ may be a two-dimensional vector where Q_{i1} represents the probability of being a mutation carrier, and Q_{i2} a non-carrier. To illustrate the computation of $\boldsymbol{Q}_i$, let L_i denote the unobserved genotype in a relative, and let L_{i0} denote the observed genotype in a proband.

Let p_A denote the population frequency of the mutation allele A and let a denote the wild type. Consider a heterozygous carrier proband with genotype Aa. Assuming Mendelian transmission, the probability of a parent of a proband being a carrier is $Q_{i1} = \Pr(L_i = AA \text{ or } Aa | L_{i0} = Aa) = \frac{1}{2}(1 + p_A)$. The probability for a sibling of a proband to be a carrier is $Q_{i1} = \Pr(L_i = AA \text{ or } Aa | L_{i0} = Aa) = -\frac{1}{4}p_A^2 + \frac{3}{4}p_A + \frac{1}{2}$. When $p_A \approx 0$, the two Q_{i1}'s are both $\frac{1}{2}$. The $\boldsymbol{Q}_i$ for other types of relatives and other types of probands (homozygous or non-carrier probands) are computed similarly; see Khoury et al. (1993) (Section 8.4) for details.

In general, for both QTL and kin-cohort studies, $\boldsymbol{Q}_i$ has a discrete distribution with a finite support, say $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$. Its probability mass function, denoted as p_Q, is determined by the experimental design. For example, in a backcross QTL study, $\boldsymbol{Q}_i$ is a two-dimensional vector that will take four possible values $(1, 0)^T$, $(\theta, 1-\theta)^T$, $(1-\theta, \theta)^T$, and $(0, 1)^T$, where θ is the known recombination fraction between the putative QTL and the flanking marker. The probability of $\boldsymbol{Q}_i$ taking these four values is determined by the marker genotype frequencies computed from the observed marker data (e.g., Table 10.4 in Wu et al. 2007). In kin-cohort studies, the distribution of $\boldsymbol{Q}_i$ is determined by the type of relatives collected (e.g., parents, siblings and children) and the distribution of the probands' genotypes (e.g., number of non-carriers probands, heterozygote probands and homozygous probands).

Lastly, $X_i = \min(T_i, C_i)$, where T_i is a subject's event time and C_i is a random continuous censoring time independent of T_i; and $\Delta_i = I(T_i \leqslant C_i)$ is the censoring indicator. We let $\boldsymbol{f}(\cdot)$ denote the p-dimensional unspecified conditional probability density function of T given genotypes in p genotype groups, and $\boldsymbol{F}(\cdot)$ denote the corresponding cumulative distribution function. Interest lies in estimating $\boldsymbol{F}(t)$ for any fixed time t. In the COHORT study, we have $p = 2$ with $F_1(t)$ and $F_2(t)$ corresponding to the age-at-death distribution for HD gene mutation carriers and non-carriers. Throughout, except when specifically pointed out, we assume that the event times $x_1, \ldots, x_n$ have no ties, and that the censoring distribution is common for all subjects. Then, letting $G(\cdot)$ denote the survival function of C and $g(\cdot)$ its corresponding density, the log-likelihood of n observations is

$$\sum_{i=1}^{n} \log\left(p_{\boldsymbol{Q}}(\boldsymbol{q}_i)\left\{\boldsymbol{q}_i^T f(x_i)G(x_i)\right\}^{\delta_i}\left[\left\{1 - q_i^T \boldsymbol{F}(x_i)\right\}g(x_i)\right\}\right]^{1-\delta_i}\right), \tag{1}$$

where we use the fact that $\boldsymbol{q}_i^T \mathbf{1}_p = 1$ with $\mathbf{1}_p$ a p-dimensional vector of ones.

2.1 The type I NPMLE and its inefficiency

The type I NPMLE was proposed in the literature to analyze kin-cohort data (Wacholder et al. 1998). It first maximizes (1) with respect to $\boldsymbol{q}_i^T \boldsymbol{f}(x_i)$'s, then recovers $\boldsymbol{F}(t)$ through a linear transformation. Although an NPMLE based estimator is usually efficient, it is not so for mixture data, and the magnitude of efficiency loss can be large.

To describe the type I NPMLE, we reformulate the maximization problem by evoking the assumption that $\boldsymbol{Q}$ has finite support $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ and by letting $s_j(x_k) = \boldsymbol{u}_j^T \boldsymbol{f}(x_k)$ and $S_j(x_k) = 1 - \boldsymbol{u}_j^T \boldsymbol{F}(x_k)$. Throughout this work, we refer to the p different genotype populations as p subpopulations, and the m different $\boldsymbol{u}_j$ values as m subgroups. In the literature, the type I NPMLE assumes random censoring, and hence the censoring distribution does not contribute information to the parameter of interest. Therefore, ignoring $G(\cdot)$ and $g(\cdot)$ in (1), it maximizes the equivalent target function

$$\sum_{j=1}^{m} \sum_{i=1}^{n} \log \left\{ s_j(x_i)^{\delta_i} S_j(x_i)^{1-\delta_i} \right\} I(\boldsymbol{q}_i = \boldsymbol{u}_j) \tag{2}$$

with respect to $s_j(x_i)$'s and subject to $\sum_{i=1}^{n} s_j(x_i) I(\boldsymbol{q}_i = \boldsymbol{u}_j) \leqslant 1$, $s_j(x_i) \geqslant 0$ for $j = 1, \ldots, m$. This is equivalent to m separate maximization problems, each concerning $s_j(\cdot)$ and $S_j(\cdot)$ only, so that the maximizers are the classical Kaplan-Meier estimators. That is,

$$\widehat{S}_j(a) = \prod_{x_i \leqslant a, \boldsymbol{q}_i = \boldsymbol{u}_j} \left\{ 1 - \frac{\delta_i}{\sum_{\boldsymbol{q}_k = \boldsymbol{u}_j} I(x_k \geqslant x_i)} \right\}$$

and $s_j(a) = S_j(a^-) - S_j(a)$ for all a. Using the linear relation $\boldsymbol{u}_j^T \boldsymbol{F}(t) = 1 - S_j(t)$ for $j = 1, \ldots, m$, we then recover the type I NPMLE estimator as

$$\widetilde{\boldsymbol{F}}(t) = \left(\boldsymbol{U}^T \boldsymbol{U} \right)^{-1} \boldsymbol{U}^T \{ \mathbf{1}_m - \widehat{\boldsymbol{S}}(t) \},$$

where $\widehat{\boldsymbol{S}}(t) = \{\widehat{S}_1(t), \ldots, \widehat{S}_m(t)\}^T$, and $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m)^T$. In this notation, $\boldsymbol{S}(t) = \mathbf{1}_m - \boldsymbol{U}\boldsymbol{F}(t)$. The consistency of the Kaplan-Meier estimator of $\boldsymbol{S}(t)$ ensures the consistency of $\widetilde{\boldsymbol{F}}(t)$. The inefficiency of $\widetilde{\boldsymbol{F}}(t)$, however, is evident considering that $\widetilde{\boldsymbol{F}}_w(t) =$

$(\boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\{\mathbf{1}_m - \widehat{\boldsymbol{S}}(t)\}$ with $\boldsymbol{\Sigma}$ denoting the variance-covariance matrix of $\widehat{\boldsymbol{S}}(t)$ yields a more efficient estimator. In this case, $\boldsymbol{\Sigma}$ is a diagonal matrix because each of the m components of $\widehat{\boldsymbol{S}}(t)$ is estimated using a distinct subset of the observations. Hence, $\widetilde{\boldsymbol{F}}_w(t)$ is a weighted version of the type I NPMLE, and this simple weighting scheme improves the estimation efficiency.

2.2 The type II NPMLE and its inconsistency

The type II NPMLE is considered an improvement over the type I NPMLE (Chatterjee and Wacholder 2001). It maximizes the same log likelihood (1), but with respect to $f(x_i)$'s subject to $\sum_{i=1}^n \boldsymbol{f}(x_i) \leqslant \mathbf{1}_p$ and $\boldsymbol{f}(x_i) \geqslant \mathbf{0}$ component-wise. Like the type I NPMLE, the type II NPMLE also assumes random censoring and ignores $G(\cdot)$ and $g(\cdot)$ in (1). In general, no closed form solution exists, and the EM algorithm is implemented to obtain the $F(x_i)$'s. Specifically, we regard the genotypes $L_i = 1, \ldots, p$ as missing data and derive the complete data log likelihood of the observations $\boldsymbol{o}_i = (L_i = l_i, X_i = x_i, \Delta_i = \delta_i), i = 1, \ldots, n$, as

$$\mathcal{L}^{\text{comp}}_{\text{type II}}\{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_n; \boldsymbol{f}(x_i), \boldsymbol{F}(x_i), i = 1, \ldots, n\} = \sum_{i=1}^n \sum_{k=1}^p I(l_i = k)\log\left[f_k(x_i)^{\delta_i}\{1 - F_k(x_i)\}^{1-\delta_i}\right].$$

The EM algorithm is an iterative procedure. At the bth iteration, we take the conditional expectation of the complete data log-likelihood given the observed data (e.g., $\{(X_i = x_i, \Delta_i = \delta_i), i = 1, \ldots, n\}$), and update the E-step via

$$\begin{aligned} & E\left[\mathcal{L}^{\text{comp}}_{\text{type II}}\{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_n; \boldsymbol{f}(x_i), \boldsymbol{F}(x_i), i = 1, \ldots, n\} | \boldsymbol{f}^{(b)}(x_i), \boldsymbol{F}^{(b)}(x_i), i = 1 \ldots, n\right] \\ = & \sum_{k=1}^p \sum_{i=1}^n \left[\frac{\delta_i q_{ik} f_k^{(b)}(x_i)}{\sum_{k=1}^p q_{ik} f_k^{(b)}(x_i)} \log f_k(x_i) + \frac{(1-\delta_i) q_{ik}\{1 - F_k^{(b)}(x_i)\}}{\sum_{k=1}^p q_{ik}\{1 - F_k^{(b)}(x_i)\}} \log\{1 - F_k(x_i)\}\right]. \end{aligned}$$

The M-step maximizes the above expression with respect to $\boldsymbol{f}(x_i)$ and $\boldsymbol{F}(x_i)$'s subject to $\boldsymbol{f}(x_i) \geqslant \mathbf{0}$ and $\mathbf{1} \geqslant \boldsymbol{F}(x_i) \geqslant \mathbf{0}$. To this end, let

$$c_{ik}^{(b)} = \delta_i \frac{q_{ik} f_k^{(b)}(x_i)}{\sum_{k=1}^p q_{ik} f_k^{(b)}(x_i)} + (1 - \delta_i) \frac{q_{ik}\{1 - F_k^{(b)}(x_i)\}}{\sum_{k=1}^p q_{ik}\{1 - F_k^{(b)}(x_i)\}}$$

denote the known quantity based on the bth iteration. Then, the M-step reduces to p separate maximization problems of the form

$$\sum_{i=1}^{n} c_{ik}^{(b)} \left[\delta_i \log f_k(x_i) + (1-\delta_i)\log\{1 - F_k(x_i)\}\right],$$

for $k = 1, \ldots, p$. Viewing this as the log likelihood of weighted observations, where the ith observation represents $c_{ik}^{(b)}$ observations of the same value, the maximizer is a modified Kaplan-Meier estimator:

$$1 - \check{F}_k^{(b+1)}(t) = \prod_{x_i \leqslant t, \delta_i = 1} \left\{1 - \frac{\sum_{j=1}^{n} I(x_j = x_i, \delta_j = 1) c_{jk}^{(b)}}{\sum_{j=1}^{n} c_{jk}^{(b)} I(x_j \geqslant x_i)}\right\} = \prod_{x_i \leqslant t, \delta_i = 1} \left\{1 - \frac{c_{ik}^{(b)}}{\sum_{j=1}^{n} c_{jk}^{(b)} I(x_j \geqslant x_i)}\right\}.$$

Iterating the E- and the M-step until convergence leads to the type II estimator.

As natural as the type II NPMLE appears, we show in Section S.1 of the Supplementary Material the surprising result that it is an inconsistent estimator of $F(t)$.

3 Proposed nonparametric estimators for censored mixture data

3.1 The IPW and the optimal augmented IPW estimators

To compensate for the deficiencies of the NPMLEs, we propose a class of nonparametric estimators based on inverse probability weighting (IPW) and its augmented version which are consistent and easy to implement. We describe these estimators in terms of their corresponding influence functions.

3.1.1 Inverse probability weighting

The notion of IPW was first introduced by Horvitz and Thompson (1952) in the context of survey sampling, and later by Robins, Rotnitzky, and Zhao (1994) in the context of missing data as a means for upweighting subjects who are under-represented because of missingness. Bang and Tsiatis (2000, 2002) used the IPW to estimate the mean and median medical costs by capturing information from patients whose medical costs were subject to right censoring. In this spirit, we elicit information from the censored observations in the mixture data with an IPW estimator. Specifically, our IPW estimator solves

$$n^{-1}\sum_{i=1}^{n}\frac{\delta_i\boldsymbol{\phi}(\boldsymbol{q}_i,x_i)}{\widehat{G}(x_i)}=0, \tag{3}$$

where $\boldsymbol{\phi}$ denotes a general influence function for non-censored mixture data corresponding to $\delta_i = 1$ ($i = 1,\ldots,n$) in (1) (see Appendix for elaborations on ϕ), and $\widehat{G}(t)$ is the Kaplan-Meier estimator of $G(t)$:

$$\widehat{G}(t)=\prod_{x_i\leqslant t}\left\{1-\frac{1-\delta_i}{\sum_{j=1}^{n}I(x_j\geqslant x_i)}\right\}.$$

The intuition behind (3) is that for any subject randomly selected from the population with $X_i = x_i$, the probability that such a subject will not be censored is $G(x_i)$. Therefore, any uncensored subject with $X_i = x_i$ can be regarded as representing $1/G(x_i)$ subjects from the population. By inversely weighting all uncensored subjects with their corresponding probabilities of not being censored, we obtain a consistent estimating equation in (3).

We now characterize the asymptotic behavior of the IPW estimator in terms of its influence function. Let $Y_i(u) = I(X_i \geqslant u)$, $Y(u) = \sum_{i=1}^{n} Y_i(u)$, $N_i^c(u) = I(X_i \leqslant u, \Delta_i = 0)$, and $\lambda^c(\cdot)$ be the hazard function for the censoring distribution. Also, let

$$M_i^c(u)=N_i^c(u)-\int_0^u I(X_i\geqslant s)\lambda^c(s)ds$$

denote the censoring martingale; and define

$$\mathcal{B}(\boldsymbol{h},u)=E\left\{\boldsymbol{h}(\cdot)|T_i\geqslant u\right\}=\frac{E\{\boldsymbol{h}(\cdot)I(T_i\geqslant u)\}}{S(u)}$$

where $\boldsymbol{h}$ is any p-length function. Then, as derived in Section S.2 of the Supplementary Material, the ith influence function for the IPW estimator is

$$\boldsymbol{\phi}_{\text{ipw}}(\boldsymbol{q}_i,x_i,\delta_i)=\boldsymbol{\phi}(q_i,t_i)-\int\frac{dM_i^c(u)}{G(u)}\left\{\boldsymbol{\phi}(\boldsymbol{q}_i,t_i)-\mathcal{B}\left(\boldsymbol{\phi},u\right)\right\}.$$

The two terms in $\boldsymbol{\phi}_{\text{ipw}}$ are uncorrelated given that $\boldsymbol{\phi}(\boldsymbol{q}_i, x_i)$ is $\mathcal{F}(0)$ measurable where $\mathcal{F}(u)$ is a filtration defined by the set of σ-algebras generated by $\sigma\{\boldsymbol{q}_i, I(C_i \leqslant r), r \leqslant u; I(T_i \leqslant$

$x), 0 \leqslant x < \infty, i = 1, \ldots, n\}$. Hence, the estimation variance of the IPW estimator is

$$\boldsymbol{V}_{\text{ipw}} = \text{cov}\{\boldsymbol{\phi}(\boldsymbol{Q}_i, T_i)\} + E\left\{\int \frac{\mathcal{B}(\boldsymbol{\phi}^{\otimes 2}, u) - \mathcal{B}(\boldsymbol{\phi}, u)^{\otimes 2}}{G^2(u)} \lambda^c(u) Y_i(u) du\right\},$$

and a consistent estimator is

$$\widehat{\boldsymbol{V}}_{\text{ipw}} = n^{-1} \sum_{i=1}^{n} \frac{\delta_i \boldsymbol{\phi}(\boldsymbol{q}_i, x_i)\boldsymbol{\phi}^T(\boldsymbol{q}_i, x_i)}{\widehat{G}(x_i)} + n^{-1} \sum_{i=1}^{n} \int \frac{\widehat{\mathcal{B}}_1(\boldsymbol{\phi}^{\otimes 2}, u) - \widehat{\mathcal{B}}_1(\boldsymbol{\phi}, u)^{\otimes 2}}{\widehat{G}^2(u)} dN_i^c(u),$$

where $\widehat{\mathcal{B}}_1(\boldsymbol{h}, u) = \dfrac{1}{n\widehat{S}(u)} \sum_{i=1}^{n} \dfrac{\delta_i \boldsymbol{h}(\boldsymbol{q}_i, x_i, \delta_i) I(x_i \geqslant u)}{\widehat{G}(x_i)}$ for an arbitrary function $\boldsymbol{h}(\boldsymbol{q}_i, x_i, \delta_i)$.

3.1.2 Augmented inverse probability weighting

Although intuitive and easy to implement, the IPW estimator is inefficient. Instead, using a modification motivated by Robins and Rotnitzky (1992), one may adjust the IPW estimator to improve its efficiency. With $\boldsymbol{\phi}$ as the complete data influence function, Robins and Rotnitzky (1992) provided the following general class of influence functions for censored data:

$$\boldsymbol{\phi}(\boldsymbol{q}_i, t_i) - \int \frac{dM_i^c(u)}{G(u)} \{\boldsymbol{\phi}(\boldsymbol{q}_i, t_i) - \mathcal{B}(\boldsymbol{\phi}, u)\} + \int \frac{dM_i^c(u)}{G(u)} [\boldsymbol{h}\{\bar{\boldsymbol{a}}_i(u), u\} - \mathcal{B}(\boldsymbol{h}, u)]. \tag{4}$$

For our mixture data problem, $\boldsymbol{a}_i(u) = \{\boldsymbol{q}_i, I(u < T_i)\}$ and $\bar{\boldsymbol{a}}_i(u)$ contains the functions $\boldsymbol{a}_i(\tilde{u})$ for all $\tilde{u} \leqslant u$. Compared to the influence function for the IPW estimator, the estimator from (4) contains an augmentation term that may improve the estimation efficiency, and is thus termed the augmented IPW (AIPW) estimator. Among all the choices for $\boldsymbol{h}$, Robins et al. (1994) and Van der Laan and Hubbard (1998) showed that

$$\begin{aligned} & \boldsymbol{h}^*_{\text{eff}}\{\bar{\boldsymbol{a}}_i(u), u\} = E\{\boldsymbol{\phi}(\boldsymbol{Q}_i, T_i) | T_i \geqslant u, \bar{\boldsymbol{a}}_i(u)\} \\ = \; & \{I(u < X_i) + I(u = X_i, \delta_i = 0)\} E\{\phi(\boldsymbol{Q}_i, T_i) | \boldsymbol{q}_i, T_i \geqslant u\} + I(u = X_i)\delta_i \boldsymbol{\phi}(\boldsymbol{q}_i, u) \end{aligned}$$

with $u \leqslant X_i$ yields the optimal efficiency. Denoting $\boldsymbol{h}_{\text{eff},i}(u) = E\{\boldsymbol{\phi}(\boldsymbol{Q}_i, T_i) | \boldsymbol{q}_i, T_i \geqslant u\}$, we have that $\boldsymbol{h}^*_{\text{eff}}\{\bar{\boldsymbol{a}}_i(u), u\}$ and $\boldsymbol{h}_{\text{eff},i}(u)$ are identical except when $u = X_i$ and $\delta_i = 1$. The functional $\boldsymbol{h}^*_{\text{eff}}$ only appears in the censoring martingale integral, so using $h_{\text{eff},i}(u)$ instead of $\boldsymbol{h}^*_{\text{eff}}\{\bar{\boldsymbol{a}}_i(u), u\}$ yields the same influence function. This simplification is of great importance because otherwise the optimal $\boldsymbol{h}^*_{\text{eff}}$ is only an interesting but impractical theoretical result. For most problems, computing $\boldsymbol{h}^*_{\text{eff}}$ is nearly impossible and would require extra modeling assumptions which prevents the estimator from achieving the efficiency bound.

In our case, however, $\boldsymbol{h}^*_{\text{eff}}$ is simple to compute and the AIPW estimator achieves the optimal efficiency. A consistent estimate uses a sample version of (4) with IPW:

$$\widehat{\boldsymbol{h}}_{\text{eff},i}(u) = \frac{\sum_{j=1}^n I(\boldsymbol{q}_j = \boldsymbol{q}_i)\boldsymbol{\phi}(\boldsymbol{q}_j, x_j)Y_j(u)\delta_j/\widehat{G}(x_j)}{\sum_{j=1}^n I(\boldsymbol{q}_j = \boldsymbol{q}_i)Y_j(u)\delta_j/\widehat{G}(x_j)}. \tag{5}$$

Because $\boldsymbol{h}_{\text{eff},i}(u)$ is not a function of C_i, the independence between the censoring and survival process gives

$$\mathcal{B}(\boldsymbol{h}_{\text{eff}}, u) \quad = \quad \frac{E\left\{\boldsymbol{h}_{\text{eff},i}(u)I(T_i \geqslant u)I(C_i \geqslant u)\right\}}{E\{I(T_i \geqslant u)I(C_i \geqslant u)\}} = \frac{E\left\{\boldsymbol{h}_{\text{eff},i}(u)Y_i(u)\right\}}{E\{Y_i(u)\}}.$$

Therefore, we can approximate $\mathcal{B}(\boldsymbol{h}_{\text{eff}}, u)$ with

$$\widehat{\mathcal{B}}(\boldsymbol{h}_{\text{eff}}, u) = \frac{\sum_{i=1}^n \boldsymbol{h}_{\text{eff},i}(u)Y_i(u)}{Y(u)},$$

which satisfies

$$\sum_{i=1}^n \int \frac{\lambda^c(u)Y_i(u)}{\widehat{G}(u)}\left\{\widehat{\boldsymbol{h}}_{\text{eff},i}(u) - \widehat{\mathcal{B}}(\widehat{\boldsymbol{h}}_{\text{eff}}, u)\right\}du = 0.$$

This enables us to obtain the optimal AIPW estimator $\widehat{F}(t)$ by solving

$$\sum_{i=1}^n \left[\frac{\delta_i \boldsymbol{\phi}(\boldsymbol{q}_i, x_i)}{\widehat{G}(x_i)} + \int \frac{dN_i^c(u)}{\widehat{G}(u)}\left\{\widehat{\boldsymbol{h}}_{\text{eff},i}(u) - \widehat{\mathcal{B}}\left(\widehat{\boldsymbol{h}}_{\text{eff}}, u\right)\right\}\right] = 0. \tag{6}$$

The AIPW estimator is very easy to implement especially comparing to many other non-parametric or semiparametric problems where the efficient estimator often involves additional model assumptions (Tsiatis and Ma 2004), solving integral equations (Rabinowitz 2000) and iterative procedures (Zhang et al. 2008).

In Section S.3 of the online Supplementary Material, we demonstrate that the AIPW estimator indeed has the efficient influence function, which corresponds to replacing $h(\cdot)$

with $\boldsymbol{h}_{\text{eff},i}(u)$ in (4). The variance of the efficient estimator is

$$\boldsymbol{V}_{\text{eff}} = \text{cov}\{\boldsymbol{\phi}(\boldsymbol{Q}_i, T_i)\} + E\int \frac{\mathcal{B}\{(\boldsymbol{\phi} - \boldsymbol{h}_{\text{eff}})^{\otimes 2}, u\}}{G^2(u)}\lambda^c(u)Y_i(u)du,$$

which is estimated consistently by

$$\widehat{\boldsymbol{V}}_{\text{eff}} = n^{-1}\sum_{i=1}^{n} \frac{\delta_i \boldsymbol{\phi}(\boldsymbol{q}_i, x_i)\boldsymbol{\phi}^T(\boldsymbol{q}_i, x_i)}{\widehat{G}(x_i)} + n^{-1}\sum_{i=1}^{n}\int \frac{\widehat{\mathcal{B}}_1\{(\boldsymbol{\phi} - \widehat{\boldsymbol{h}}_{\text{eff}})^{\otimes 2}, u\}}{\widehat{G}^2(u)} dN_i^c(u).$$

3.1.3 Subgroup-specific censoring

The IPW estimator (3) and the AIPW estimator (6) are designed for the case when the censoring distribution $G(\cdot)$ is common for all subjects in m subgroups. When this is not the case, subgroup-specific censoring distributions, $\widetilde{G}_j(t), j = 1, \ldots, m$, should be used. Specifically, $\widehat{G}(t)$ is replaced by the subgroup-specific Kaplan-Meier estimators

$$\widetilde{G}_j(t) = \prod_{\substack{x_i \leqslant t \\ \boldsymbol{q}_i = \boldsymbol{u}_j}} \left\{1 - \frac{1 - \delta_i}{\sum_{\boldsymbol{q}_k = \boldsymbol{u}_j} I(x_k \geqslant x_i)}\right\}$$

for $j = 1, \ldots, m$. Consequently, the IPW estimating equation (3) changes to

$$n^{-1}\sum_{j=1}^{m}\sum_{i:\boldsymbol{q}_i = \boldsymbol{u}_j} \frac{\delta_i \boldsymbol{\phi}(\boldsymbol{q}_i, x_i)}{\widetilde{G}_j(x_i)} = 0,$$

and the corresponding estimated variance $\widehat{V}_{\text{ipw}}$ changes analogously with summation $\sum_{j=1}^{m}\sum_{i:\boldsymbol{q}_i=\boldsymbol{u}_j}$ replacing $\sum_{i=1}^{n}$, and a subgroup-specific

$$\widehat{\mathcal{B}}_{1j}(\boldsymbol{h}, u) = \frac{1}{\#\{i : \boldsymbol{q}_i = \boldsymbol{u}_j\}\widehat{S}_j(u)} \sum_{i:\boldsymbol{q}_i = \boldsymbol{u}_j} \frac{\delta_i \boldsymbol{h}(\boldsymbol{q}_i, x_i, \delta_i) I(x_i \geqslant u)}{\widetilde{G}_j(x_i)}$$

replacing the pooled $\widehat{\mathcal{B}}_1(\boldsymbol{h}, u)$.

Similar changes apply to the AIPW estimator. Specifically, in (5), (6) and in the expression of $\widehat{\boldsymbol{V}}_{\text{eff}}$, we replace $\sum_{i=1}^{n}$ with $\sum_{j=1}^{m}\sum_{i:\boldsymbol{q}_i=\boldsymbol{u}_j}$, replace $\widehat{G}(t)$ with $\widetilde{G}_j(t)$, and replace $\widehat{\mathcal{B}}(\boldsymbol{h}_{\text{eff}}, u)$ with

$$\widehat{\mathcal{B}}_j(\boldsymbol{h}_{\text{eff}}, u) = \frac{\sum_{i:\ \boldsymbol{q}_i = \boldsymbol{u}_j} \boldsymbol{h}_{\text{eff},i}(u) Y_i(u)}{\sum_{i:\ \boldsymbol{q}_i = \boldsymbol{u}_j} Y_i(u)}.$$

It is worth noting that if we erroneously treat the censoring distribution as common when in fact it is not, the IPW estimator will not be consistent any more because the corresponding influence function no longer has mean zero. On the other hand, the AIPW estimator will still be consistent, although less efficient. This is a direct consequence of the double robustness property of the AIPW estimator, in that the validity of the full data influence function $\boldsymbol{\phi}(\boldsymbol{q}, t)$ alone guarantees consistency of the AIPW estimator. However, since the efficiency of AIPW is achieved when the correct censoring model is used, treating the censoring distribution as identical across subgroups when they are not leads to efficiency loss. The issue of subgroup-specific censoring and the performance of IPW, AIPW and their modified versions are illustrated in simulation studies in Section 4.3.

3.2 An imputation (IMP) estimator

Lipsitz et al. (1999) proposed a conditional estimating equation for regression with missing covariates by conditioning the complete data estimating equation on the observed data. Similarly, with censored observations, we replace the unknown complete data influence function with its conditional expectation given that the event happens after the observed censoring time. Doing so yields the following imputed estimating equation:

$$0 = \sum_{i=1}^{n} \left[\delta_i \boldsymbol{\phi}(\boldsymbol{q}_i, x_i) + (1 - \delta_i) E\left\{\boldsymbol{\phi}(\boldsymbol{Q}_i, T_i) | T_i > x_i, \boldsymbol{q}_i\right\}\right] = \sum_{i=1}^{n} \left\{\delta_i \boldsymbol{\phi}(\boldsymbol{q}_i, x_i) + (1 - \delta_i) \boldsymbol{h}_{\mathrm{eff},i}(x_i)\right\}.$$

In practice, with $\widehat{\boldsymbol{h}}_{\mathrm{eff},i}(\cdot)$ as in (5), we obtain the IMP estimator by solving

$$0 = \sum_{i=1}^{n} \left\{\delta_i \boldsymbol{\phi}(\boldsymbol{q}_i, x_i) + (1 - \delta_i) \widehat{\boldsymbol{h}}_{\mathrm{eff},i}(x_i)\right\}.$$

While in many cases the imputation method could lead to bias if the model of the missingness is mis-specified, it is straightforward to see that our proposed imputation estimator is always consistent. In practice, we often, but not always, observe that it performs competitively in comparison with the optimal AIPW estimator. For inferences, we derive the influence function of the IMP estimator in Section S.4 of the Supplementary Material and show that it has a complex form containing nested conditional expectations and hence is hardly useful practically. Asymptotic analysis for imputation based estimation is often complex and can be rather involved even in parametric imputation procedures (Wang and Robins 1998; Robins and Wang 2000), which partially explains why the bootstrap method is usually favored in its inference.

4 Simulations

4.1 Simulation design

We conducted comprehensive Monte Carlo simulation studies to illustrate the finite sample performance of four groups of estimators, yielding a total of eleven different estimators. The first three groups of estimators include the IPW, optimal AIPW and IMP estimators based on (a) the complete data ordinary least square (OLS) influence function; (b) the complete data weighted least square (WLS) influence function using the inverse of the variance as weights; and (c) the efficient (EFF) influence function. See the Appendix for the exact forms of the influence functions. The fourth group of estimators contains the two NPMLEs.

The primary goal of the simulation studies is to compare the bias and efficiency of the eleven estimators of the distribution function of an outcome subject to censoring in each genotype population without directly observing the genotypes. In the first two simulation studies, the distribution function $\boldsymbol{F}(t)$ is a two-dimensional vector (i.e., $p = 2$ subpopulations). In the first simulation experiment, we set $F_1(t) = \{1 - \exp(-t/4)\}/\{1 - \exp(-2.5)\}$ on the interval $(0, 10)$ and $F_2(t) = F_1(t)^{0.98}$ on the interval $(0, 5)$. In the second experiment, we set $F_1(t) = [\{1 - \exp(-t/4)\}/\{1 - \exp(-2.5)\}]^{0.5}$ on $(0, 10)$ and $F_2(t) = \{1 - \exp(-t/2)\}/\{1 - \exp(-2.5)\}$ on $(0, 5)$. Thus, the distributions in the first experiment belong to the proportional hazards model family, while they do not in the second. In both experiments, we let each random mixture proportion $\boldsymbol{q}_i$ be one of $m = 4$ different possible vector values: $(1, 0)^T$, $(0.6, 0.4)^T$, $(0.2, 0.8)^T$ and $(0.16, 0.84)^T$. Our sample size was 500 and we generated a uniform censoring distribution to achieve moderate (20%) and high (50%) censoring rates.

Our third simulation experiment mimics the COHORT data. With $p = 2$, we set $F_1(t) = [1 + \exp\{-(t - 63)/7\}]^{-0.9}/0.995$ on $(0, 100)$, and $F_2(t) = 0.0007t$ on $(0, 53]$ and $F_2(t) = 0.022 + [1 + \exp\{-(t - 68)/7.5\}]^{-2}$ on $(53, 100)$. These distributions roughly mimic the estimated cumulative risk of death for HD gene mutation carriers and non-carriers, respectively, in the COHORT study (Figure 1). Analogous to the COHORT study, we used sample size $n = 4500$, generated $m = 6$ different mixture proportions: $(0,1)^T$, $(0.5,0.5)^T$, $(0.97,0.03)^T$, $(0.75,0.25)^T$, $(0.25,0.75)^T$, and $(1,0)^T$; and censored 65% of the observations with a uniformly distributed censoring process.

For each of the three experiments under different censoring rates, we evaluated all eleven estimators at different t values. First, we ran 1000 Monte Carlo simulations to evaluate the pointwise bias, defined as $\widehat{\boldsymbol{F}}(t) - \boldsymbol{F}(t)$, at $t = 2.5$ in the first experiment (Table 1), at $t = 1.5$ in the second experiment (Table S.1, Supplementary Material), and at $t = 70$ in the third experiment (Table S.2, Supplementary Material). The corresponding estimated standard errors for the IPW and AIPW estimators were based on $\widehat{\boldsymbol{V}}_{\text{ipw}}$ and $\widehat{\boldsymbol{V}}_{\text{eff}}$ given in Section 3.1.1 and Section 3.1.2, respectively, whereas bootstrap estimates were used to quantify the variability for the IMP estimator and the NPMLEs. Next, we evaluated the biases of the estimators across the entire range of t values through an integrated absolute bias (IAB), defined as $\int_0^\infty |\bar{F}_j(t) - F_j(t)|dt$, $j = 1, 2$, where $\bar{F}_j(t)$ is the average estimated curves over multiple data sets. The results are in Table S.3 (upper half) and Tables S.2, S.3 (upper half) of the Supplementary Material. In the first two experiments, the IAB was computed on a grid set with an increment of $\Delta = 0.1$ as $\sum_{i=1}^{100} |\bar{F}_1(x_i) - F_1(x_i)|\Delta$ and $\sum_{i=1}^{50} |\bar{F}_2(x_i) - F_2(x_i)|\Delta$, where $\bar{F}_j(x_i)$ $(j = 1, 2)$ denotes the mean estimated distribution from 1000 data sets. In the third experiment, it was computed on a grid set with an increment of $\Delta = 2$ on (0,100) as $\sum_{i=1}^{50} |\bar{F}_j(x_i) - F_j(x_i)|\Delta$ $(j = 1, 2)$, where $\bar{F}_j(x_i)$ denotes the mean estimated distribution from 250 data sets.

4.2 Simulation results

The results in Table 1 and Tables S.1 and S.2 (Supplementary Material) indicate that all the nonparametric estimators we propose have ignorable finite sample biases, while the type II NPMLE has much larger bias. At high censoring rates, the bias for the type II NPMLE is much greater and the coverage probability is much lower than the pre-specified nominal level. Moreover, the bias is not a finite sample effect since even at a sample size of $n = 4500$, the bias persists. Despite its asymptotic consistency, the type I NPMLE also shows substantial bias when the censoring rate increases. This is because in the estimation procedure of the type I NPMLE, the mixture nature of the model is not taken advantage of at the maximization step. The Kaplan-Meier estimation in some subgroups could be based on very small sample sizes which can make the overall estimation unreliable.

Compared to the proposed estimators, the type I NPMLE has, for the most part, larger estimation variability, and the increased variability is rather substantial for the $F_2(t)$ estimation. In particular, the WLS based estimators have much less variability than the type I NPMLE. Ma and Wang (2012) showed that the type I NPMLE for uncensored data belongs to the WLS family with weights $w_i = 1/n_i$, where n_i is the number of subjects who share the same $\boldsymbol{q}_i$. In other words, the type I NPMLE essentially downweights individuals belonging to large subgroups. Since such a weighting strategy is highly undesirable, it is not surprising to see that the weights in the WLS provide an improvement over those in the type I NPMLE.

In contrast, the three proposed nonparametric estimators have satisfactorily small biases and are more efficient compared to the type I NPMLE. The optimal AIPW and IMP estimators both improve upon IPW in terms of estimation efficiency. When the censoring rate is moderate, IMP and AIPW perform similarly, while when the censoring rate increases, the superiority of the optimal AIPW over IMP becomes more notable. The similarity of the results in the first three groups of estimators suggests that the estimation efficiency is not sensitive to the choice of the non-censored data influence function $\boldsymbol{\phi}$. The same insensitivity of estimation efficiency to the choice of influence function is also observed for the non-censored data case (Ma and Wang 2012). This phenomenon proves beneficial especially in the censored data analysis since Robins and Rotnitzky (1992) remarked that the best complete data influence function does not necessarily yield an optimal censored data influence function, and finding the optimal member usually requires a computationally intensive procedure. Finally, the estimated standard error matches reasonably well with the empirical standard error, while the 95% coverage probability is close to the nominal level.

In Figure 1, we examine the entire estimated survival curve $\mathbf{1} - \widehat{\boldsymbol{F}}(t)$ in experiment three which mimics the COHORT study. In particular, we display results from the imputation estimator (EFFIMP) and the AIPW estimator (EFFAIPW), both based on the full data efficient influence function, as representatives of the proposed estimators, and compare them with the two NPMLEs. To evaluate these estimators, we plotted the true survival curves and the mean estimated survival curves from 250 simulated data sets along with the 95% pointwise confidence bands. The EFFIMP and EFFAIPW estimators perform satisfactorily throughout the entire range of t, while the type I NPMLE starts to exhibit large variability and small sample estimation bias as time progresses. This confirms our previous observation that the type I NPMLE suffers from the small subgroup sample size difficulty and the instability of the Kaplan-Meier estimation procedure near the maximum event time. The type II NPMLE also shows a non-ignorable bias for a wide range of t's.

To illustrate the overall bias of the eleven estimators across the entire range of t values, we further provide the IAB for all three experiments in Table S.3 (upper half) and Tables S.2, S.3 (upper half). Nearly all estimators have very small IAB, whereas the type II NPMLE can yield a bias ten times larger than the other estimators. For the same estimator in each experiment, the IAB also tends to increase with larger censoring rate.

4.3 Subgroup-specific censoring

We now examine the performance of the original IPW and AIPW estimators proposed in Sections 3.1.1 and 3.1.2, as well as their modified versions in Section 3.1.3, when the true censoring distribution is different across different subgroups. We extend the first simulation by generating the censoring times from the proportional hazards distribution

$$G(t \mid \boldsymbol{q}_i) = 1 - \exp\{-\gamma_1 t^{\gamma_2} \exp(\gamma_3)\},$$

where we set $\gamma_3 = 0$ if $\boldsymbol{q}_i = (1, 0)^T$ or $\boldsymbol{q}_i = (0.6, 0.4)^T$, and set $\gamma_3 = 0.2$ if $\boldsymbol{q}_i = (0.2, 0.8)^T$ or $\boldsymbol{q}_i = (0.16, 0.84)^T$. Here, γ_1, γ_2 remain the same across the different subgroups and were chosen to achieve respectively moderate (20%) and high (50%) censoring proportions. The survival times and $\boldsymbol{q}_i$ were generated as in the first simulation experiment in Section 4.1. We first implemented the original IPW and AIPW estimators, which ignore the subgroup-specific censoring pattern and simply use a pooled censoring distribution $\widehat{G}(t)$. We then implemented the modified IPW and AIPW estimators, which incorporate the subgroup-specific censoring distributions through obtaining $\widetilde{G}_j(t)$ as described in Section 3.1.3. As in our earlier analyses, we investigated the pointwise bias at $t = 2.5$, as well as the IAB across the entire range of t.

Table 3 shows the pointwise bias of the IPW and AIPW estimators when using a pooled estimated censoring distribution, and a subgroup-specific censoring distribution. It is evident that the IPW has substantial bias if the pooled censoring estimate is used, indicating that the original IPW is not applicable when the censoring is subgroup-specific. However, its bias is substantially reduced as soon as the subgroup-specific censoring is taken into account. The AIPW, on the other hand, is robust to mis-specification and has small bias regardless of whether a pooled or subgroup-specific censoring estimate is used. For the AIPW, ignoring the subgroup-specific censoring pattern only incurs a small efficiency loss, mostly for the OLS-based and WLS-based estimators. For the EFF-based AIPW, the efficiency loss is minimal.

The IAB (lower half of Table S.3) further indicates that when the pooled censoring distribution is used, the IPW has much larger overall bias than the AIPW. However, after incorporating the subgroup-specific censoring distribution, the modified IPW and AIPW have similar magnitudes of the IAB. In a separate analysis where we extended the second simulation experiment in Section 4.1, we also found similar behaviors for the pointwise bias (Table S.4) and IAB (lower half of Tables S.3).

5 Analysis of COHORT data

Data from the COHORT study consist of 4587 relatives of the proband participants who have different mixing proportions for being carriers or non-carriers of the HD gene mutation. Computation of these mixing proportions is discussed in Section 2, Wacholder et al. (1998) and Wang et al. (2008). The event time of interest is age of death, and roughly 68% of the data is censored. A main research interest is estimating the age-at-death distribution or the survival function for carriers and non-carriers to assess the effects of HD mutation on survival. The severity of Huntington's disease warrants that non-carriers tend to live longer, so we expect to see lower survival rates for the carrier group.

Since it is well known that survival rates differ by gender, we stratified the COHORT data by gender (2367 males and 2220 females), and analyzed the effects of HD gene mutation on the male and female subpopulations. The quantity of interest, $\mathbf{1}-\boldsymbol{F}(t)$, is a four-dimensional vector (i.e., $p=4$) where $1-F_1(t)$ and $1-F_2(t)$ denote the survival functions for male non-carriers and carriers respectively, and $1-F_3(t)$ and $1-F_4(t)$ denote analogous functions for females. Furthermore, the mixture proportions $\boldsymbol{Q}_i$ were four-dimensional vectors with the first two components corresponding to mixture proportions for male non-carriers and carriers, and the last two components for female non-carriers and carriers. To estimate $\mathbf{1} - \boldsymbol{F}(t)$, we implemented several theoretically consistent nonparametric estimators, including the type I NPMLE and the full data efficient influence function based AIPW estimator (EFFAIPW) as representatives of the already existing and newly proposed methods respectively.

To examine the performance of these estimators, we first compare the results for the male and female non-carrier groups to the general male and female US populations in 2003 (Arias 2006). These survival rates should be similar since the risk in non-carriers for both genders would reflect the general population if there is minimal ascertainment bias in family members. Figure S.1 and Tables 4, 5 (lowest panel) indicate that the EFFAIPW outperforms the type I NPMLE in capturing the behavior of the general male and female US populations. In fact, comparing the non-carrier female estimates and general female population, the EFFAIPW has an IAB less than half of that of the type I NPMLE. Likewise, for the non-carrier males, the EFFAIPW has an IAB about half of that of the type I NPMLE. Hence, the EFFAIPW appears to be a more reasonable estimator for analysis.

The EFFAIPW depicted in Figure S.1 shows a steep difference between the estimated survival rates of the carrier and non-carrier groups for both genders. In addition, the bottom left figure suggests that male carriers tend to have only slightly lower survival rates than female carriers. For example, at age 65, male carriers have a cumulative risk of death of 46.9% (95% CI: 40.9%, 53%) whereas female carriers have a cumulative risk of death of 43.0% (95% CI: 37.4%, 48.7%). This slight difference in combination with the overlapping 95% confidence band (not shown here) suggests that HD mutation affects males and females equally. The observed lack of gender effects from EFFAIPW agrees with some earlier studies which also did not find a gender effect in either the mean survival times of HD patients (Harper 1996), or the progression of Huntington's disease (Marder et al. 2000). In contrast, the type I NPMLE suggests that male carriers have much better survival rates than female carriers, and sometimes even slightly better survival rates than non-carriers, a behavior contradictory to the existing clinical literature.

The upper panel of Table 5 further presents the area under the survival curves, which can be interpreted as the expected years of life. Hence, the difference of areas under two survival curves represents the expected years of life lost of one compared to the other. Based on the EFFAIPW, the estimated expected years of life lost for mutation carriers compared to non-carriers is 9.06 years in males and 12.76 in females. In contrast, the type I NPMLE estimates longer expected years of life for male carriers, which is unreasonable.

Our further investigation reveals that the poor performance of the type I NPMLE on the COHORT data is partially due to small sample sizes in several $\boldsymbol{Q}_i$ subgroups. When we remove 211 subjects pertaining to three subgroups with sample sizes no more than 3% of the data, the behavior of the type I NPMLE improves (Figure S.1 in Supplementary Material). The improvement is reflected in both the IAB between non-carriers and the US population and the survival rates for carriers in both genders. However, there is still a large difference between the non-carrier female type I NPMLE estimate in terms of expected life lost compared to the US population (Table 5, middle panel).

Lastly, in Table 6 we present the estimated conditional probabilities of surviving the next several years in 5-year intervals for a subject alive at a given age. These probabilities are based on the EFFAIPW and are stratified by gender and mutation status. For example, a 35-year old female carrier has a 94.71% chance of surviving the next 10 years, and 81.76% chance of surviving the next 20 years, compared to 99.63% and 96.81% in a female non-carrier. Such probabilities are useful for patients when interpreting mutation testing results, and may help them make lifetime decisions such as having children.

In conclusion, using the more reliable EFFAIPW estimator, our analysis suggests that mutation carriers tend to have much lower survival rates than non-carriers, and the mutation equally affects survival rates in both genders. These survival rates are the first in the literature obtained from a sample of family members, and they highlight the deleterious effects of HD mutation on survival. The estimated survival rates in non-carriers closely resemble that of the US population, which illustrates minimal ascertainment bias and reflects the advantage of analyzing family members whose information is not used in the initial recruitment of probands.

6 Discussion

We propose two IPW based estimators and an IMP estimator for censored mixture data, among which the AIPW achieves the optimal efficiency based on a given complete data influence function. These estimators are easy to compute and do not involve any iterative procedures. When the sample size is small and the censoring rate is moderate, the IMP estimator can sometimes compete or even outperform the asymptotically optimal AIPW estimator. We also point out the surprising results of the inefficiency of the type I NPMLE and the inconsistency of the type II NPMLE proposed in the literature. Our finite sample simulations suggest that the efficiency loss of the type I NPMLE and the bias of the type II estimator can be quite substantial, and the finite sample bias of the type I NPMLE can be non-ignorable when the subgroup sample size is small or the estimation region is close to the upper end of the distribution support. Caution should be applied in interpreting inconsistency of the type II NPMLE, which occurs when a pure nonparametric model is used. Parametric models and semiparametric models such as the Cox proportional hazards model with a nonparametric baseline or piecewise exponential model are expected to be consistent (Zeng and Lin 2007).

In a special case when the data arise from a single distribution (i.e., $p = 1$), the IPW, AIPW, IMP and the two NPMLEs are all equivalent to the familiar Kaplan-Meier estimator. This indicates the complexity arising from the mixture nature. Through extensive simulation studies, we demonstrate that the proposed AIPW has satisfactory small bias and is more efficient than the type I NPMLE even when the censoring rate is high.

The optimal AIPW estimator also provides reasonable survival rate estimates for both genders and different mutation status in the COHORT study. Since genetic testing of HD mutation is commercially available, the estimated survival rates are useful in genetic counseling when a subject, with a family history of Huntington's disease, needs to decide on whether to undergo genetic testing. Understanding the mortality rates associated with a positive testing result may make the subject more inclined to determine his/her mutation status and seek treatments. In addition, in a future work it may be of interest to estimate the survival rates as a function of the number of CAG repeats in carriers.

In some genetic studies, the relatives are included through their probands, and there might be potential ascertainment bias. If the HD gene mutation carrier probands are randomly sampled from the population of all carriers, then the estimation from the relatives can be generalized to the population of all carriers. This corresponds to no ascertainment bias. However, when there is heterogeneity in the survival function of HD gene mutation carriers (e.g., there exists another gene influencing the survival function) and if the carrier probands are not a representative sample of the population of all HD mutation carriers, then estimation based on their relatives may be biased (Begg 2002). For example, if there is a second gene that decreases survival in HD mutation carriers, then over-sampling of probands with the second gene may lead to an upward bias of the risk of death, and under-sampling would lead to a downward bias. However, in the analysis of the COHORT data, the estimated survival function in non-carriers is reasonably close to the general population estimates obtained from the census data. This is an indication that the COHORT sample is not likely to be subject to severe ascertainment bias. Otherwise, the non-carrier distribution estimated from the COHORT relative data would be very different from the general population due to the distorted distribution of additional risk factors.

Finally, since the survival distribution is well known to be different between two genders in the general population, we carried out the COHORT analysis separately for each gender. It may be desirable to test the gender difference among the HD gene mutation carriers/non-carriers. This amounts to testing $H_0 : \boldsymbol{F}^1(t) = \boldsymbol{F}^2(t)$ at all t versus $H_1 : \boldsymbol{F}^1(t) \neq \boldsymbol{F}^2(t)$ for at least one t, where $\boldsymbol{F}^1(\cdot)$ is the vector of CDFs for male and $\boldsymbol{F}^2(\cdot)$ for female. Among various methods of performing the test, a convenient choice is permutation. Specifically, we compute the test statistic $v_0 = \sup_t ||\widehat{\boldsymbol{F}}^1(t) - \widehat{\boldsymbol{F}}^2(t)||$ from the observed data, where $|| \cdot ||$ is the L_2-norm. Since under the null hypothesis, the two genders have identical distributions, we can randomly permute the gender variable to obtain a permuted sample. Perform such permutation B times for some large B, and recompute the test statistic v_b based on the bth permuted sample, $b = 1, \ldots B$. The p-value is then $\sum_{b=1}^{B} I(v_b \geqslant v_0)/B$. If interest only lies in the gender difference in the carrier population, one may extract the corresponding component in $\boldsymbol{F}^1(t)$ and $\boldsymbol{F}^2(t)$ to perform the test.

Appendix: Influence function of consistent estimators with complete data

When there is no censoring (i.e., $\delta_i = 1$ for all i in (1)), Ma and Wang (2012) adopted a pure nonparametric model of the genotype-specific distributions without assuming any parametric form of the density function. They proposed a general class of consistent estimators and identified the efficient member of the class. The complete set of all influence functions of the consistent estimators for $F(t)$ can be characterized as (Ma and Wang 2012)

$$\left\{\boldsymbol{\phi}(\boldsymbol{q},s):\boldsymbol{\phi}(\boldsymbol{q},s)=\boldsymbol{d}(\boldsymbol{q},s)-\boldsymbol{F}(t)-\boldsymbol{B}\mathbf{1}_p,\int\boldsymbol{d}(\boldsymbol{Q},s)Q^T p_{\boldsymbol{Q}}(\boldsymbol{Q})d\mu(\boldsymbol{Q})=\boldsymbol{I}(s\leqslant t)\boldsymbol{I}_p+\boldsymbol{B}\right\},$$

where $\boldsymbol{I}_p$ is a p-dimensional identity matrix, $\boldsymbol{d}(\boldsymbol{q},s)$ is a vector of real functions (for qualified choices of $\boldsymbol{d}(\boldsymbol{q},s)$ see Ma and Wang 2012), $\boldsymbol{B}$ is an arbitrary $p\times p$ constant matrix, and $\mathbf{1}_p$ is a p-dimensional vector with all elements being one.

It is useful to identify several typical members in this class, such as the ordinary least-square estimator (OLS) and the weighted least-square estimator (WLS). Ma and Wang (2012) showed that for uncensored data, the OLS is derived by viewing the $\boldsymbol{q}_i$'s as covariates and $I(T_i\leqslant t)$ as response variables where the covariates and the responses are linked by $F(t)$ via a linear regression model

$$Y_i\equiv I(T_i\leqslant t)=\boldsymbol{q}_i^T\boldsymbol{F}(t)+e_i,$$

with $E(e_i|\boldsymbol{q}_i)=0$, $i=1,\dots,n$. It is straightforward that the e_i's are independent conditional on $\boldsymbol{q}_i$'s, and have variances $v_i=\boldsymbol{q}_i^T\boldsymbol{F}(t)\{1-\boldsymbol{q}_i^T\boldsymbol{F}(t)\}$. The WLS is then defined by using the inverse of the variances v_i as weights in a weighted least square.

Both the OLS and the WLS correspond to special members of the family of all influence functions. Specifically, the OLS has the influence function

$$\boldsymbol{\phi}_{\mathrm{OLS}}(\boldsymbol{q},s)=\{E(\boldsymbol{Q}\boldsymbol{Q}^T)\}^{-1}\boldsymbol{q}\left\{I(s<t)-\boldsymbol{q}^T\boldsymbol{F}(t)\right\},$$

and WLS has the influence function

$$\boldsymbol{\phi}_{\mathrm{WLS}}(\boldsymbol{q},s)=\{E(W\boldsymbol{Q}\boldsymbol{Q}^T)\}^{-1}w\boldsymbol{q}\left\{I(s<t)\boldsymbol{q}^T\boldsymbol{F}(t)\right\}.$$

where W is a random weight variable. Furthermore, by projecting an influence function onto the tangent space, Ma and Wang (2012) derived the efficient influence function corresponding to a semiparametric efficient estimator:

$$\boldsymbol{\phi}_{\mathrm{EFF}}(\boldsymbol{q},s)=\frac{\{I(s<t)\boldsymbol{I}_p-K\}\boldsymbol{A}^{-1}(s)\boldsymbol{q}}{\boldsymbol{q}^T\boldsymbol{f}(s)},$$

where $\boldsymbol{A}(s) = \int \frac{\boldsymbol{Q}\boldsymbol{Q}^T p_{\boldsymbol{Q}}(\boldsymbol{Q})}{\boldsymbol{Q}^T \boldsymbol{f}(s)} d\mu(\boldsymbol{Q})$, and $\boldsymbol{K} = \int_{T_1}^{T_2} I(s < t)\boldsymbol{A}^{-1}(s)ds \left\{ \int_{T_1}^{T_2} \boldsymbol{A}^{-1}(s)ds \right\}^{-1}$.

The form $\boldsymbol{\phi}_{\text{EFF}}$ is known, but may contain unknown nuisance parameters such as the density $\boldsymbol{f}(\cdot)$. As before, we assume $\boldsymbol{f}(\cdot)$ is completely unspecified (nonparametric), and thus is an infinite dimensional nuisance parameter. Ma and Wang (2012) showed that substituting consistent estimators for the nuisance parameters in $\boldsymbol{\phi}_{\text{EFF}}$ and solving for $\boldsymbol{F}(t)$ leads to a semiparametric efficient estimator which reaches the semiparametric efficiency bound in the sense of Bickel et al. (1993).

Supplementary Material

There is an online Supplementary Material available at the Journal website.

References

Arias, E. (2006). United States Life Tables, 2003. *National Vital Statistics Report*, **54**, 1-40.

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

Bang, H. and Tsiatis, A. A. (2002). Median Regression with Censored Cost Data. *Biometrics*, **58**, 643-649.

Begg, C. B. (2002). On the Use of Familial Aggregation in Population-Based Case Probands for Calculating Penetrance. *Journal of the National Cancer Institute*, **94**, 1221-1226.

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

Chatterjee, N. and Wacholder, S. (2001). A Marginal Likelihood Approach for Estimating Penetrance from Kin-cohort Designs. *Biometrics*, **57**, 245-252.

Diao, G. and Lin, D. (2005). Semiparametric Methods for Mapping Quantitative Trait Loci with Censored Data. *Biometrics*, **61**, 789-798.

Dorsey E. Ray, Beck Christopher, Adams Mary, et al and the Huntington Study Group (2008). TREND-HD Communicating Clinical Trial Results to Research Participants.

Arch Neurol, **65**, 1590-1595.

Fine, J., Zhou, F., and Yandell, B. (2004). Nonparametric Estimation of the Effects of Quantitative Trait Loci. *Biostatistics*, **5**, 501-513.

Foroud, T., Gray, J., Ivashina, J., and Conneally, P. (1999). Differences in Duration of Huntingtons Disease based on Age at Onset. *J. Neurol. Neurosurg Psychiatry*, **66**, 52-56.

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

Horvitz, D. G. and Thompson, D.J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, **47**, 663-685.

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

Huntington Study Group. (2011). HSG: Huntington Study Group, Seeking Treatments that Make a Difference for Huntington Disease. Available at http://www.huntington-study-group.org.

Khoury, M., Beaty, H. and Cohen, B. (1993). *Fundamentals of Genetic Epidemiology.* New York: Oxford University Press.

Lander, E. S. and Botstein, D. (1989). Mapping Mendelian Factors Underlying Quantiative Traits Using RFLP Linkage Maps. *Genetics*, **121**, 743-756.

Langbehn, D., Brinkman, R., Falush, D., Paulsen, J., and Hayden, M. (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.

Li, H., Yang, P., and Schwartz, A. G. (1998). Analysis of Age of Onset Data from Case-Control Family Studies. *Biometrics*, **54**, 1030-1039.

Lipsitz, S.R., Ibrahim, J.G., and Zhao, L.P. (1999). A Weighted Estimating Equation for Missing Covariate Data with Properties Similar to Maximum Likelihood. *Journal of the American Statistical Association*, **94**, 1147-1160.

Ma, Y. and Wang, Y. (2012). Efficient Semiparametric Estimation for Mixture Data. *E*lectronic Journal of Statistics, **6**, 710-737.

Mai, P. L., Chatterjee, N., Hartge, P., Tucker, M., Brody, L., Struewing, J. P., and Wacholder, S. (2009). Potential Excess Mortality in BRCA1/2 Mutation Carriers beyond Breast, Ovarian, Prostate, and Pancreatic Cancers, and Melanoma. *PLoS ONE*, **4**, e4812, DOI: 10.1371/journal.pone.0004812.

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 Group, T. H. S. (2000). Rate of Functional Decline in Huntingtons Disease. *Neurology*, **369**, 452-458.

Rabinowitz, D. (2000). Computing the Efficient Score in Semi-parametric Problems. *Statistica Sinica*, **10**, 265-280.

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.

Robins, J. M., Rotnitzky, A. and Zhou, L. P. (1994). Estimation of Regression Coefficients when some Regressors are Not Always Observed. *Journal of the American Statistical Association*, **89**, 846-866.

Robins, J. M. and Wang, N. (2000). Inference for Imputation Estimators, *Biometrika*, **87**, 113-124.

Rubinsztein, D. C., Leggo, J., Coles, R., Almqvist, E., et al. (1996). Phenotypic Characterization of Individuals with 30-40 CAG Repeats in the Huntington Disease (HD) Gene Reveals HD Cases with 36 repeats and Apparently Normal Elderly Individuals with 36-39 Repeats. *American Journal of Human Genetics*, **59**, 16-22.

Struewing, J.P., Hartge, P., Wacholder, S., Baker, S.M., Berlin, M., McAdams, M., Timmerman, M.M., Brody, L.C., Tucker, M.A. (1997). The Risk of Cancer Associated with Specific Mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *New Engl. J. Med.*

336, 1401-1408.

Tsiatis, A. A., and Ma, Y. (2004), Locally Efficient Semiparametric Estimators for Functional Measurement Error Model. *Biometrika*, **91**, 835-848.

Van der Laan, M. J. and Hubbard, A. E. (1998). Locally Efficient Estimation of the Survival Distribution with Right-Censored Data and Covariates when Collection of Data is Delayed. *Biometrika*, **85**, 771-783.

Wacholder, S., Hartge, P., Struewing, J., Pee, D., McAdams, M., Brody, L. and Tucker, M. (1998). The Kin-cohort Study for Estimating Penentrance. *American Journal of Epidemiology.* **148**, 623-630.

Wang, N. and Robins, J. M. (1998). Large Sample Inference in Parametric Multiple Imputation. *Biometrika*, **85**, 935-948.

Wang, Y., Clark, L. N., Louis, E. D., Mejia-Santana, H., Harris, J., Cote, L. J., Waters, C., Andrews, H., Ford, B., Frucht, S., Fahn, S., Ottman, R., Rabinowitz, D., and Marder, K. (2008). Risk of Parkinson Disease in Carriers of Parkin Mutations: Estimation using the Kin-Cohort Method. *Arch. Neurol.*, **65**, 467-474.

Wang, Y., Clark, L. N., Marder, K. and Rabinowitz, D. (2007). Non-parametric Estimation of Genotype-specific Age-at-onset Distributions From Censored Kin-cohort Data. *Biometrika*, **94**, 403-414.

Wu, R., Ma, C., and Casella, G. (2007). *Statistical Genetics of Quantitative Traits: Linkage, Maps, and QTL.* New York: Springer.

Wu, R., Ma, C., Chang, M., Littell, R., Wu, S., Yin T, Huang, M., Wang, M., and Casella, G. (2002). A Logistic Mixture Model for Characterizing Genetic Determinants Conserved Synteny in Rat and Mouse for a Blood Pressure Causing Differentiation in Growth Trajectories. *Genet. Res.* **79**, 235-245.

Wu, R., Zeng, Z., McKend, S., and OMalley, D. (2000). The case for Molecular Mapping in Forest Tree Breeding. *Plant Breeding Reviews*, **19**, 41-68.

Yu, Z. and Lin, X. (2008) Nonparametric Regression using Local Kernel Estimating Equations for Correlated Failure Time Data. *Biometrika*, **95**, 123-137.

Zeng, D. and Lin, D. (2007). Maximum Likelihood Estimation in Semiparametric Models with Censored Data (with discussion). *Journal of the Royal Statistical Society, Series B*, **69**, 507-564.

Zhang, M., Tsiatis, A. A. and Davidian, M. (2008). Improving Efficiency of Inferences in Randomized Clinical Trials using Auxiliary Covariates. *Biometrics*, **64**, 707-715.

Zhao, W., and Wu, R. (2008). Wavelet-based Nonparametric Functional Mapping of Longitudinal Curves. *Journal of the American Statistical Association*, **103**, 714-725.

Table 1: Simulation 1. Bias, empirical standard deviation (emp sd), average estimated standard deviation (est sd), 95% coverage (95% cov) of 11 estimators. Sample size $n = 500$, 20% and 50% censoring rate, 1000 simulations.

		$F_1(t) = 0.5063$				$F_2(t) = 0.5132$		
Estimator	bias	emp sd	est sd	95% cov	bias	emp sd	est sd	95% cov
				Group 1: OLS based, censoring rate =20%				
IPW	0.0013	0.0424	0.0427	0.9460	-0.0026	0.0408	0.0418	0.9590
AIPW	0.0003	0.0396	0.0402	0.9450	-0.0021	0.0390	0.0393	0.9520
IMP	0.0004	0.0396	0.0401	0.9460	-0.0022	0.0389	0.0392	0.9500
				Group 2: WLS based, censoring rate =20%				
IPW	0.0013	0.0424	0.0427	0.9450	-0.0026	0.0408	0.0418	0.9590
AIPW	0.0003	0.0396	0.0402	0.9440	-0.0021	0.0390	0.0393	0.9520
IMP	0.0004	0.0396	0.0401	0.9460	-0.0022	0.0389	0.0392	0.9500
				Group 3: EFF based, censoring rate =20%				
IPW	0.0011	0.0427	0.0431	0.9520	-0.0029	0.0418	0.0433	0.9630
AIPW	0.0004	0.0399	0.0404	0.9480	-0.0022	0.0393	0.0399	0.9540
IMP	0.0004	0.0398	0.0403	0.9430	-0.0022	0.0391	0.0394	0.9500
				Group 4: NPMLE, censoring rate =20%				
type I	0.0000	0.0462	0.0466	0.9470	0.0007	0.0881	0.0897	0.9230
type II	-0.0135	0.0364	0.0364	0.9300	0.0075	0.0308	0.0313	0.9390
				Group 1: OLS based, censoring rate =50%				
IPW	0.0043	0.0733	0.0683	0.9290	0.0004	0.0722	0.0668	0.9290
AIPW	0.0010	0.0452	0.0459	0.9530	-0.0026	0.0452	0.0448	0.9450
IMP	0.0043	0.0486	0.0496	0.9580	0.0012	0.0484	0.0480	0.9480
				Group 2: WLS based, censoring rate =50%				
IPW	0.0044	0.0732	0.0683	0.9290	0.0004	0.0722	0.0668	0.9280
AIPW	0.0010	0.0452	0.0459	0.9520	-0.0026	0.0451	0.0448	0.9460
IMP	0.0043	0.0486	0.0496	0.9590	0.0012	0.0484	0.0480	0.9470
				Group 3: EFF based, censoring rate =50%				
IPW	-0.0021	0.0740	0.0702	0.9290	0.0005	0.0799	0.0744	0.9300
AIPW	-0.0006	0.0465	0.0464	0.9500	-0.0008	0.0470	0.0455	0.9480
IMP	0.0031	0.0493	0.0508	0.9610	0.0026	0.0495	0.0489	0.9480
				Group 4: NPMLE, censoring rate =50%				
type I	0.0011	0.0531	0.0537	0.9410	0.0001	0.1058	0.1030	0.9190
type II	-0.0405	0.0407	0.0417	0.8350	0.0312	0.0381	0.0382	0.8750

Table 2: Simulation 1. Integrated absolute bias (IAB) Upper half of table: the true censoring distribution is independent of $\boldsymbol{Q}_i$, $G(t)$ is estiimated using a common Kaplan-Meier estimator of the censoring distribution. Lower half of table: the true censoring distribution is subgroup-specific, $G(t)$ is estimated using a common Kaplan-Meier estimator (denoted by †), or a subgroup-specific Kaplan-Meier estimator (denoted by *).

	Censoring rate			
	20%		50%	
Estimator	$F_1(t)$	$F_2(t)$	$F_1(t)$	$F_2(t)$
	Group 1: OLS based			
IPW	0.0103	0.0060	0.0332	0.0096
AIPW	0.0091	0.0055	0.0305	0.0077
IMP	0.092	0.0056	0.0348	0.0146
	Group 2: WLS based			
IPW	0.0102	0.0060	0.0345	0.0095
AIPW	0.0370	0.0055	0.1062	0.0092
IMP	0.0092	0.0056	0.0357	0.0145
	Group 3: EFF based			
IPW	0.0095	0.0062	0.0170	0.0106
AIPW	0.0090	0.0053	0.0462	0.0070
IMP	0.0090	0.0055	0.0325	0.0195
	Group 4: NPMLE			
type I	0.0104	0.0116	0.0174	0.0337
type II	0.0996	0.0374	0.2483	0.1367
	Group 1: OLS based			
IPW†	0.0922	0.0632	0.0651	0.0412
AIPW†	0.0110	0.0045	0.0317	0.0101
IPW*	0.0159	0.0052	0.0181	0.0073
AIPW*	0.0148	0.0052	0.0240	0.0079
	Group 2: WLS based			
IPW†	0.0923	0.0632	0.0658	0.0413
AIPW†	0.0517	0.0045	0.1024	0.0111
IPW*	0.0169	0.0051	0.0218	0.0074
AIPW*	0.0591	0.0051	0.1031	0.0092
	Group 3: EFF based			
IPW†	0.1104	0.0791	0.0642	0.0495
AIPW†	0.0122	0.0054	0.0392	0.0094
IPW*	0.0151	0.0053	0.0149	0.0071
AIPW*	0.0112	0.0073	0.0351	0.0068

Table 3: Simulation 1. Bias, empirical standard deviation (emp sd), average estimated standard deviation (est sd), 95% coverage (95% cov) of 11 estimators. The censoring distribution is subgroup-specific, and $G(t)$ is estimated using a common Kaplan-Meier estimator (denoted by †), or a subgroup-specific Kaplan-Meier estimator (denoted by *). Sample size $n = 500$, 20% and 50% censoring rate, 1000 simulations.

		$F_1(t) = 0.5063$				$F_2(t) = 0.5132$		
Estimator	bias	emp sd	est sd	95% cov	bias	emp sd	est sd	95% cov
			Group 1: OLS based, censoring rate =20%					
IPW†	-0.0132	0.0439	0.0443	0.9440	0.0131	0.0433	0.0428	0.9360
AIPW†	0.0007	0.0386	0.0399	0.9520	-0.0018	0.0380	0.0381	0.9490
IPW*	0.0019	0.0395	0.0395	0.9370	-0.0012	0.0384	0.0389	0.9530
AIPW*	0.0018	0.0391	0.0391	0.9430	-0.0012	0.0383	0.0384	0.9490
			Group 2: WLS based, censoring rate =20%					
IPW†	-0.0132	0.0439	0.0443	0.9440	0.0131	0.0433	0.0428	0.9360
AIPW†	0.0007	0.0386	0.0399	0.9520	-0.0018	0.0380	0.0381	0.9490
IPW*	0.0019	0.0395	0.0395	0.9400	-0.0012	0.0384	0.0389	0.9530
AIPW*	0.0018	0.0390	0.0391	0.9430	-0.0012	0.0383	0.0384	0.9500
			Group 3: EFF based, censoring rate =20%					
IPW†	-0.0153	0.0459	0.0452	0.9350	0.0163	0.0446	0.0450	0.9360
AIPW†	0.0004	0.0392	0.0401	0.9470	-0.0014	0.0387	0.0386	0.9470
IPW*	0.0020	0.0392	0.0396	0.9390	-0.0014	0.0384	0.0393	0.9550
AIPW*	0.0013	0.0393	0.0393	0.9390	-0.0007	0.0387	0.0390	0.9500
			Group 1: OLS based, censoring rate =50%					
IPW†	-0.0077	0.0708	0.0682	0.9320	0.0130	0.0729	0.0663	0.9230
AIPW†	0.0014	0.0448	0.0472	0.9570	-0.0027	0.0463	0.0442	0.9380
IPW*	0.0032	0.0502	0.0474	0.9410	-0.0028	0.0492	0.0477	0.9380
AIPW*	0.0021	0.0473	0.0449	0.9410	-0.0021	0.0471	0.0455	0.9330
			Group 2: WLS based, censoring rate =50%					
IPW†	-0.0076	0.0707	0.0682	0.9310	0.0130	0.0729	0.0663	0.9220
AIPW†	0.0014	0.0448	0.0472	0.9550	-0.0027	0.0462	0.0442	0.9380
IPW*	0.0033	0.0504	0.0474	0.9390	-0.0028	0.0492	0.0477	0.9370
AIPW*	0.0022	0.0473	0.0449	0.9420	-0.0021	0.0471	0.0455	0.9330
			Group 3: EFF based, censoring rate =50%					
IPW†	-0.0148	0.0753	0.0707	0.9250	0.0169	0.0794	0.0736	0.9240
AIPW†	0.0008	0.0463	0.0477	0.9570	-0.0021	0.0487	0.0449	0.9350
IPW*	0.0020	0.0481	0.0466	0.9480	-0.0024	0.0485	0.0477	0.9400
AIPW*	0.0003	0.0472	0.0454	0.9470	-0.0003	0.0483	0.0462	0.9390s

Table 4: Estimated survival rates and 95% confidence intervals (in parentheses) based on EFFAIPW and type I NPMLE for carrier and non-carrier groups in the COHORT data stratified by gender. Survival rates are compared to Kaplan-Meier estimated survival rates for the general male and female US populations (USpop) in 2003.

		Non-Carrier		Carrier	
Age	USpop	EFFAIPW	type I NPMLE	EFFAIPW	type I NPMLE
		Males			
30	97.2 (97.1, 97.3)	97.5 (96.6, 98.3)	98.1 (96.8, 99.4)	98.7 (98.2, 99.3)	99.7 (99.4, 99.9)
35	96.5 (96.4, 96.6)	97.4 (96.6, 98.3)	97.9 (96.8, 99.0)	97.2 (96.2, 98.2)	99.4 (98.9, 99.9)
40	95.6 (95.5, 95.7)	96.8 (95.7, 98.0)	96.5 (94.2, 98.7)	94.1 (92.4, 95.8)	98.4 (97.6, 99.2)
45	94.3 (94.1, 94.4)	95.5 (94.0, 97.1)	95.1 (92.1, 98.1)	91.2 (89.0, 93.4)	97.4 (96.1, 98.7)
50	92.3 (92.1, 92.4)	93.9 (92.0, 95.8)	93.8 (90.0, 97.5)	86.4 (83.6, 89.3)	93.4 (88.8, 98.1)
55	89.4 (89.2, 89.6)	89.7 (86.9, 92.5)	88.9 (83.5, 94.3)	77.9 (74.0, 81.9)	86.5 (78.8, 94.3)
60	85.5 (85.3, 85.7)	83.2 (79.5, 87.0)	82.4 (76.1, 88.7)	66.5 (61.4, 71.5)	82.4 (72.4, 92.5)
65	79.9 (79.6, 80.1)	77.4 (72.9, 81.9)	72.6 (65.8, 79.3)	53.1 (47.0, 59.1)	76.9 (65.8, 88.0)
70	72.0 (71.7, 72.3)	68.1 (62.7, 73.5)	65.4 (56.3, 74.5)	40.3 (33.6, 47.0)	64.4 (43.5, 85.3)
75	61.3 (61.0, 61.6)	59.4 (53.3, 65.6)	54.8 (45.5, 64.2)	27.3 (20.5, 34.0)	58.9 (39.3, 78.5)
		Females			
30	98.4 (98.4, 98.5)	98.0 (97.1, 98.9)	98.7 (98.1, 99.3)	98.9 (98.1, 99.7)	99.5 (99.1, 99.9)
35	98.1 (98.0, 98.2)	97.8 (96.8, 98.7)	97.5 (95.6, 99.4)	97.7 (96.6, 98.9)	99.0 (98.4, 99.5)
40	97.6 (97.5, 97.7)	97.7 (96.7, 98.8)	96.5 (93.6, 99.4)	96.3 (94.8, 97.8)	98.2 (97.5, 98.9)
45	96.7 (96.6, 96.9)	97.4 (96.2, 98.6)	94.3 (89.8, 98.7)	92.6 (90.5, 94.7)	96.2 (95.3, 97.1)
50	95.5 (95.4, 95.7)	96.1 (94.4, 97.8)	90.3 (85.0, 95.7)	88.9 (86.2, 91.7)	93.8 (92.4, 95.3)
55	93.8 (93.6, 94.0)	94.7 (92.7, 96.7)	89.5 (84.1, 94.9)	79.9 (76.3, 83.5)	89.9 (88.2, 91.6)
60	91.2 (91.1, 91.4)	93.0 (90.3, 95.6)	86.6 (80.5, 92.7)	69.1 (64.5, 73.8)	83.7 (81.6, 85.8)
65	87.3 (87.1, 87.5)	91.0 (87.6, 94.4)	84.9 (79.2, 90.7)	57.0 (51.3, 62.6)	75.9 (73.3, 78.5)
70	81.6 (81.3, 81.8)	84.0 (79.3, 88.6)	74.5 (67.3, 81.6)	42.8 (36.5, 49.2)	68.7 (66.2, 71.1)
75	73.3 (73.0, 73.6)	73.9 (68.0, 79.8)	62.8 (55.0, 70.6)	33.2 (26.6, 39.9)	63.6 (60.7, 66.5)

Table 5: Summary statistics for the COHORT data and the general male and female US populations (USpop) in 2003 from ages 20-90 years.

	Males		Females	
	Non-Carrier	Carrier	Non-Carrier	Carrier
	Expected years of life (area under the survival curves)			
USpop	55.5559		59.8726	
EFFAIPW	54.8759	45.8229	59.6289	46.8695
type I NPMLE	53.6082	57.7125	57.0451	52.1563
type I NPMLE*	56.5764	47.3792	61.1127	47.2930
	Expected years of life lost compared to the US population			
EFFAIPW	0.6800	9.7330	0.2436	13.0031
type I NPMLE	1.9477	-2.1566	2.8275	7.7162
type I NPMLE*	-1.0205	8.1767	-1.2401	12.5796
	IAB between an estimator and the US population			
EFFAIPW	1.4026	10.0315	1.2957	13.1922
type I NPMLE	2.6079	3.9104	2.9016	8.1144
type I NPMLE *	1.1071	8.7317	1.8435	12.7560

*: Computed under a sub-sample by removing subjects in $\boldsymbol{Q}_i$ groups with small sample sizes.

Table 6: Estimated conditional probabilities of survival for the COHORT data based on the EFFAIPW estimator.

	Conditional probability individual alive in the next # of years							
Current	5 years	10 years	15 years	20 years	5 years	10 years	15 years	20 years
age	Male Carriers				Male Non-Carriers			
30	0.9843	0.9533	0.9239	0.8753	0.9996	0.9937	0.9803	0.9636
35	0.9686	0.9387	0.8893	0.8017	0.9941	0.9808	0.9640	0.9206
40	0.9692	0.9182	0.8277	0.7061	0.9866	0.9698	0.9261	0.8594
45	0.9474	0.8540	0.7285	0.5818	0.9830	0.9387	0.8711	0.8099
50	0.9015	0.7690	0.6142	0.4666	0.9550	0.8862	0.8239	0.7251
55	0.8530	0.6813	0.5176	0.3498	0.9280	0.8628	0.7593	0.6628
60	0.7987	0.6068	0.4101	0.2122	0.9297	0.8182	0.7142	0.4984
	Female Carriers				Female Non-Carriers			
30	0.9887	0.9744	0.9363	0.8997	0.9977	0.9975	0.9940	0.9806
35	0.9855	0.9471	0.9100	0.8176	0.9998	0.9963	0.9829	0.9689
40	0.9610	0.9233	0.8296	0.7177	0.9965	0.9831	0.9691	0.9511
45	0.9608	0.8633	0.7469	0.6153	0.9865	0.9725	0.9544	0.9346
50	0.8985	0.7773	0.6404	0.4814	0.9858	0.9674	0.9473	0.8737
55	0.8651	0.7127	0.5358	0.4156	0.9814	0.9610	0.8863	0.7802
60	0.8239	0.6194	0.4804	0.1971	0.9792	0.9031	0.7950	0.6116

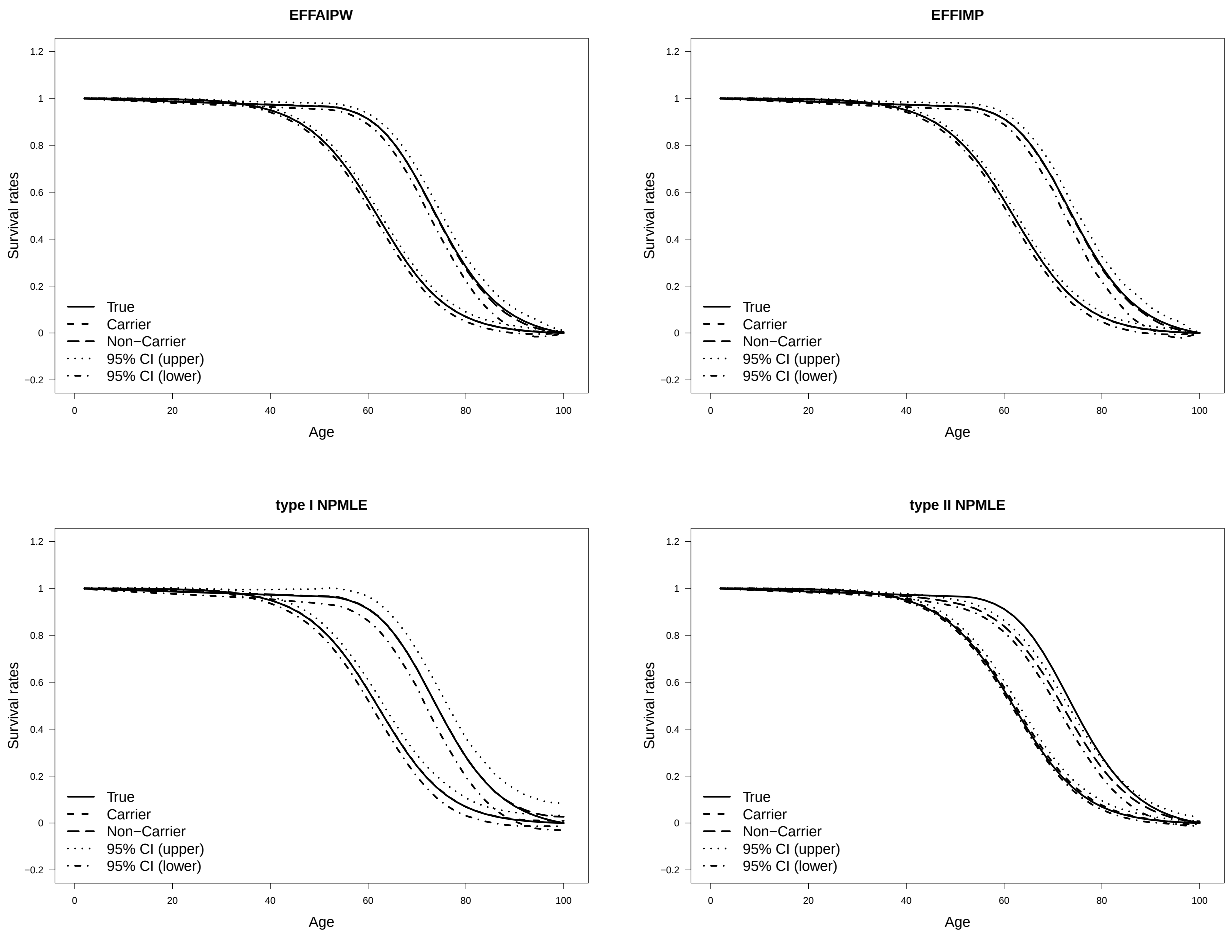

Figure 1: Simulation 3. True survival curve (solid) and the mean of 250 simulations at each time point (short-dashed for carrier group, long-dashed for non-carrier group), 95% pointwise confidence band (upper band dotted, lower band dash-dotted) of the estimated survival curves. The mean and true survival curves are indistinguishable in EFFIMP and EFFAIPW estimators. Sample size is 4500, censoring rate is 65%.

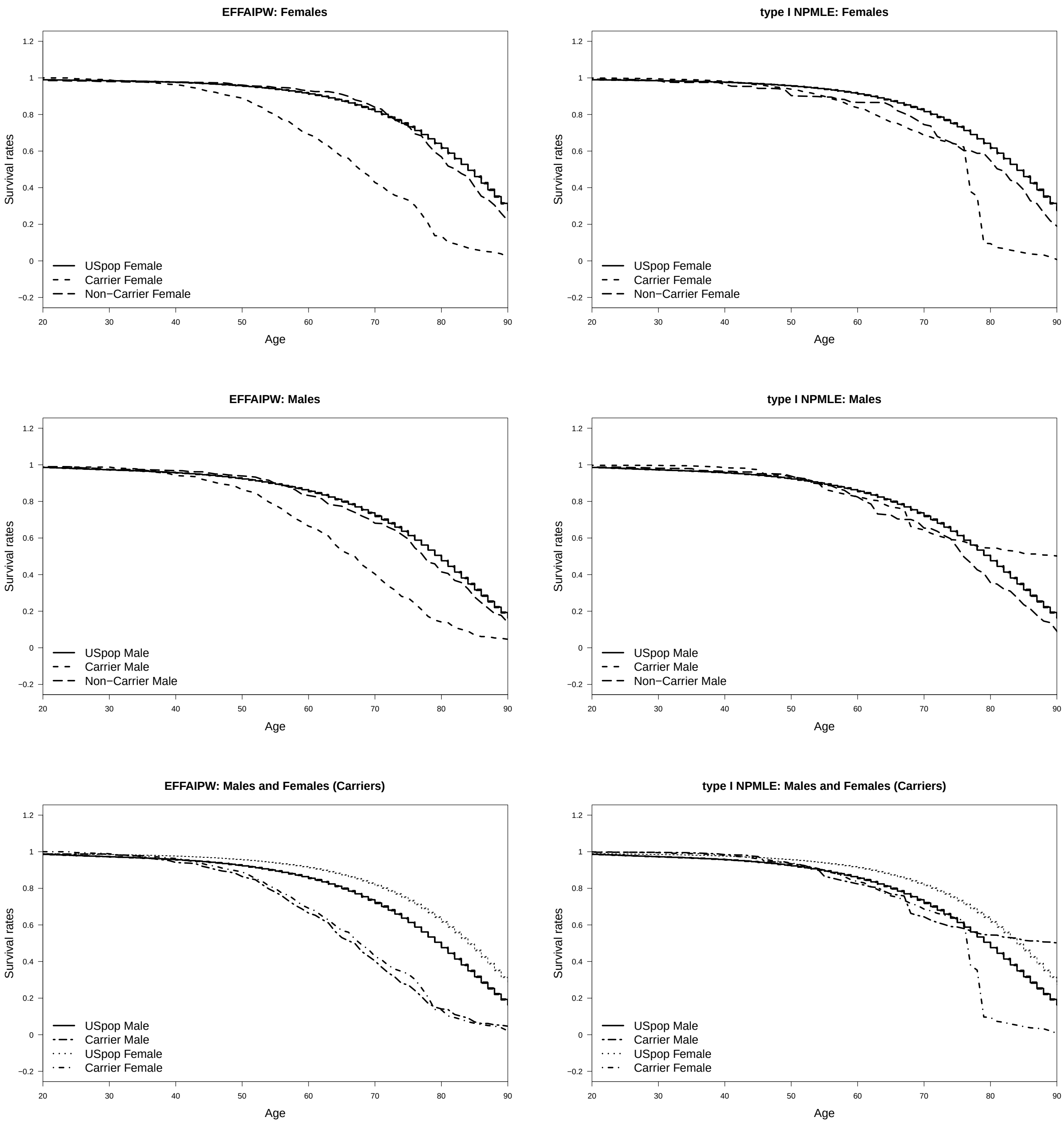

Figure 2: Estimated survival curves for the COHORT data stratified by gender using the EFFAIPW estimator (left) and the type I NPMLE (right). The curves "USpop Male" and "USpop Female" correspond to Kaplan-Meier estimated survival rates for the general male and female US populations in 2003. Bottom two figures compare the estimated survival curves for the male and female HD gene mutation carriers to that of the general male and female US populations.

Supplementary Material for "Nonparametric estimation for censored mixture data with application to the Cooperative Huntington's Observational Research Trial"

Yuanjia Wang, Tanya P. Garcia and Yanyuan Ma

S.1 Inconsistency of the type II NPMLE

From Section 2.2, it is easy to see that the type II optimization should only assign positive weights to the observed event times, and the estimating procedure proceeds to form

$$\check{\boldsymbol{F}}(t) = \sum_{i=1}^{n} \delta_i I(x_i < t) \check{\boldsymbol{f}}(x_i).$$

If a set of $\boldsymbol{f}(x_i)$'s is the maximizer of (1) under the type II NPMLE constraints, then $\boldsymbol{q}_i^T \boldsymbol{f}(x_i)$'s are all non-negative and satisfy $0 \leqslant \boldsymbol{q}_i^T \boldsymbol{F}(x_i) \leqslant 1$. Thus, if we denote $H(x_i; \boldsymbol{q}_1, \ldots, \boldsymbol{q}_n) = \boldsymbol{q}_i^T \boldsymbol{F}(x_i)$ and $h(x_i; \boldsymbol{q}_1, \ldots, \boldsymbol{q}_n) = \boldsymbol{q}_i^T \boldsymbol{f}(x_i)$ for all $i = 1, \ldots, n$, then the maximization is obtained at the MLE estimator of the hypothetical H. Note that H depends on $\boldsymbol{q}_i$'s which take m different values, and we view these m values as known parameters. For notational simplicity, we write $H(x; \boldsymbol{q}_1, \ldots, \boldsymbol{q}_n)$ as $H(x)$. Thus, if one thinks of $\boldsymbol{u}_j^T \boldsymbol{F}(x)$ as a particular mixture of $\boldsymbol{F}(x)$, then H contains all m mixtures. In addition, H is not necessarily a valid distribution function since it may not be monotone. With the maximum of H as $\check{H}$, one can then recover the $\boldsymbol{F}(x_i)$'s through $\check{\boldsymbol{F}}(x_i) = (\boldsymbol{q}_i^T \boldsymbol{r}_i)^{-1} \boldsymbol{r}_i \check{H}_i$ for some length p vector $\boldsymbol{r}_i$ (i.e., $\boldsymbol{r}_i = \boldsymbol{F}(x_i)$). If a suitable selection of r_i's exists to ensure $0 \leqslant (\boldsymbol{q}_i^T \boldsymbol{r}_i)^{-1} \boldsymbol{r}_i \check{H}_i \leqslant 1$ and monotonicity, then we have a solution of the type II optimization problem.

Assuming to the contrary this solution is consistent, then the resulting $\check{\boldsymbol{F}}(t)$ would satisfy

$$\boldsymbol{u}_j^T \check{\boldsymbol{F}}(t) = 1 - \check{S}_j(t)$$

for all $j = 1, \ldots, m$, where $\check{S}_j(t)$ is a consistent estimator of $S_j(t)$. Here $S_j(t)$ has the same definition as in the type I NPMLE and is the survival function of the observations that have the common mixing proportion, $\boldsymbol{u}_j$. Obviously, $\check{\boldsymbol{F}}(t) = \check{\boldsymbol{F}}(x_i)$, where x_i is the largest x value that satisfies $x_i \leqslant t$ and $\delta_i = 1$. Without confusion, we assume the corresponding $\boldsymbol{q}_i = \boldsymbol{u}_j$, thus we have

$$\boldsymbol{u}_j^T \check{\boldsymbol{F}}(t) = \check{H}_i = 1 - \check{S}_j(t).$$

Because $\check{S}_j$ is a consistent estimator of S_j, we must have $\check{H}_i$ converging to $1 - S_j(t)$. However, this leads to a contradiction since $1 - S_j(t) = \boldsymbol{u}_j^T \boldsymbol{F}(t)$ is the distribution function corresponding to the jth mixing vector $\boldsymbol{u}_j$, while $\check{H}(t)$ aims at estimating the hypothetic H function which contains all m different mixtures.

It is not always possible or easy to find the $\check{H}_i$'s or to identify the $\boldsymbol{r}_i$'s. For this reason, the type II NPMLE is hardly ever solved through obtaining $\check{H}_i$'s and $\boldsymbol{r}_i$'s. Instead, the EM algorithm introduced in Section 2.2 is used to obtain the $\boldsymbol{f}(x_i)$'s. However, the above explanation reveals the underlying reason why the type II NPMLE fails. Conceptually, identifying the $F_1, \ldots, F_p$ functions is equivalent to identifying the $S_1, \ldots, S_m$ functions. The type II NPMLE maximizes the product of m different likelihoods, $S_1, \ldots, S_m$, formed by all observations with respect to a collection of parameters, while in fact each of these parameters should concern only one of these likelihoods formed by a subset of the observations. The type II NPMLE cannot identify correct parameters and observations for each kind of likelihood.

For example, consider an uncensored observation where $\boldsymbol{Q}_i = (0.5, 0.5)$ and $T_i = t_i$. The event time t_i will carry an equal weight with the type II estimator for each of the two subpopulations. However, if this observation belongs to the first subpopulation, then the nonparametric estimator of the CDF of the first population, $\widehat{F}_1$, should have a jump at t_i but not the $\widehat{F}_2$. In other words, t_i should not be included in the support when estimating F_2 nonparametrically. The type II NPMLE will let both $\widehat{F}_1$ and $\widehat{F}_2$ have a jump at t_i. Likewise, if this observation belongs to the second subpopulation, then t_i should not be included in the support when estimating F_1 nonparametrically. Since we do not observe which subpopulation an observation comes from, it is difficult for the type II NPMLE to correctly identify the support for each subpopulation. In contrast, the type I NPMLE does not suffer from this difficulty: when regarding $S_j = \boldsymbol{u}_j^T \boldsymbol{F}$ as an unknown parameter, the type I correctly assigns the support of S_j to include all observations with $\boldsymbol{q}_i = \boldsymbol{u}_j$. Such a computation is feasible as all $\boldsymbol{q}_i$ are observed.

To help further understand the mistreatment of different conditional likelihoods of the type II NPMLE, one may consider maximizing a marginal likelihood. Denoting $n_j = \sum_{i=1}^n I(\boldsymbol{q}_i = \boldsymbol{u}_j)$, $j = 1, \ldots, m$, this would correspond to maximizing

$$\sum_{i=1}^n \log \sum_{j=1}^m n_j \left\{\boldsymbol{u}_j^T \boldsymbol{f}(x_i)\right\}^{\delta_i} \{1 - \boldsymbol{u}_j^T \boldsymbol{F}(x_i)\}^{1-\delta_i}$$

with respect to $\boldsymbol{f}(x_i)$'s. This estimator uses all observed x_i's to form a common marginal likelihood, hence is consistent. However it completely ignores the pair information provided in the data hence could be highly inefficient.

S.2 Influence function of the IPW estimator

To derive the influence function of the IPW estimator, we first note the following several useful facts (Bang and Tsiatis 2000; Robins and Rotnitzky 1992)

$$\begin{aligned} Y(t) &= n\widehat{G}(t^-)\widehat{S}(t^-) \\ \frac{\widehat{G}(t) - G(t)}{G(t)} &= -\int_0^t \frac{\widehat{G}(u^-)}{G(u)} \frac{dM^c(u)}{Y(u)} \\ \frac{\delta_i}{G(x_i)} &= 1 - \int \frac{dM_i^c(u)}{G(u)}. \end{aligned}$$

Using the above relations, we expand the IPW estimator as

$$\begin{aligned} n^{-1/2} \sum_{i=1}^n \frac{\delta_i \boldsymbol{\phi}(\boldsymbol{q}_i, x_i)}{\widehat{G}(x_i)} &= n^{-1/2} \sum_{i=1}^n \frac{\delta_i \boldsymbol{\phi}(\boldsymbol{q}_i, t_i)}{G(x_i)} + n^{-1/2} \sum_{i=1}^n \frac{\delta_i \boldsymbol{\phi}(\boldsymbol{q}_i, t_i)}{\widehat{G}(x_i)} \int_0^{x_i} \frac{\widehat{G}(u^-)}{G(u)} \frac{dM^c(u)}{Y(u)} \\ &= n^{-1/2} \sum_{i=1}^n \frac{\delta_i \boldsymbol{\phi}(\boldsymbol{q}_i, t_i)}{G(x_i)} + n^{-1/2} \sum_{i=1}^n \int \frac{\widehat{\mathcal{B}}_1\{\boldsymbol{\phi}, u\} \widehat{S}(u) dM_i^c(u)}{G(u)\widehat{S}(u^-)} \\ &= n^{-1/2} \sum_{i=1}^n \boldsymbol{\phi}(\boldsymbol{q}_i, t_i) - n^{-1/2} \sum_{i=1}^n \int \frac{\{\boldsymbol{\phi}(\boldsymbol{q}_i, t_i) - \mathcal{B}(\boldsymbol{\phi}, u)\} dM_i^c(u)}{G(u)} + o_p(1). \end{aligned}$$

Because the complete data influence function $\boldsymbol{\phi}(\boldsymbol{q}_i, t_i)$ has the general form of $\boldsymbol{d}(\boldsymbol{q}_i, t_i) - \boldsymbol{F}(t)$ (Ma and Wang 2012), we have that $\partial \boldsymbol{\phi} / \partial \boldsymbol{F}^T(t) = -\boldsymbol{I}_p$. This, in combination with exchanging integration and differentiation of the above expansion, implies that the ith influence function for the IPW is $\boldsymbol{\phi}_{\text{ipw}}$ as stated. The two terms in $\boldsymbol{\phi}_{\text{ipw}}$ are uncorrelated because $\boldsymbol{\phi}(\boldsymbol{q}_i, t_i)$ are $\mathcal{F}(0)$ measurable. Therefore we can compute the variance of the IPW estimator as

$$\begin{aligned}
\boldsymbol{V}_{\text{ipw}} &= \text{cov}\{\boldsymbol{\phi}(\boldsymbol{Q}_i, T_i)\} + E\left[\int \frac{\{\boldsymbol{\phi}(\boldsymbol{Q}_i, T_i) - \mathcal{B}(\boldsymbol{\phi}, u)\}^{\otimes 2}}{G^2(u)} \lambda^c(u) Y_i(u) du\right] \\
&= \text{cov}\{\boldsymbol{\phi}(\boldsymbol{Q}_i, T_i)\} + E\left\{\int \frac{\mathcal{B}(\boldsymbol{\phi}, u)^{\otimes 2}}{G^2(u)} \lambda^c(u) Y_i(u) du\right\} \\
&\quad + \left[E\int \frac{E\left[\left\{\boldsymbol{\phi}(\boldsymbol{Q}_i, T_i)^{\otimes 2} - 2\boldsymbol{\phi}(\boldsymbol{Q}_i, T_i)\mathcal{B}(\boldsymbol{\phi}, u)^T\right\} I(T_i \geqslant u) | C_i\right] I(C_i \geqslant u)}{G^2(u)} \lambda^c(u) du\right] \\
&= \text{cov}\{\boldsymbol{\phi}(\boldsymbol{Q}_i, T_i)\} + E\left\{\int \frac{\mathcal{B}(\boldsymbol{\phi}^{\otimes 2}, u) - \mathcal{B}(\boldsymbol{\phi}, u)^{\otimes 2}}{G^2(u)} \lambda^c(u) Y_i(u) du\right\}.
\end{aligned}$$

S.3 Influence function of the AIPW estimator

From (6) we have

$$\begin{aligned}
0 &= n^{-1/2} \sum_{i=1}^n \frac{\delta_i \boldsymbol{\phi}(\boldsymbol{q}_i, x_i)}{\widehat{G}(x_i)} + n^{-1/2} \sum_{i=1}^n \int \frac{dN_i^c(u)}{\widehat{G}(u)} \left\{\widehat{\boldsymbol{h}}_{\text{eff},i}(u) - \widehat{\mathcal{B}}\left(\widehat{\boldsymbol{h}}_{\text{eff}}, u\right)\right\} \\
&= n^{-1/2} \sum_{i=1}^n \frac{\delta_i \boldsymbol{\phi}(\boldsymbol{q}_i, x_i)}{\widehat{G}(x_i)} + n^{-1/2} \sum_{i=1}^n \int \frac{dM_i^c(u)}{\widehat{G}(u)} \left\{\widehat{\boldsymbol{h}}_{\text{eff},i}(u) - \widehat{\mathcal{B}}\left(\widehat{\boldsymbol{h}}_{\text{eff}}, u\right)\right\} \\
&= n^{-1/2} \sum_{i=1}^n \boldsymbol{\phi}(\boldsymbol{q}_i, t_i) - n^{-1/2} \sum_{i=1}^n \int \frac{\left\{\boldsymbol{\phi}(\boldsymbol{q}_i, t_i) - \boldsymbol{h}_{\text{eff},i}(u)\right\} dM_i^c(u)}{G(u)} + o_p(1),
\end{aligned}$$

where the last equality follows from $\mathcal{B}(\boldsymbol{\phi} - \boldsymbol{h}_{\text{eff}}, u) = 0$. Similar to the IPW case, the two terms in the influence function are uncorrelated which suggests that we can compute the variance of the efficient estimator as

$$\begin{aligned}
\boldsymbol{V}_{\text{eff}} &= \text{cov}\{\boldsymbol{\phi}(\boldsymbol{Q}_i, T_i)\} + E\left[\int \frac{\left\{\boldsymbol{\phi}(\boldsymbol{Q}_i, T_i) - \boldsymbol{h}_{\text{eff},i}(u)\right\}^{\otimes 2}}{G^2(u)} \lambda^c(u) Y_i(u) du\right] \\
&= \text{cov}\{\boldsymbol{\phi}(\boldsymbol{Q}_i, T_i)\} + E\int \frac{\mathcal{B}\{(\boldsymbol{\phi} - \boldsymbol{h}_{\text{eff}})^{\otimes 2}, u\}}{G^2(u)} \lambda^c(u) Y_i(u) du.
\end{aligned}$$

S.4 Influence function of the imputation estimator

We now analyze the asymptotic properties of the imputation estimator.

$$
\begin{aligned}
0 &= n^{-1/2}\sum_{i=1}^{n}\left\{\delta_i\boldsymbol{\phi}(\boldsymbol{q}_i,x_i)+(1-\delta_i)\widehat{\boldsymbol{h}}_{\mathrm{eff},i}(x_i)\right\}\\
&= n^{-1/2}\sum_{i=1}^{n}\left\{\delta_i\boldsymbol{\phi}(\boldsymbol{q}_i,x_i)+(1-\delta_i)\boldsymbol{h}_{\mathrm{eff},i}(x_i)\right\}+n^{-1/2}\sum_{i=1}^{n}(1-\delta_i)\left\{\widehat{\boldsymbol{h}}_{\mathrm{eff},i}(x_i)-\boldsymbol{h}_{\mathrm{eff},i}(x_i)\right\}.
\end{aligned}
$$

We now inspect the last term. In our approximation in (5), $\widehat{\boldsymbol{h}}_{\mathrm{eff},i}(u)$ is estimated using weighted sample averages, with the subset of data that have a common $\boldsymbol{q}_i$ value. We now analyze $\widehat{\boldsymbol{h}}_{\mathrm{eff},i}(u)$ in the kth subsample. For notational simplicity, we assume the first n_k observations have the common q value. We have

$$
\begin{aligned}
&\widehat{\boldsymbol{h}}_{\mathrm{eff},i}(x_i)-\boldsymbol{h}_{\mathrm{eff},i}(x_i)\\
=\;&\frac{n_k^{-1}\sum_{j=1}^{n_k}\boldsymbol{\phi}(\boldsymbol{q}_j,x_j)I(x_j\geqslant x_i)\delta_j/\widehat{G}(x_j)}{n_k^{-1}\sum_{j=1}^{n_k}I(x_j\geqslant x_i)\delta_j/\widehat{G}(x_j)}-\frac{E\{\boldsymbol{\phi}(\boldsymbol{q}_i,T)I(T>x_i)\|\boldsymbol{q}_i,x_i\}}{E\{I(T>x_i)\|\boldsymbol{q}_i,x_i\}}\\
=\;&\frac{n_k^{-1}\sum_{j=1}^{n_k}\boldsymbol{\phi}(\boldsymbol{q}_j,x_j)I(x_j\geqslant x_i)\delta_j/\widehat{G}(x_j)-E\{\boldsymbol{\phi}(\boldsymbol{q}_i,T)I(T>x_i)\|\boldsymbol{q}_i,x_i\}}{E\{I(T>x_i)\|\boldsymbol{q}_i,x_i\}}\\
&-\boldsymbol{h}_{\mathrm{eff},i}(x_i)\frac{n_k^{-1}\sum_{j=1}^{n_k}I(x_j\geqslant x_i)\delta_j/\widehat{G}(x_j)-E\{I(T>x_i)\|\boldsymbol{q}_i,x_i\}}{E\{I(T>x_i)\|\boldsymbol{q}_i,x_i\}}+o_p(n_k^{-1/2}).
\end{aligned}
$$

Using derivations similar to the IPW analysis, we first get some basic facts. For any function $\boldsymbol{f}(\boldsymbol{q}_i,x_i)$, we have

$$
\begin{aligned}
\sum_{j=1}^{n_k}\frac{\delta_j\boldsymbol{f}(\boldsymbol{q}_j,x_j)}{\widehat{G}(x_j)} &= \sum_{j=1}^{n_k}\frac{\delta_j\boldsymbol{f}(\boldsymbol{q}_j,t_j)}{G(x_j)}+n^{-1}\sum_{j=1}^{n_k}\frac{\delta_j\boldsymbol{f}(\boldsymbol{q}_j,x_j)}{\widehat{G}(x_j)}\int\frac{Y_j(u)}{G(u)}\frac{dM^c(u)}{\widehat{S}(u^-)}\\
&= \sum_{j=1}^{n_k}\frac{\delta_j\boldsymbol{f}(\boldsymbol{q}_j,t_j)}{G(x_j)}+\frac{n_k}{n}\int\frac{E\{\boldsymbol{f}(\widetilde{\boldsymbol{q}}_k,T)I(T\geqslant u)\|\widetilde{\boldsymbol{q}}_k\}dM^c(u)}{G(u)S(u)}+o_p(n_k^{1/2}).
\end{aligned}
$$

Here we use $\widetilde{\boldsymbol{q}}_k$ to represent the common $\boldsymbol{q}_i$ value in the kth group. Using the above result, we have

$$
\begin{aligned}
& n_k^{-1}\sum_{j=1}^{n_k}\boldsymbol{\phi}(\boldsymbol{q}_j,x_j)I(x_j\geqslant x_i)\delta_j/\widehat{G}(x_j)-E\{\boldsymbol{\phi}(\boldsymbol{q}_i,T)I(T>x_i)\|\boldsymbol{q}_i,x_i\}\\
= \;& n_k^{-1}\sum_{j=1}^{n_k}\frac{\delta_j\boldsymbol{\phi}(\boldsymbol{q}_j,t_j)I(t_j\geqslant x_i)}{G(x_j)}-E\{\boldsymbol{\phi}(\boldsymbol{q}_i,T)I(T>x_i)\|\boldsymbol{q}_i,x_i\}\\
& +\frac{1}{n}\int\frac{E\{\boldsymbol{\phi}(\boldsymbol{q}_i,T)I(T\geqslant x_i)I(T\geqslant u)\|\boldsymbol{q}_i,x_i\}dM^c(u)}{G(u)S(u)}+o_p(n_k^{-1/2}),\\
& n_k^{-1}\sum_{j=1}^{n_k}I(x_j\geqslant x_i)\delta_j/\widehat{G}(x_j)-E\{I(T>x_i)\|\boldsymbol{q}_i,x_i\}\\
= \;& n_k^{-1}\sum_{j=1}^{n_k}\frac{\delta_jI(t_j\geqslant x_i)}{G(x_j)}-E\{I(T>x_i)\|\boldsymbol{q}_i,x_i\}\\
& +\frac{1}{n}\int\frac{E\{I(T\geqslant x_i)I(T\geqslant u)\|\boldsymbol{q}_i,x_i\}dM^c(u)}{G(u)S(u)}+o_p(n_k^{-1/2}).
\end{aligned}
$$

Inserting these forms, we have

$$
\begin{aligned}
& \widehat{\boldsymbol{h}}_{\text{eff},i}(x_i)-\boldsymbol{h}_{\text{eff},i}(x_i)\\
= \;& \frac{n_k^{-1}\sum_{j=1}^{n_k}\delta_j\boldsymbol{\phi}(\boldsymbol{q}_j,t_j)I(t_j\geqslant x_i)/G(x_j)}{E\{I(T>x_i)\|\boldsymbol{q}_i,x_i\}}-\frac{n_k^{-1}\sum_{j=1}^{n_k}\boldsymbol{h}_{\text{eff},i}(x_i)\delta_jI(t_j\geqslant x_i)/G(x_j)}{E\{I(T>x_i)\|\boldsymbol{q}_i,x_i\}}\\
& +\frac{1}{E\{I(T>x_i)\|\boldsymbol{q}_i,x_i\}}\frac{1}{n}\int\frac{E\{\boldsymbol{\phi}(\boldsymbol{q}_i,T)I(T\geqslant x_i)I(T\geqslant u)\|\boldsymbol{q}_i,x_i\}dM^c(u)}{G(u)S(u)}\\
& -\frac{\boldsymbol{h}_{\text{eff},i}(x_i)}{E\{I(T>x_i)\|\boldsymbol{q}_i,x_i\}}\frac{1}{n}\int\frac{E\{I(T\geqslant x_i)I(T\geqslant u)\|\boldsymbol{q}_i,x_i\}dM^c(u)}{G(u)S(u)}+o_p(n_k^{-1/2}).
\end{aligned}
$$

Summing up the n_k such terms in the kth group, exchanging the summation on i and j, and writing $\boldsymbol{a}(\boldsymbol{q}_i,t_i)=E\{\boldsymbol{h}_{\text{eff},i}(C)I(C\leqslant t_i)\|\boldsymbol{q}_i\}$, we obtain

$$
\begin{aligned}
\sum_{i=1}^{n_k}(1-\delta_i)\left\{\widehat{\boldsymbol{h}}_{\text{eff},i}(x_i)-\boldsymbol{h}_{\text{eff},i}(x_i)\right\} \;=\; & \sum_{i=1}^{n_k}\frac{\delta_i\boldsymbol{\phi}(\boldsymbol{q}_i,t_i)}{G(x_i)}\{1-G(t_i)\}-\sum_{i=1}^{n_k}\frac{\delta_i}{G(x_i)}\boldsymbol{a}(\boldsymbol{q}_i,t_i)\\
& +\frac{n_k}{n}\int E[\boldsymbol{\phi}(\boldsymbol{q}_i,T)\{1-G(T)\}I(T\geqslant u)\|\widetilde{\boldsymbol{q}}_k]\frac{dM^c(u)}{G(u)S(u)}\\
& -\frac{n_k}{n}\int E\left\{\boldsymbol{a}(\widetilde{\boldsymbol{q}}_k,T)I(T\geqslant u)\|\widetilde{\boldsymbol{q}}_k\right\}\frac{dM^c(u)}{G(u)S(u)}+o_p(n_k^{1/2}).
\end{aligned}
$$

In the above derivation, we used the fact that the censoring survival function in the group is the same as the global survival function $G(t)$. Now summing up all the m groups, we have

$$\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n}(1-\delta_i)\left\{\widehat{\boldsymbol{h}}_{\text{eff},i}(x_i)-\boldsymbol{h}_{\text{eff},i}(x_i)\right\} \\
= \; & n^{-1/2}\sum_{i=1}^{n}\frac{\delta_i\boldsymbol{\phi}(\boldsymbol{q}_i,t_i)}{G(x_i)} - n^{-1/2}\sum_{i=1}^{n}\delta_i\boldsymbol{\phi}(\boldsymbol{q}_i,t_i) - n^{-1/2}\sum_{i=1}^{n}\frac{\delta_i}{G(x_i)}\boldsymbol{a}(\boldsymbol{q}_i,t_i) \\
& +n^{-1/2}\sum_{i=1}^{n}\int\frac{\mathcal{B}\left(\boldsymbol{\phi},u\right)}{G(u)}dM_i^c(u) - n^{-1/2}\sum_{i=1}^{n}\int\frac{\mathcal{B}\left\{\boldsymbol{\phi}(\boldsymbol{q},t)G(t),u\right\}}{G(u)}dM_i^c(u) \\
& -n^{-1/2}\sum_{i=1}^{n}\int\frac{\mathcal{B}\left\{\boldsymbol{a}(\boldsymbol{q},t),u\right\}}{G(u)}dM_i^c(u)+o_p(1).
\end{aligned}$$

Thus, we have obtained

$$\begin{aligned}
0 \; = \; & n^{-1/2}\sum_{i=1}^{n}\left\{\boldsymbol{\phi}(\boldsymbol{q}_i,t_i)-\boldsymbol{a}(\boldsymbol{q}_i,t_i)\right\} - n^{-1/2}\sum_{i=1}^{n}\int\frac{\left\{\boldsymbol{\phi}(\boldsymbol{q}_i,t_i)-\boldsymbol{a}(\boldsymbol{q}_i,t_i)-\mathcal{B}(\boldsymbol{\phi}-\boldsymbol{a},u)\right\}dM_i^c(u)}{G(u)} \\
& +n^{-1/2}\sum_{i=1}^{n}(1-\delta_i)\boldsymbol{h}_{\text{eff},i}(x_i) - n^{-1/2}\sum_{i=1}^{n}\int\frac{\mathcal{B}\left\{\boldsymbol{\phi}(\boldsymbol{q},t)G(t),u\right\}}{G(u)}dM_i^c(u)+o_p(1).
\end{aligned}$$

Using similar arguments as in the IPW and AIPW cases,

$$\begin{aligned}
& \left\{\boldsymbol{\phi}(\boldsymbol{q}_i,t_i)-\boldsymbol{a}(\boldsymbol{q}_i,t_i)\right\} - \int\frac{\left\{\boldsymbol{\phi}(\boldsymbol{q}_i,t_i)-\boldsymbol{a}(\boldsymbol{q}_i,t_i)-\mathcal{B}(\boldsymbol{\phi}-\boldsymbol{a},u)\right\}dM_i^c(u)}{G(u)} \\
& +(1-\delta_i)\boldsymbol{h}_{\text{eff},i}(x_i) - \int\frac{\mathcal{B}\left\{\boldsymbol{\phi}(\boldsymbol{q},t)G(t),u\right\}}{G(u)}dM_i^c(u)
\end{aligned}$$

is the ith influence function of the imputation estimator.

References

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

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.

Table S.1: Simulation 2. Bias, empirical standard deviation (emp sd), average estimated standard deviation (est sd), 95% coverage (95% cov) of 11 estimators. Sample size $n = 500$, 20% and 50% censoring rate, 1000 simulations.

		$F_1(t) = 0.5837$				$F_2(t) = 0.5748$		
Estimator	bias	emp sd	est sd	95% cov	bias	emp sd	est sd	95% cov
				Group 1: OLS based, censoring rate =20%				
IPW	0.0080	0.0475	0.0433	0.9130	-0.0022	0.0449	0.0425	0.9300
AIPW	0.0063	0.0418	0.0390	0.9210	-0.0014	0.0399	0.0389	0.9390
IMP	0.0010	0.0390	0.0393	0.9470	-0.0014	0.0392	0.0387	0.9360
				Group 2: WLS based, censoring rate =20%				
IPW	0.0080	0.0475	0.0433	0.9130	-0.0023	0.0449	0.0425	0.9300
AIPW	0.0063	0.0418	0.0390	0.9210	-0.0014	0.0399	0.0389	0.9390
IMP	0.0010	0.0390	0.0393	0.9460	-0.0014	0.0392	0.0387	0.9370
				Group 3: EFF based, censoring rate =20%				
IPW	0.0068	0.0469	0.0431	0.9150	-0.0021	0.0448	0.0423	0.9290
AIPW	0.0066	0.0419	0.0391	0.9260	-0.0019	0.0404	0.0388	0.9360
IMP	0.0021	0.0395	0.0394	0.9380	-0.0024	0.0395	0.0388	0.9350
				Group 4: NPMLE, censoring rate =20%				
type I	0.0003	0.0451	0.0459	0.9470	-0.0017	0.0883	0.0890	0.9170
type II	-0.0220	0.0353	0.0361	0.9110	0.0160	0.0317	0.0312	0.9070
				Group 1: OLS based, censoring rate =50%				
IPW	0.0036	0.0845	0.0779	0.9230	0.0026	0.0817	0.0766	0.9280
AIPW	0.0006	0.0479	0.0477	0.9440	-0.0019	0.0499	0.0486	0.9330
IMP	0.0059	0.0541	0.0561	0.9600	0.0048	0.0557	0.0554	0.9450
				Group 2: WLS based, censoring rate =50%				
IPW	0.0037	0.0846	0.0778	0.9220	0.0025	0.0817	0.0765	0.9260
AIPW	0.0006	0.0481	0.0477	0.9440	-0.0020	0.0499	0.0486	0.9320
IMP	0.0059	0.0541	0.0561	0.9590	0.0048	0.0556	0.0554	0.9450
				Group 3: EFF based, censoring rate =50%				
IPW	-0.0040	0.0837	0.0780	0.9180	0.0016	0.0805	0.0762	0.9300
AIPW	-0.0023	0.0485	0.0479	0.9470	0.0009	0.0501	0.0488	0.9320
IMP	0.0040	0.0546	0.0573	0.9630	0.0067	0.0560	0.0560	0.9430
				Group 4: NPMLE, censoring rate =50%				
type I	0.0006	0.0558	0.0569	0.9540	-0.0038	0.1162	0.1097	0.8990
type II	-0.0488	0.0433	0.0446	0.8040	0.0375	0.0417	0.0417	0.8720

Table S.2: Simulation 3. Bias, empirical standard deviation (emp sd), average estimated standard deviation (est std), 95% coverage (95% cov), and IAB of 11 estimators. Sample size $n = 4500$, 65% censoring rate, 1000 simulations. (IAB based on 250 simulations.)

	$F_1(t) = 0.7580$				$F_2(t) = 0.3446$				IAB	
Estimator	bias	emp sd	est sd	95% cov	bias	emp sd	est sd	95% cov	$F_1(t)$	$F_2(t)$
Group 1: OLS based										
IPW	0.0004	0.0155	0.0151	0.9400	0.0031	0.0272	0.0267	0.9430	0.0911	0.2757
AIPW	0.0005	0.0136	0.0131	0.9380	0.0027	0.0253	0.0245	0.9500	0.0899	0.2356
IMP	0.0004	0.0142	0.0139	0.9410	0.0024	0.0256	0.0256	0.9500	0.0930	0.2664
Group 2: WLS based										
IPW	0.0004	0.0155	0.0151	0.9420	0.0030	0.0273	0.0267	0.9440	0.0903	0.2781
AIPW	0.0004	0.0135	0.0131	0.9400	0.0026	0.0254	0.0244	0.9480	0.0951	0.2572
IMP	0.0003	0.0142	0.0139	0.9410	0.0023	0.0257	0.0256	0.9480	0.0879	0.2765
Group 3: EFF based										
IPW	0.0003	0.0156	0.0151	0.9490	0.0026	0.0269	0.0263	0.9490	0.0900	0.2678
AIPW	-0.0005	0.0138	0.0131	0.9360	0.0036	0.0252	0.0243	0.9460	0.0841	0.3086
IMP	0.0000	0.0142	0.0139	0.9380	0.0025	0.0257	0.0254	0.9510	0.0839	0.2905
Group 4: NPMLE										
type I	0.0015	0.0242	0.0241	0.9490	-0.0028	0.0410	0.0414	0.9500	0.1710	0.2788
type II	-0.0141	0.0133	0.0133	0.8170	0.0892	0.0211	0.0213	0.0120	0.5061	2.6427

Table S.3: Simulation 2. Integrated absolute bias (IAB). Upper half of table: the true censoring distribution is independent of q, $G(t)$ is estimated using a common Kaplan-Meier estimator of the censoring distribution. Lower half of table: the true censoring distribution is subgroup-specific, $G(t)$ is estimated using a common Kaplan-Meier estimator (denoted by †), or a subgroup-specific Kaplan-Meier estimator (denoted by *).

	Censoring rate			
	20%		50%	
Estimator	$F_1(t)$	$F_2(t)$	$F_1(t)$	$F_2(t)$
	Group 1: OLS based			
IPW	0.0746	0.0043	0.0334	0.0094
AIPW	0.0666	0.0044	0.0336	0.0307
IMP	0.0375	0.0040	0.0310	0.0195
	Group 2: WLS based			
IPW	0.0740	0.0085	0.0333	0.0226
AIPW	0.0797	0.0047	0.0364	0.0402
IMP	0.0416	0.0043	0.0483	0.0212
	Group 3: EFF based			
IPW	0.0669	0.0044	0.0172	0.0090
AIPW	0.0692	0.0060	0.0622	0.0229
IMP	0.0474	0.0073	0.0325	0.0191
	Group 4: NPMLE			
type I	0.0373	0.0194	0.0333	0.2461
type II	0.1308	0.0644	0.1475	0.1173
	Group 1: OLS based			
IPW†	0.0505	0.0543	0.0447	0.0321
AIPW†	0.0470	0.0073	0.0433	0.0321
IPW*	0.0654	0.0153	0.0137	0.0094
AIPW*	0.0586	0.0132	0.0317	0.0280
	Group 2: WLS based			
IPW†	0.0501	0.0534	0.0454	0.0467
AIPW†	0.0620	0.0080	0.0400	0.0415
IPW*	0.0656	0.0135	0.0146	0.0215
AIPW*	0.0689	0.0126	0.0335	0.0354
	Group 3: EFF based			
IPW†	0.0546	0.0535	0.0370	0.0320
AIPW†	0.0474	0.0084	0.0651	0.0245
IPW*	0.0608	0.0161	0.0137	0.0085
AIPW*	0.0631	0.0096	0.0596	0.0188

Table S.4: Simulation 2. Bias, empirical standard deviation (emp sd), average estimated standard deviation (est sd), 95% coverage (95% cov) of 11 estimators. The censoring distribution is subgroup-specific, and $G(t)$ is estimated using a common Kaplan-Meier estimator (denoted by †), or a subgroup-specific Kaplan-Meier estimator (denoted by *). Sample size $n = 500$, 20% and 50% censoring rate, 1000 simulations.

	$F_1(t) = 0.5837$				$F_2(t) = 0.5748$			
Estimator	bias	emp sd	est sd	95% cov	bias	emp sd	est sd	95% cov
	Group 1: OLS based, censoring rate =20%							
IPW†	-0.0092	0.0494	0.0459	0.9260	0.0132	0.0479	0.0441	0.9080
AIPW†	0.0053	0.0413	0.0390	0.9280	-0.0020	0.0394	0.0376	0.9350
IPW*	0.0072	0.0429	0.0394	0.9180	0.0008	0.0405	0.0388	0.9360
AIPW*	0.0065	0.0418	0.0381	0.9180	0.0013	0.0397	0.0379	0.9360
	Group 2: WLS based, censoring rate =20%							
IPW†	-0.0092	0.0495	0.0458	0.9250	0.0132	0.0479	0.0441	0.9080
AIPW†	0.0054	0.0413	0.0390	0.9260	-0.0020	0.0394	0.0376	0.9350
IPW*	0.0073	0.0430	0.0394	0.9180	0.0007	0.0406	0.0389	0.9340
AIPW*	0.0065	0.0418	0.0381	0.9180	0.0012	0.0397	0.0379	0.9360
	Group 3: EFF based, censoring rate =20%							
IPW†	-0.0112	0.0502	0.0459	0.9240	0.0133	0.0470	0.0438	0.9110
AIPW†	0.0055	0.0415	0.0391	0.9260	-0.0023	0.0404	0.0376	0.9300
IPW*	0.0063	0.0419	0.0391	0.9270	0.0013	0.0399	0.0386	0.9420
AIPW*	0.0070	0.0413	0.0382	0.9220	0.0006	0.0398	0.0379	0.9360
	Group 1: OLS based, censoring rate =50%							
IPW†	-0.0028	0.0805	0.0764	0.9270	0.0063	0.0814	0.0744	0.9240
AIPW†	0.0009	0.0464	0.0488	0.9560	-0.0024	0.0510	0.0480	0.9260
IPW*	0.0028	0.0530	0.0496	0.9440	-0.0022	0.0545	0.0528	0.9440
AIPW*	0.0016	0.0490	0.0462	0.9460	-0.0016	0.0520	0.0499	0.9360
	Group 2: WLS based, censoring rate =50%							
IPW†	-0.0027	0.0805	0.0764	0.9280	0.0063	0.0815	0.0744	0.9220
AIPW†	0.0009	0.0465	0.0488	0.9570	-0.0025	0.0510	0.0480	0.9260
IPW*	0.0031	0.0535	0.0498	0.9430	-0.0025	0.0545	0.0529	0.9430
AIPW*	0.0017	0.0493	0.0461	0.9480	-0.0017	0.0520	0.0499	0.9360
	Group 3: EFF based, censoring rate =50%							
IPW†	-0.0098	0.0829	0.0769	0.9140	0.0050	0.0792	0.0742	0.9210
AIPW†	-0.0012	0.0471	0.0491	0.9580	-0.0004	0.0516	0.0482	0.9270
IPW*	0.0013	0.0522	0.0495	0.9390	-0.0016	0.0542	0.0527	0.9460
AIPW*	-0.0011	0.0484	0.0464	0.9470	0.0008	0.0519	0.0499	0.9370

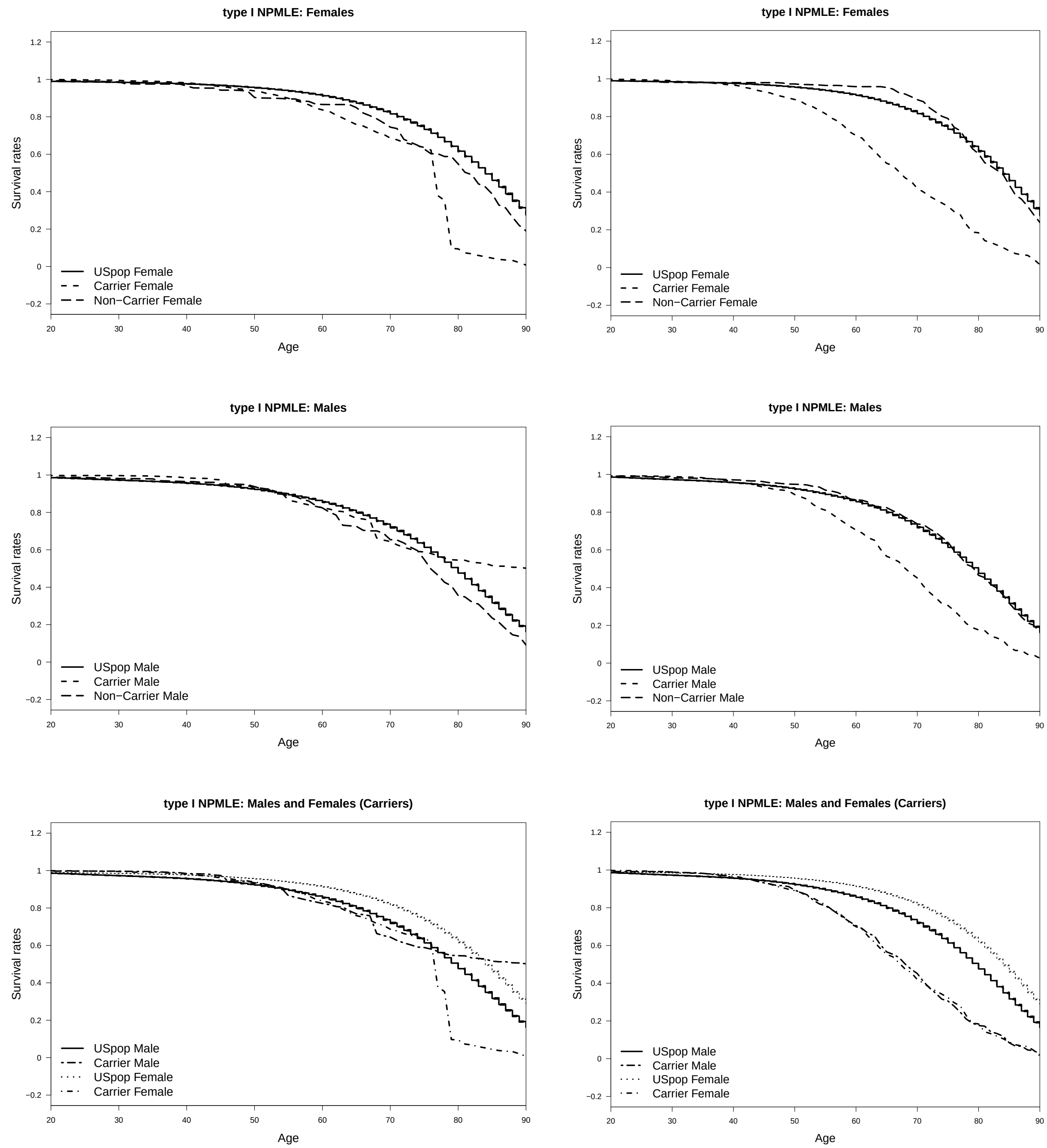

Figure S.1: Estimated type I NPMLE survival curves for the COHORT data stratified by gender using the full sample (left panels) and sub-sample (right panels; removing subjects in $\boldsymbol{Q}_i$ groups with small sample sizes). The curves "USpop Male" and "USpop Female" correspond to Kaplan-Meier estimated survival rates for the general male and female US populations in 2003. Bottom two figures compare the estimated survival curves for the male and female HD gene mutation carrier groups to that of the general male and female US populations.