Journal of the
Royal Statistical Society
SERIES C
Applied
Statistics

Appl. Statist. (2014)
63, *Part* 1, *pp.* 1–23

Estimating disease onset distribution functions in mutation carriers with censored mixture data

Yanyuan Ma

Texas A&M University, College Station, USA

and Yuanjia Wang

Columbia University, New York, USA

[Received September 2012. Final revision April 2013]

Summary. We consider non-parametric estimation of disease onset distribution functions in multiple populations by using censored data with unknown population identifiers. The problem is motivated from studies aiming at estimating the age-specific disease risk distribution in deleterious mutation carriers for genetic counselling and design of therapeutic intervention trials to modify disease progression (i.e. to slow down the development of symptoms and to delay the onset of disease). In some of these studies, the distribution of disease risk in participants assumes a mixture form. Although the population identifiers are missing, study design and scientific knowledge allow calculation of the probability of a subject belonging to each population. We propose a general family of weighted least squares estimators and show that existing consistent non-parametric methods belong to this family. We identify a computationally effortless estimator in the family, study its asymptotic properties and show its significant gain in efficiency compared with the existing estimators in the literature. The application to a large genetic epidemiological study of Huntington's disease reveals information on the age-at-onset distribution of Huntington's disease which sheds light on some clinical hypotheses.

Keywords: Huntington's disease; Mixture observations; Penetrance function; Risk prediction; Unknown population label

1. Introduction

In some scientific studies, it is of interest to estimate the distribution function of an outcome by using data arising from a mixture of multiple populations with unknown population identifiers. For example, in Huntington's disease (HD) research, one of the major goals is to estimate the distribution of the age at onset of HD (subject to censoring) in HD gene mutation carriers. Accurate estimation of the distribution function in carriers is important for genetic counselling, which is a process of informing patients or relatives at risk of an inherited disorder on the consequences and nature of the disorder, the probability of developing it and advising on care management and family planning. It is also useful in designing clinical trials of therapeutics modifying disease progression, and it provides estimation of positive and negative predictive values of a genetic test (Heagerty and Zheng, 2005). In some studies such as the 'Cooperative Huntington's observational research trial' (COHORT) (Dorsey *et al.*, 2012), initial participants (probands) underwent a clinical evaluation and were genotyped for HD mutation. Through a systematic family history interview, they also reported ages at onset of disease of their rela-

Address for correspondence: Yuanjia Wang, Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 West 168th Street, New York, NY 10032, USA.
E-mail: yuanjia.wang@columbia.edu

0035–9254/14/63001

tives. However, most of the relatives are not genotyped and their mutation status is unknown. Thus, relatives are a sample of a mixture of carrier and non-carrier populations with unknown population identifiers, and the probability that a subject belongs to a population is calculated on the basis of Mendelian inheritance. Such a design where probands are genotyped and provide disease onset times (subject to censoring) of their relatives through a family history interview is commonly applied to study the distribution of a disease in mutation carriers (Marder *et al.*, 2003; Wang *et al.*, 2007, 2008; Dorsey *et al.*, 2012). Note that, here and throughout the text, we refer to the collection of all subjects with a particular genetic variant (such as carrying the mutation or carrying the wild type) as a population.

Another example of studies collecting data with similar structure is quantitative trait locus studies. Quantitative trait loci (QTL) are hypothesized specific chromosomal regions containing genes that make significant contributions to the expression of a complex trait. QTL are generally identified by comparing the linkage (degree of covariation) of polymorphic molecular markers and phenotypic trait measurements. These polymorphic molecular markers are called flanking markers. In a QTL study, subjects are genotyped at known locations along their genome, and the goals are to determine the location of the gene influencing manifestation of a quantitative trait. The genotypes at the typed markers are known for a subject, but they are missing for locations in between markers. Under a standard interval mapping framework (Lander and Botstein, 1989; Wu *et al.*, 2007), a subject's phenotype trait distribution is a mixture of QTL genotype-specific distributions, where the mixing proportions are obtained from the design of experiment, location and genotypes at the flanking markers and genetic distance between the markers and the QTL (see, for example, Wu *et al.* (2007)). In many cases, the quantitative outcome of interest, such as the time to flowering of a plant (Ferreira *et al.*, 1995; Lin and Wu, 2006), is subject to right censoring.

The research goal of both types of study can be formulated as estimating distribution functions for censored outcomes arising from multiple populations although for some subjects it is unknown from which population they are drawn. The probability that an observation belongs to each population can be calculated through taking into account the scientific knowledge and the experiment design. Modelling the distribution in each population parametrically, e.g. through a Gaussian mixture model (McLachlan and Peel, 2000), and proceeding with the usual maximum likelihood estimation is one choice. To be more flexible and to leave the distribution in each population completely model free, Wacholder *et al.* (1998) investigated a non-parametric model and proposed a non-parametric maximum likelihood estimator (NPMLE). Two other non-parametric estimators were developed. One aimed at overcoming some limitations of the original NPMLE such as ensuring monotonicity (Chatterjee and Wacholder, 2001), and the other aimed at improving estimation efficiency (Fine *et al.*, 2004). Since the proposal in Chatterjee and Wacholder (2001) is also an NPMLE, to distinguish it from the original estimator in Wacholder *et al.* (1998), the original proposal is named NPMLE1 and the modified version NPMLE2 here. The estimator in Fine *et al.* (2004) exploits the independence assumption between the censoring times and event times, and hence is named IND.

When using these methods to analyse the COHORT data, we observe that the existing non-parametric methods are inadequate. For example, when using IND and NPMLE1 to estimate the cumulative risk of HD of the HD mutation carriers, a cumulative risk greater than 1 was obtained at ages 65 years and older by using IND, whereas NPMLE1 provided an estimation of a risk less than 0.4 at all ages (Figs 1(b) and 1(c)). These results do not agree with the clinical literature on HD (e.g. Langbehn *et al.* (2004)). This observation, together with the established result that both IND and NPMLE1 are consistent estimators (Fine *et al.* 2004; Wacholder *et al.*,

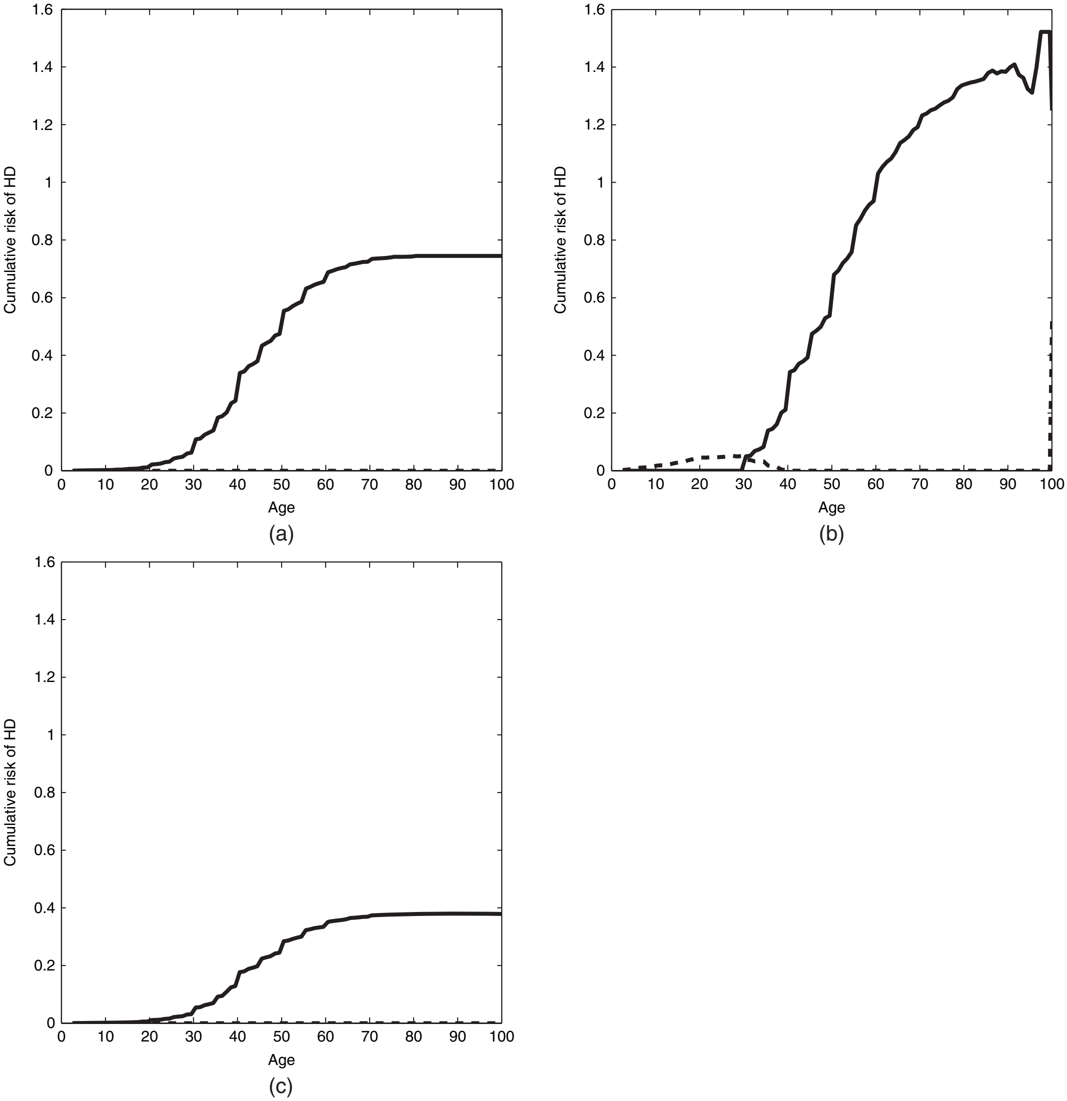


Fig. 1. COHORT study: cumulative risk of onset of HD based on (a) weighted least squares, (b) IND and (c) NPMLE1 (———, risk of the carrier population; ------, risk of the non-carrier population)

1998), motivated this work to examine variability and efficiency of NPMLE1 and IND. In fact, our analysis in Section 3 reveals the inefficiency of these methods, in the sense that the estimation variability can be further reduced and, using an improved estimator, results that are consistent with clinical knowledge can be obtained (see Fig. 1(a)). In addition, IND requires each subject to have a positive probability of being observed to have an event at all time points of interest, which is not satisfied in the COHORT study and many other chronic disease studies. In the COHORT study, not every family member will eventually develop HD before death; therefore the probability of being censored is 1 for these subjects. Finally, NPMLE2 is not a consistent estimator (Ma and Wang, 2012; Wang *et al.*, 2012).

To provide valid estimation and to improve stability and efficiency, we propose a general family of weighted least squares (WLS) type estimators. We derive the asymptotically optimal member of this family and identify a computationally efficient estimator that has competitive performance compared with the optimal member. We demonstrate the relationship of the WLS family with the existing methods and with a class of imputation-based methods that have not been proposed for this type of problems in the literature.

The rest of this paper is organized as follows. In Section 2, we describe the motivating example (the COHORT study) in detail and examine some initial analysis results comparing IND, NPMLE1 and WLS. In Section 3, we propose the WLS family, identify the recommended estimator within this family and derive its asymptotic properties for inference. We study its relationship to the existing estimators and provide insights on the limitations of the existing estimators. In Section 4, we carry out simulation studies to demonstrate the finite sample properties to illustrate our theoretical findings. In Section 5, we provide further analyses with COHORT data to estimate the age-at-onset distribution for HD gene mutation carriers from family members who may not be genotyped. We examine the connection between the estimated risk function and positive predictive value of the HD mutation test. Lastly, we conclude this work with some discussions in Section 6.

2. Description and initial analysis of the data

Before we introduce the methods proposed, we first describe the motivating study. HD is an autosomal dominant neurodegenerative disease that is caused by an unstable expansion of trinucleotide repeat 'C–A–G' (CAG repeat) at the ITI5 gene on chromosome 4 which codes for a protein named huntingtin (Huntingtons Disease Collaborative Research Group, 1993). Subjects with a CAG repeat length of 36 or more are considered to be HD mutation positive (i.e. CAG expanded) and the majority of them will develop HD in the life course if not censored by death, whereas subjects with a CAG repeat length less than 36 do not develop HD (Rubinsztein *et al.*, 1996; Nance *et al.*, 1998). The COHORT study is an observational study that was designed to collect clinical and genetic data from a sample of symptomatic and presymptomatic HD mutation carriers and their family members. Details of the study design are discussed in Dorsey *et al.* (2012).

In the COHORT study, the initial participants (probands) were followed for 5 years and provided information on whether their family members had experienced HD in past years or developed HD during the follow-up years. The age at onset was recorded if a relative had experienced the disease and age at death recorded if a relative had died. 4735 relatives from 786 families were included in the analysis. The total number of relatives who had experienced HD is 1184. Most of the relatives were not genotyped. However, since each relative is genetically related to the probands, the relationship information and the proband's mutation status are used to obtain $m=6$ distinct values of the probabilities of carrying the HD mutation under the Mendelian transmission assumption. These six probability values for a relative to be a carrier are 0, 0.25, 0.5, 0.75, 0.97 and 1, with 1329 (relatives of non-carrier probands), 141 (grandparents of carrier probands with one CAG expanded allele), 2010 (parents or siblings of carrier probands with one CAG expanded allele), two (siblings of carrier probands with two CAG expanded alleles), 1183 (relatives of carrier probands with one CAG expanded allele and developed HD) and 70 (relatives with a confirmed CAG expanded allele) observations in the corresponding group. These relatives' current ages are distributed between 10 and 100 years. We are interested in estimating the distribution of age at onset of HD for HD mutation carriers (CAG lengths 36 or longer) using exclusively the relative data.

We now first show the results of analysing the COHORT data by the existing consistent estimators IND and NPMLE1, and compare them with the WLS method that will be proposed in Section 3. From Fig. 1, we see that IND provides highly non-smooth estimates for the carrier group at several ages (30 and 90 years) and has an estimate that is much larger than one at older ages. It also provided some positive estimates for the non-carrier group, which is inconsistent with clinical knowledge, since subjects without HD mutation do not develop HD (Rubinsztein *et al.*, 1996; Nance *et al.*, 1998). The performance problems for IND are encountered because, in some of the m subgroups, the estimation is not valid because of the smaller censoring process support than the event process support, and subsequently the estimates in such groups adversely influence the overall estimates when they are combined to form IND. NPMLE1 provides an estimated cumulative risk of below 40% at age 80 years, which may be too low compared with the existing clinical literature (e.g. Langbehn *et al.* (2004)). The unsatisfactory performance of NPMLE1 can be related to the small sample sizes in some groups. Although the Kaplan–Meier estimator is not accurate in these groups, the corresponding result is not downweighted in NPMLE1. Further investigation of these methods shows that they are consistent estimators, which suggests that estimation variability related to inefficiency may have given rise to unexpected estimates in practice. The theoretical examination in the next section presents some explanations of the limitation of both methods in terms of efficiency. Finally, we show that the proposed WLS estimates of the cumulative risk of HD are 33.9% (95% confidence interval [32.0%, 35.8%]) by age 40 years and 74.5% (95% confidence interval [73.9%, 76.0%]) by age 80 years for carriers. These results are within the same range as the weighted averages of estimates provided in Langbehn *et al.* (2004) for the population with CAG lengths between 36 and 41 and the population with CAG lengths greater than 41.

3. A family of weighted least squares estimators

We now introduce methods to address the research goal of estimating distribution functions in studies such as the COHORT study. Suppose that there are p populations ($p=2$ in the COHORT study, the carrier and non-carrier populations) and, in the jth population, the time to event of interest (such as onset of HD in the COHORT study) has differentiable cumulative distribution functions $F_j(t)$, $j=1,\ldots,p$. The corresponding probability density functions are $f_1(t),\ldots,f_p(t)$. Let $\mathbf{F}(t)=(F_1(t),\ldots,F_p(t))^{\mathrm{T}}$ and $\mathbf{f}(t)=(f_1(t),\ldots,f_p(t))^{\mathrm{T}}$. Assume that the ith ($i=1,\ldots,n$) subject is randomly sampled from these p populations, where the probability that this observation belongs to the kth population is q_{ik} for $k=1,\ldots,p$. Thus, we can write the ith observation as $(\mathbf{q}_i,S_i)$, where $\mathbf{q}_i=(q_{i1},\ldots,q_{ip})^{\mathrm{T}}$, and S_i is a random event time. Further assume that the n observations are independent of each other; hence the event times within each population are independent. Note that $\Sigma_{k=1}^{p}q_{ik}=1$, and in most applications, including both QTL analysis and proband–family studies, the $\mathbf{q}_i$s are known quantities computed on the basis of knowledge in a study design (e.g. QTL experiment design or the relationship of a relative to the proband).

In all studies of interest here, $\mathbf{q}_i$ takes only $m<\infty$ different vector values which we denote by $\mathbf{u}_1,\ldots,\mathbf{u}_m$, and we assume that there are r_j observations corresponding to each of the $\mathbf{u}_j$s for $j=1,\ldots,m$, so that $\Sigma_{j=1}^{m}r_j=n$. For example, in the COHORT data that were described in Section 2, $m=6$ and the $\mathbf{u}_j$ and r_js were specified. Assume further that the ith observation is censored at C_i, and the censoring times are independent of the survival times. Note that we also allow the situation that the censoring distribution has smaller support than the support of the event times. In summary, an observation subject to censoring can be written as $(\mathbf{q}_i,Y_i,\delta_i)$, where $Y_i = \min(C_i, S_i)$ and $\delta_i = I(S_i \leqslant C_i)$. The observations are assumed to be ordered so that $Y_1 < Y_2 \ldots < Y_{n-1} < Y_n$. Our interest lies in estimating the p distribution functions $F_1(t), \ldots, F_p(t)$ and making inference.

To illustrate these notations by using the studies that we introduced in Section 1 and 2, note that S_i can be the age at onset of an event (e.g. the time to onset of HD). For the HD study, q_{ik} is the probability that the ith relative carries the kth genotype at the HD gene given the proband's genotype status, and $F_k(t)$ is the distribution function of S_is within the subjects with the kth genotype. An autosomal dominant disease yields $p = 2$, and an additive genetic model yields $p = 3$. Each of the p components of $\mathbf{F}(t)$ thus captures the probability of developing a disease by a certain age for subjects with a certain mutation status. For $p = 3$ the first and second components of $\mathbf{F}$(t) are referred to as the penetrance function for homozygous or heterozygous mutation respectively in the genetics literature. In the QTL studies, q_{ik} is the probability that a subject carries the kth genotype at the QTL given genotypes at the flanking markers. The dimension p depends on the experimental design; for example, $p = 2$ for a back-cross experiment and $p = 3$ for an intercross experiment. To see this, assume that the parental generation has alleles MM and mm; then the first generation (F1) has genotype Mm. The F1-generation is then crossed with the parental generation and each back-cross individual has probability 0.5 of having genotype Mm and probability 0.5 of having genotype mm (i.e. $p = 2$). Intercross individuals result from crossing F1-individuals and therefore have genotypes MM, Mm or mm (i.e. $p = 3$). In either situation, since genotypes may not be observed, the distribution of S_i is a mixture of $F_1, \ldots, F_p$, i.e. $\mathbf{q}_i^{\mathrm{T}}\mathbf{F}$.

Taking advantage of the finiteness of m, we propose first to estimate the distribution of the outcomes in each of the m fixed mixture groups, and then we use a familiar WLS estimate to retrieve the distribution $\mathbf{F}$. Specifically, denote $H_j(t) = \mathbf{u}_j^{\mathrm{T}}\mathbf{F}(t)$ for $j = 1, \ldots, m$, and let $\mathbf{H}(t) = (H_1(t), \ldots, H_m(t))^{\mathrm{T}}$. Obviously, $H_j(t)$ is a valid cumulative distribution function and can be estimated by using all observations with $\mathbf{q}_i = \mathbf{u}_j$ for $i = 1, \ldots, n$. For convenience, the collection of observations with $\mathbf{q}_i = \mathbf{u}_j$ is denoted (Y_{ji}, δ_{ji}) for $i = 1, \ldots, r_j$, and we also assume that they are ordered so that $Y_{j1} < Y_{j2} < \ldots < Y_{jr_j-1} < Y_{jr_j}$ for all $j = 1, \ldots, m$. Denote an estimated distribution function as $\hat{H}_j(t)$ and let $\hat{\mathbf{H}}(t) = (\hat{H}_1(t), \ldots, \hat{H}_m(t))^{\mathrm{T}}$. Denote the matrix $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$. From $\mathbf{H}(t) = \mathbf{U}^{\mathrm{T}}\mathbf{F}(t)$, we easily obtain a WLS family of estimators,

$$\hat{\mathbf{F}}(t) = (\mathbf{U}\mathbf{W}\mathbf{U}^{\mathrm{T}})^{-1}\mathbf{U}\mathbf{W}\hat{\mathbf{H}}(t), \tag{1}$$

where $\mathbf{W}$ is an $m \times m$ weight matrix.

3.1. The proposed estimator and its inference

Within the WLS family, we propose to use a diagonal matrix, which is denoted $\mathbf{R}$, with $r_1, \ldots, r_m$ as diagonal elements, as the weight matrix and use a classical Kaplan–Meier estimator in the jth group to obtain $\hat{H}_j(t)$ for $j = 1, \ldots, m$. The resulting estimator has a simple form:

$$\begin{aligned}\hat{\mathbf{F}}(t) &= (\mathbf{U}\mathbf{R}\mathbf{U}^{\mathrm{T}})^{-1}\mathbf{U}\mathbf{R}\hat{\mathbf{H}}(t)\\ &= \left(\sum_{j=1}^{m} r_j \mathbf{u}_j \mathbf{u}_j^{\mathrm{T}}\right)^{-1} \sum_{j=1}^{m} r_j \mathbf{u}_j \hat{H}_j(t).\end{aligned} \tag{2}$$

Because the Kaplan–Meier estimator is known to be root n consistent (Kaplan and Meier, 1958), we can easily obtain that the estimator $\hat{\mathbf{F}}(t)$ is root n consistent. The asymptotic covariance of $\hat{F}(t)$ can be estimated as

$$\widehat{\mathrm{cov}}\{\hat{\mathbf{F}}(t)\}=\left(\sum_{j=1}^{m}r_j\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\right)^{-1}\left\{\sum_{j=1}^{m}r_j^2\hat{\sigma}_j^2(t)\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\right\}\left(\sum_{j=1}^{m}r_j\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\right)^{-1},$$

where

$$\hat{\sigma}_j^2(t)=\{1-\hat{H}_j(t)\}^2\sum_{Y_{ji}\leqslant t}\delta_{ji}/\{(r_j-i)(r_j-i+1)\}.$$

This result provides an easy way to perform hypothesis testing. For example, to test $H_0:\mathbf{a}^{\mathrm{T}}\mathbf{F}(t)=c$ *versus* $H_1:\mathbf{a}^{\mathrm{T}}\mathbf{F}(t)\neq c$ or $H_1:\mathbf{a}^{\mathrm{T}}\mathbf{F}(t)<c$ or $H_1:\mathbf{a}^{\mathrm{T}}\mathbf{F}(t)>c$ for any length p vector $\mathbf{a}$ and any constant c, the Wald-type test statistic is

$$T=\{\mathbf{a}^{\mathrm{T}}\hat{\mathbf{F}}(t)-c\}/[\mathbf{a}^{\mathrm{T}}\widehat{\mathrm{cov}}\{\hat{\mathbf{F}}(t)\}\mathbf{a}]^{1/2}.$$

The statistic T has a standard normal distribution under hypothesis H_0. When $\mathbf{a}=(1,\ -1,\ 0,\ldots,0)^{\mathrm{T}}$ and $c=0$, this corresponds to testing whether the subjects from the first and second population have the same distribution at t, which is a research question that is often encountered in practice.

It is also possible to perform the test simultaneously at several different t-values. For example, let $\mathbf{t}=(t_1,\ldots,t_l)^{\mathrm{T}}$ and assume that $t_1<\ldots<t_l$. Let $F(t)$ be a $p\times l$ matrix with jth column corresponding to time $t_j:\mathbf{F}(\mathbf{t})=(\mathbf{F}(t_1),\ldots,\mathbf{F}(t_l))$. Let $\mathbf{c}$ be a length l vector. Suppose that we wish to test $H_0:\mathbf{a}^{\mathrm{T}}\mathbf{F}(\mathbf{t})=\mathbf{c}^{\mathrm{T}}$ *versus* $H_1:\mathbf{a}^{\mathrm{T}}\mathbf{F}(\mathbf{t})\neq\mathbf{c}^{\mathrm{T}}$. Denote the Kaplan–Meier estimator $\hat{\mathbf{H}}_j(\mathbf{t})=(\hat{H}_j(t_1),\ldots,\hat{H}_j(t_l))$ and its variance–covariance matrix as $\mathbf{V}_j(\mathbf{t})$. Using the asymptotic properties of the Kaplan–Meier estimator (Kaplan and Meier, 1958), we know that $\mathbf{V}_j(\mathbf{t})$ can be estimated by $\hat{\mathbf{V}}_j(\mathbf{t})$, where the (a,b)th entry is

$$\hat{V}_{j,a,b}=\{1-\hat{H}_j(t_a)\}\{1-\hat{H}_j(t_b)\}\sum_{Y_{ji}\leqslant t_a}\delta_{ji}/\{(r_j-i)(r_j-i+1)\}\qquad\text{for any }1\leqslant a\leqslant b\leqslant l.$$

Thus, we can form the test statistic

$$T=\{\mathbf{a}^{\mathrm{T}}\hat{\mathbf{F}}(\mathbf{t})-\mathbf{c}^{\mathrm{T}}\}\left[\sum_{j=1}^{m}\{\mathbf{a}^{\mathrm{T}}(\mathbf{URU}^{\mathrm{T}})^{-1}\mathbf{URe}_j\}^2\hat{\mathbf{V}}_j\right]^{-1}\{\mathbf{a}^{\mathrm{T}}\hat{\mathbf{F}}(\mathbf{t})-\mathbf{c}^{\mathrm{T}}\}^{\mathrm{T}},$$

where $\mathbf{e}_j$ is a length m vector with 1 on the jth entry and 0 elsewhere. Under hypothesis H_0, T has a χ^2-distribution with degrees of freedom l. The motivation of this statistic is to standardize $\mathbf{a}^{\mathrm{T}}\hat{F}(\mathbf{t})$. A direct calculation yields

$$\begin{aligned}\mathbf{a}^{\mathrm{T}}\hat{\mathbf{F}}(\mathbf{t})&=\mathbf{a}^{\mathrm{T}}\{\hat{\mathbf{F}}(t_1),\ldots,\hat{\mathbf{F}}(t_l)\}=\mathbf{a}^{\mathrm{T}}\{(\mathbf{URU}^{\mathrm{T}})^{-1}\mathbf{UR}\hat{\mathbf{H}}(t_1),\ldots,(\mathbf{URU}^{\mathrm{T}})^{-1}\mathbf{UR}\hat{\mathbf{H}}(t_l)\}\\&=\mathbf{a}^{\mathrm{T}}(\mathbf{URU}^{\mathrm{T}})^{-1}\mathbf{UR}\{\hat{\mathbf{H}}(t_1),\ldots,\hat{\mathbf{H}}(t_l)\}=\sum_{j=1}^{m}\mathbf{a}^{\mathrm{T}}(\mathbf{URU}^{\mathrm{T}})^{-1}\mathbf{URe}_j\hat{\mathbf{H}}_j(\mathbf{t}).\end{aligned}$$

Because the m different groups do not overlap, this yields the variance

$$\sum_{j=1}^{m}\{\mathbf{a}^{\mathrm{T}}(\mathbf{URU}^{\mathrm{T}})^{-1}\mathbf{URe}_j\}^2\mathbf{V}_j.$$

A useful case in practice is when $\mathbf{a}=(1,\,-1,0,\,\ldots,0)^{\mathrm{T}}$ and $\mathbf{c}=0$. This corresponds to testing whether the first and second populations have the same distribution simultaneously at all values in the vector $\mathbf{t}$.

Testing $H_0:\mathbf{a}^{\mathrm{T}}\mathbf{F}(t)=c(t)$ at all t-values is also possible, where $c(t)$ is an arbitrary deterministic function of t. From Breslow and Crowley (1974), $r_j^{1/2}\{\hat{H}_j(t)-H_j(t)\}$ converges weakly to a Gaussian process for $j=1,\ldots,m$ with mean 0 and an explicit covariance function. Thus $R(t)=$

$\mathbf{a}^{\mathrm{T}}\mathbf{F}(t)-c(t)$ as a linear combination of the $\hat{H}_j(t)$s also has the similar property of converging weakly to a Gaussian process. One can form a test statistic such as a Kolmogorov–Smirnov-type statistic $\sup_{t\in[0,\tau]} R(t)$ (Fleming *et al.*, 1980) or $\int_0^{\tau} R(t)\,\mathrm{d}t$ (Pepe and Fleming, 1989) and derive their asymptotic null distributions.

However, the asymptotic distributions might not always be suitable to use in practice. One reason is that the approximation at the large value of t can be quite imprecise. The second reason is that the above asymptotic results are valid only in the region $H_j(t)<1$. In practice, some of the populations might have a smaller support than others. Hence, depending on the $\mathbf{u}_j$-values, for the same t-value, some $H_j(t)$ might be smaller than 1 whereas others might be 1. This creates complications in practice, especially because it is often not known which $H_j(t)$ has what support. The third reason is that only when r_j is large will the asymptotic expression be a close approximation. However, in practice, some of the r_j-values can be quite small. Because of these reasons, we propose to use an alternative permutation approach when the asymptotic results are not suitable.

When $p=2$, a test of interest is whether there is a difference between distributions of mutation carriers and non-carriers, i.e. $H_0: F_1(t)=F_2(t)$, either at a finite set of t-values or for the entire range. A permutation strategy can be used (Churchill and Doerge, 1994) in this case. Specifically, we permute the (Y_i,δ_i) pairs and couple them with $\mathbf{q}_1,\ldots,\mathbf{q}_n$-values to create a permuted sample, and we use estimator (2) to obtain a new estimate of $\mathbf{F}(t)$ and a permuted test statistic $\hat{F}_1(t)-\hat{F}_2(t)$. Repeat this process a sufficiently large number of times to obtain the empirical distribution of $\hat{F}_1(t)-\hat{F}_2(t)$ under hypothesis H_0.

In what follows, we further explore the WLS family of estimators (1) and provide a justification for our recommendation (2). We also show that the two existing methods NPMLE1 and IND are non-ideal members of the WLS family.

3.2. Choice of group estimation

We first study the competing methods in estimating $\hat{H}_j(t)$ for $j=1,\ldots,m$ in family (1). It is easy to see that $H_j(t)$ is the distribution function of S_is for the collection of observations that satisfy $\mathbf{q}_i=\mathbf{u}_j$. Thus, estimation within the $\mathbf{u}_j$-group is a classical problem of estimating distribution functions with randomly censored data. The familiar Kaplan–Meier estimator is known to be the maximum likelihood estimator in this setting (Kaplan and Meier, 1958; Wellner, 1982) and hence provides the most efficient estimate for each $H_j(t)$. Thus this is the optimal choice. An additional advantage is that, other than the independent censoring assumption, no additional requirement needs to be imposed on the relationship between the censoring process and the event process for the Kaplan–Meier estimator to be valid.

NPMLE1 in Wacholder *et al.* (1998) proceeds by performing non-parametric maximum likelihood in each of the m groups, and recovering $\hat{\mathbf{F}}(t)$ via $\hat{\mathbf{F}}(t)=(\mathbf{U}\mathbf{U}^{\mathrm{T}})^{-1}\mathbf{U}\hat{H}(t)$. Hence it is a member of the WLS family. It makes the choice of using Kaplan–Meier estimation in estimating $\hat{H}_j(t)$.

IND proposed in Fine *et al.* (2004) makes a different choice in estimating $H_j(t)$ and then recovers $\hat{\mathbf{F}}(t)$ via $\hat{\mathbf{F}}(t)=(\mathbf{URU}^{\mathrm{T}})^{-1}\mathbf{UR}\hat{\mathbf{H}}(t)$. Hence it is also a member of the WLS family. To estimate $\mathbf{H}(t)$, IND exploits the independence of the event process and the censoring process, and uses the relationship $\Pr(Y_i>t)=\Pr(S_i>t)\Pr(C_i>t)$. The IND estimates $H_j(t)$ through

$$\hat{H}_j(t)=1-\frac{1}{\hat{G}(t)}\left\{\frac{1}{\sum_{i=1}^{n} I(\mathbf{q}_i=\mathbf{u}_j)}\sum_{i=1}^{n} I(\mathbf{q}_i=\mathbf{u}_j)\,I(Y_i\geqslant t)\right\},$$

where $\hat{G}(t)$ is a Kaplan–Meier estimate of the survival function, $G(t)=\Pr(C_i>t)$, of the censoring process. This method originates from Ying *et al.* (1995). However, it has several limitations compared with a direct Kaplan–Meier estimator of $H_j(t)$. First, the method can only be used in the region where $G(t)>0$. Therefore in the situations where the censoring process has a smaller support than the event process, and if t is larger than the upper limit of the possible censoring time, the method ceases to be valid. This is so with the HD study data and in our second simulation. Second, it is less efficient than maximum likelihood estimation, which is reflected in our simulation results. Third, it is not easy to obtain a variance estimate of IND. Finally, although a Kaplan–Meier estimation is avoided in the estimation related to the event process, it is still used in the estimation of the censoring process. Hence it does not provide a computational advantage.

3.3. Choice of weights

Although the Kaplan–Meier estimator in each group (i.e. $\hat{H}_j(t)$) is asymptotically efficient (Wellner, 1982), it does not necessarily guarantee that $\hat{F}(t)$ is asymptotically efficient. A good choice of weights improves efficiency. Since, for different j-values, the $\hat{H}_j(t)$s are estimated by using distinct observations, $\hat{H}_1(t),\ldots,\hat{H}_m(t)$ are mutually independent. Thus, the optimal weight matrix $\mathbf{W}$ should be diagonal. Let the diagonal elements of $\mathbf{W}$ be $w_1,\ldots,w_m$. The estimation variance of the WLS family (1) is

$$\operatorname{cov}\{\hat{F}(t)\}=\left(\sum_{j=1}^{m} w_j\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\right)^{-1}\left\{\sum_{j=1}^{m} w_j^2\sigma_j^2(t)\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\right\}\left(\sum_{j=1}^{m} w_j\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\right)^{-1},$$

where $\sigma_j^2(t)$ is the variance of the estimator $\hat{H}_j(t)$. Thus, theoretically, by letting $w_j=1/\sigma_j^2(t)$, we would obtain the optimal weights in terms of estimation efficiency within the WLS family.

Although this is the optimal weighting strategy in theory, in practice, we observe that it is often suboptimal. We provide several explanations. First, $\sigma_j^2(t)$ is not known and can only be estimated in practice. Although asymptotically this estimate itself does not cause a loss of efficiency for any WLS estimator, it creates numerical instability in finite samples. This instability can be especially harmful when some groups contain very few observations, because the estimation of $\sigma_j^2(t)$ can be noisy. Second, occasionally, it may happen that in one of the groups, say the j_0th group, the last observation is not censored and its observed event time $S_{r_{j_0}}$ is smaller than t, which is the time at which we are interested in estimating $\mathbf{H}_{j_0}(t)$. In this case, the Kaplan–Meier estimator yields $\hat{H}_{j_0}(t)=1$, and the estimated variance $\hat{\sigma}_{j_0}^2(t)=0$. Although this can be handled numerically either by assigning an upper limit on the weight w_{j_0} or by solving a constrained least squares problem instead of directly implementing WLS, the numerical instability that is caused by this phenomenon still persists. Intrinsically, this is caused by the fact that we cannot assess $\sigma_{j_0}^2(t)$ sufficiently well at the upper limit of the data. For example, $\hat{\sigma}_{j_0}^2(t)=0$ might be a suitable estimate of the variability of $\hat{H}_{j_0}(t)$ if the event process indeed has the support to the left of t, and it might not be by chance that all the observations are to the left of t.

Since Ma and Wang (2012) observed that, in the absence of censoring, equally weighting each observation and weighting each observation by its inverse variance exhibited very little difference in terms of estimation efficiency, we propose simply to assign equal weights to each observation in the same $\mathbf{u}_j$-group. This results in the weight choice of $w_j=r_j$ in equation (2), which is a direct result from the fact that, if each observation has a weight of 1, then the group with r_j observations receives a weight of r_j. This weighting strategy is simple and extremely stable in computation. In the simulations in Section 4, we did not find any better weighting scheme other than this simple choice, even including the theoretically optimal weights calculated by using the true $H_j(t)$ functions.

Inspecting the weighting choice of IND, we find that it is the same as our proposal of $w_j = r_j$. NPMLE1 in contrast makes the choice of assigning $w_j = 1$ for $j = 1, \ldots, m$. This choice unnecessarily downweights the observations belonging to the larger groups and consequently diminishes the advantage of the more accurately estimated $H_j(t)$s. In practice, we find that this choice leads to a substantial loss of efficiency. In addition, it can also be vulnerable to some degenerated groups. For example, when a group contains only one observation, the Kaplan–Meier estimate in this group is certainly not reliable, yet this estimate is allowed to enter the final estimator of $\hat{\mathbf{F}}(t)$ with the same importance as the other estimates that can greatly influence the final result.

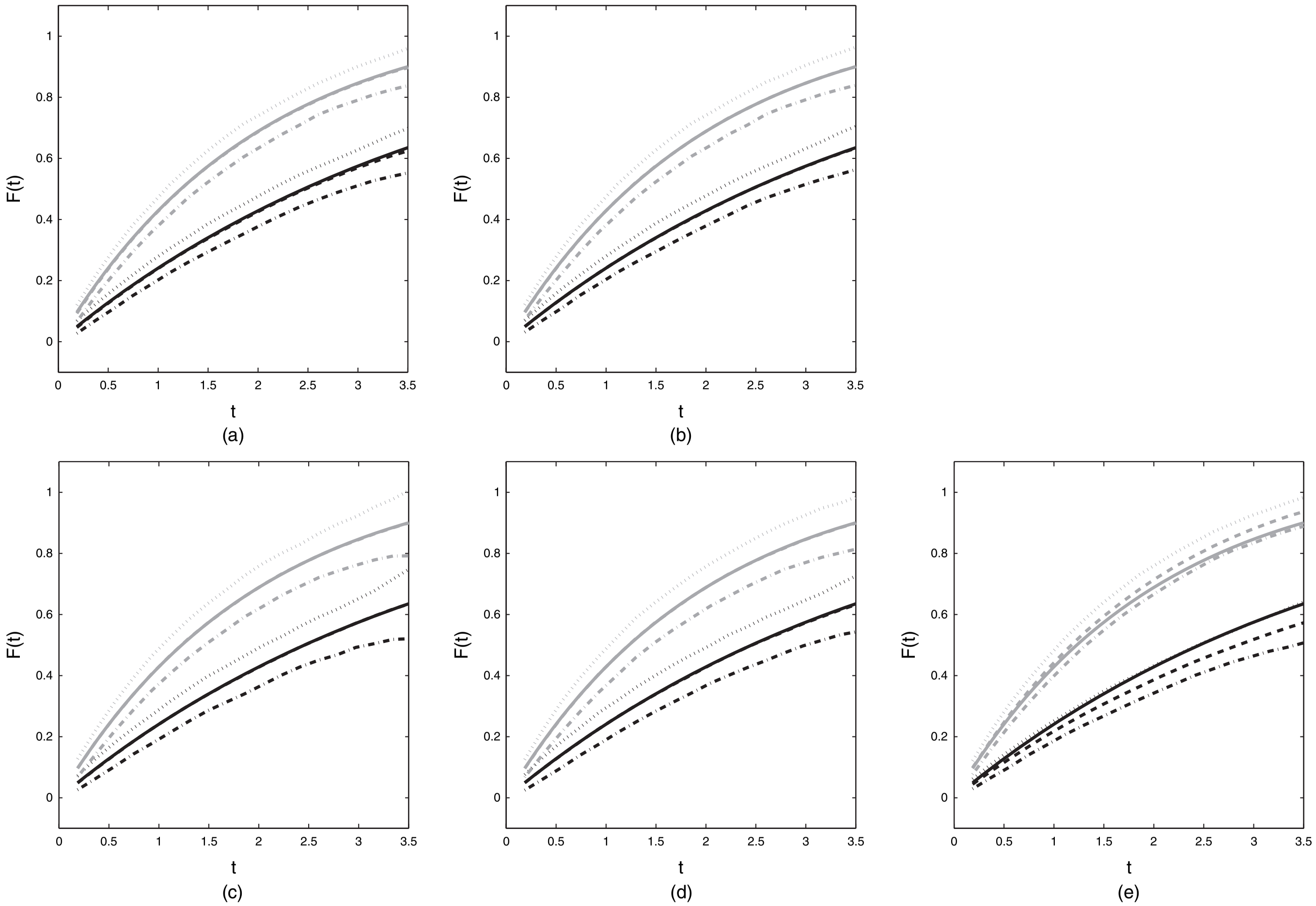
4. Simulation study

We performed two simulation studies to illustrate the finite sample performance of several estimation and inference procedures discussed above. In the first study, we generated a total of 1000 repetitions, each with the sample size $n = 1000$. The data were generated from a mixture of $p = 2$ different populations, with $m = 3$ different mixing probability vectors. The first two mixing groups contain approximately 40% and 5% of the observations, and the remaining 55% observations are in the third mixing group. The true population distributions are both truncated exponential, with support [0, 10] and [0, 5]. We generated censoring times from a uniform distribution between 0 and 3.9. This results in an approximately 50% censoring rate. We performed both estimation and testing under this simulation design. When investigating

Table 1. First simulation study: estimation bias and empirical standard errors in estimating distribution functions $\mathbf{F}(t)$ at three different t-values by using five different estimators†

Method	*bias(F_1)*	*SD(F_1)*	*bias(F_2)*	*SD(F_2)*
$t = 0.9750,\ \mathbf{F}(t) = (0.2357,\ 0.4203)^{\mathrm{T}}$				
Oracle	−0.0010	0.0229	−0.0009	0.0229
WLS	0.0004	0.0229	0.0004	0.0229
IND	−0.0000	0.0289	0.0009	0.0289
NPMLE1	0.0002	0.0310	0.0002	0.0310
NPMLE2	−0.0222	0.0200	0.0136	0.0200
$t = 1.9500,\ \mathbf{F}(t) = (0.4203,\ 0.6785)^{\mathrm{T}}$				
Oracle	−0.0034	0.0297	−0.0008	0.0297
WLS	−0.0014	0.0298	0.0005	0.0298
IND	−0.0013	0.0387	0.0004	0.0387
NPMLE1	−0.0018	0.0383	0.0001	0.0383
NPMLE2	−0.0418	0.0268	0.0251	0.0268
$t = 2.9250,\ \mathbf{F}(t) = (0.5651,\ 0.8371)^{\mathrm{T}}$				
Oracle	−0.0048	0.0349	−0.0019	0.0349
WLS	−0.0007	0.0347	−0.0001	0.0347
IND	0.0003	0.0479	−0.0012	0.0479
NPMLE1	−0.0016	0.0444	−0.0010	0.0444
NPMLE2	−0.0551	0.0324	0.0337	0.0324

†The results are based on 1000 simulations with sample size $n = 1000$.



(a)

(b)

(c)

(d)

(e)

Fig. 2. First simulation study—true cumulative distribution function (———) and the mean (------), and 95% pointwise confidence bands (· · · · · ·, -·-·-·-) of the estimated cumulative distribution functions (the grey and the black curves represent the two distinct populations; the oracle and WLS estimators are almost identical and have superior performance; IND, NPMLE1 and NPMLE2 have either wider bands or are biased): (a) oracle; (b) WLS; (c) IND; (d) NPMLE1; (e) NPMLE2

Table 2. First simulation study: estimation bias, empirical standard error, average estimated standard error and coverage of 95% confidence intervals for estimating $\mathbf{F}(t)$ by using WLS†

t	$\mathbf{F}(t)$	*bias(F)*	*sd(F)*	$\widehat{sd}(F)$	*95% confidence interval*
Under $H_0 : F_1(t) = F_2(t)$					
0.9750	0.4203	−0.0010	0.0264	0.0262	0.9360
	0.4203	0.0004	0.0300	0.0291	0.9400
1.9500	0.6785	−0.0002	0.0287	0.0280	0.9440
	0.6785	0.0003	0.0308	0.0310	0.9470
2.9250	0.8371	−0.0002	0.0276	0.0276	0.9520
	0.8371	−0.0010	0.0308	0.0306	0.9410
Under $H_1 : F_1(t) \neq F_2(t)$					
0.9750	0.2357	0.0004	0.0229	0.0225	0.9500
	0.4203	0.0004	0.0288	0.0284	0.9460
1.9500	0.4203	−0.0014	0.0298	0.0290	0.9380
	0.6785	0.0005	0.0327	0.0320	0.9490
2.9250	0.5651	−0.0007	0.0347	0.0348	0.9500
	0.8371	−0.0001	0.0336	0.0338	0.9480

†The empirical standard error is the sample standard deviation of 1000 estimates from 1000 simulations; the estimated standard errors are calculated from the asymptotic variance formula of the general WLS estimators.

the type I error rate under hypothesis H_0, we set both distributions to be the same truncated exponential with support [0, 5], while keeping everything else unchanged. This results in a censoring rate of about 40%.

We implemented our proposed WLS method, as well as the existing methods including IND, NPMLE1 and NPMLE2, where NPMLE2 is obtained through maximizing

$$\sum_{i=1}^{n} \log\{\mathbf{q}_i^{\mathrm{T}} \mathbf{f}(Y_i)\}^{\delta_i} \log\{1 - \mathbf{q}_i^{\mathrm{T}} \mathbf{F}(Y_i)\}^{1-\delta_i}$$

with respect to $\mathbf{F}(Y_i)$s by treating $\mathbf{F}(t)$ as a piecewise constant monotonically increasing function. For illustration, we also provided the oracle WLS method, where the optimal weights $w_j = 1/\mathrm{var}\{\hat{H}_j(t)\}$ are used, with $\hat{H}_j(t)$ being the Kaplan–Meier estimator in the jth group, and the variance is calculated by plugging in the true distribution functions $H_j(t)$ in the variance formula. The estimation results at three representative time points are provided in Table 1, where they are at the beginning, middle and end of the range of the observed time points. The estimation results for the entire curves are depicted in Fig. 2.

NPMLE2 shows a very large bias in comparison with all the other methods, whereas both IND and NPMLE1 have larger estimation variability in comparison with the proposed WLS method. For example, at $t = 1.95$, the bias of NPMLE2 is about 25–30 times larger than the other three consistent estimators (WLS, IND and NPMLE1), and the empirical standard errors of IND and NPLME1 are 30% and 29% larger than that of WLS respectively. The gain in efficiency is more notable towards the higher end of the t-values. When $t = 2.92$, the improvement in empirical standard errors of the proposed WLS estimator over IND and NPMLE1 is 38% and 30% respectively. The WLS estimator has very small biases, and its estimation variance is about the same as the oracle WLS (the difference is 2% or less).

Table 3. First simulation study: empirical rejection rates for single, multiple and curve testing at various nominal levels by using WLS†

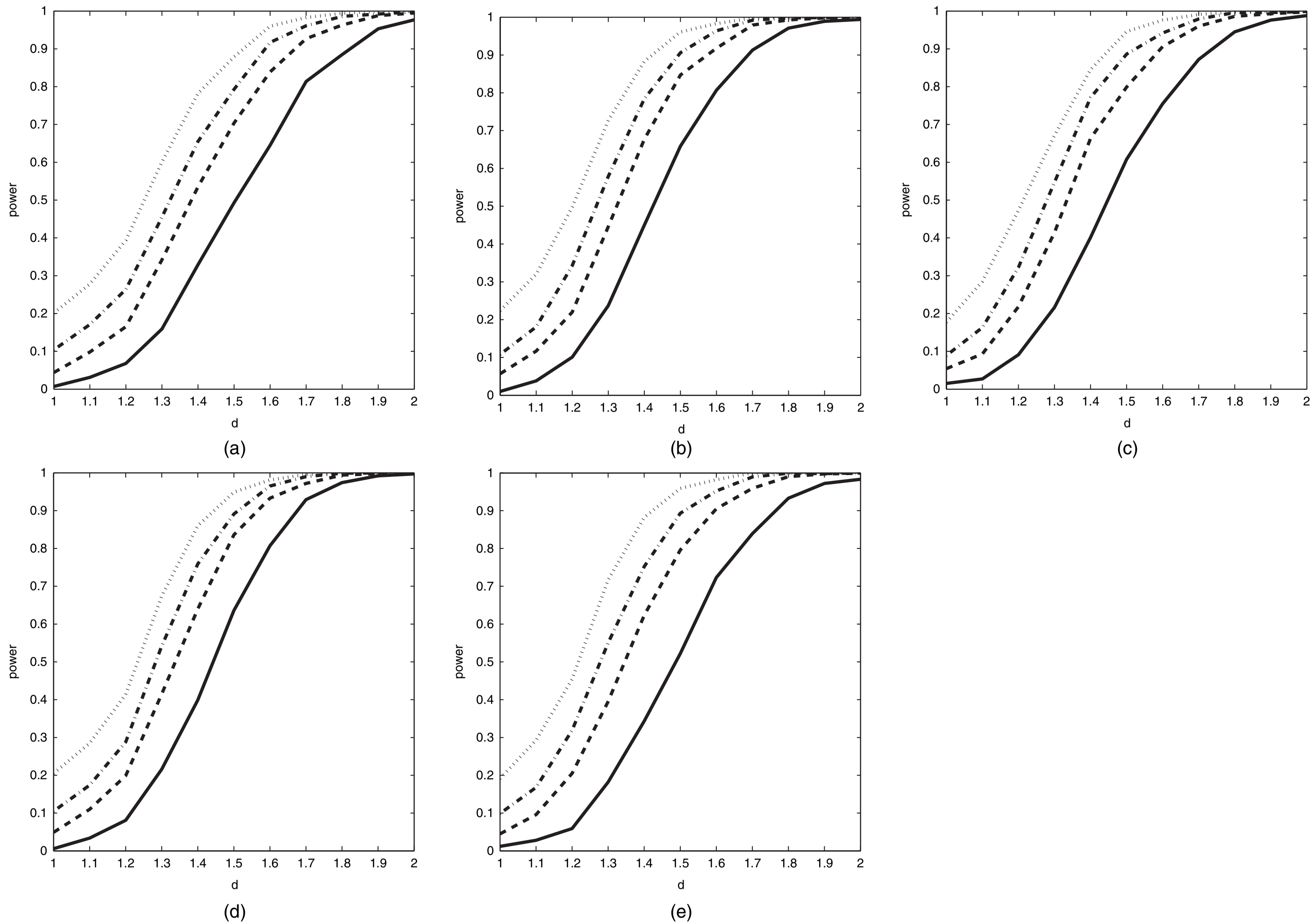
t-value	*Results for the following nominal levels:*			
	0.01	*0.05*	*0.1*	*0.2*
Under $H_0 : F_1(t) = F_2(t)$				
0.9750	0.0120	0.0470	0.1050	0.2060
1.9500	0.0130	0.0610	0.0970	0.1830
2.9250	0.0080	0.0510	0.0970	0.2050
Multiple t	0.0090	0.0540	0.1070	0.2080
Curve	0.0150	0.0620	0.1070	0.2090
Under $H_1 : F_1(t) \neq F_2(t)$				
0.9750	0.9790	0.9950	0.9980	0.9980
1.9500	0.9950	0.9990	1.0000	1.0000
2.9250	0.9920	0.9970	0.9990	1.0000
Multiple t	0.9990	1.0000	1.0000	1.0000
Curve	0.9780	0.9970	0.9990	0.9990

†Multiple t is the result of testing $F_1(t) = F_2(t)$ at the three listed t-values simultaneously. Curve is the result of testing $F_1(t) = F_2(t)$ at all t. Results are based on 1000 simulations with sample size $n = 1000$.

We also examined several tests based on the proposed WLS estimator. We report estimation and the single-point, multiple-point and curve testing results in Table 2 and Table 3. Table 2 shows the finite sample bias of the estimated cumulative distribution functions, their empirical standard errors, average estimated standard errors and 95% confidence interval coverage at the three representative time points under the null and the alternative hypotheses. It is seen that the estimation biases are small, the estimation standard errors are well estimated and the 95% confidence interval coverages are close to their nominal level. For the single- and multiple-time-point testing, we used the test statistics that were proposed in Section 3.1 and their asymptotic null distributions to compute p-values. For testing the entire difference between two distribution functions, we used the test statistic $\sup_t |F_1(t) - F_2(t)|$ and performed 1000 permutations to compute its p-value. It is seen from Table 3 that the type I error rates of all three tests adhere to their nominal levels. In addition, the power of the three tests is comparable.

To gain a more comprehensive understanding of the power performance of these tests, we further adjusted the first component of $\mathbf{F}(t)$ to be a truncated exponential on $[0, 5d]$ and let d gradually change from 1 to 2. We plotted the power of the tests as a function of d in Fig. 3. As expected, the power increases when d increases. In other words, the power increases when the two components in $\mathbf{F}(t)$ separate from each other.

We perform a second set of simulation studies to illustrate the finite sample performance of the estimators in the situation that resembles the COHORT data. In this study, we generated 5000 observations from $p = 2$ populations, with $m = 6$ distinct $\mathbf{q}$-values exactly the same as in the COHORT data. We set the different group values to be 1500, 2000, 1200, 200, 98 and 2, and set the censoring process to have smaller support than the event process, with censoring rate 75%. These are all designed to be similar to the COHORT data. The results of the five estimators of $\mathbf{F}(t)$ are in Fig. 4. Once again, it is clear that NPMLE2 is severely biased, whereas



(a)

(b)

(c)

(d)

(e)

Fig. 3. First simulation study—power of the tests $H_0 : F_1(t) = F_2(t)$ as a function of d at (a) $t = 0.975$, (b) $t = 1.95$, (c) $t = 2.925$, (d) multiple t and (e) curve t (larger d indicates a larger deviation from H_0): ——, level = 0.01; - - - - - -, level = 0.05; · - · - · -, level = 0.1; ⋯⋯, level = 0.2

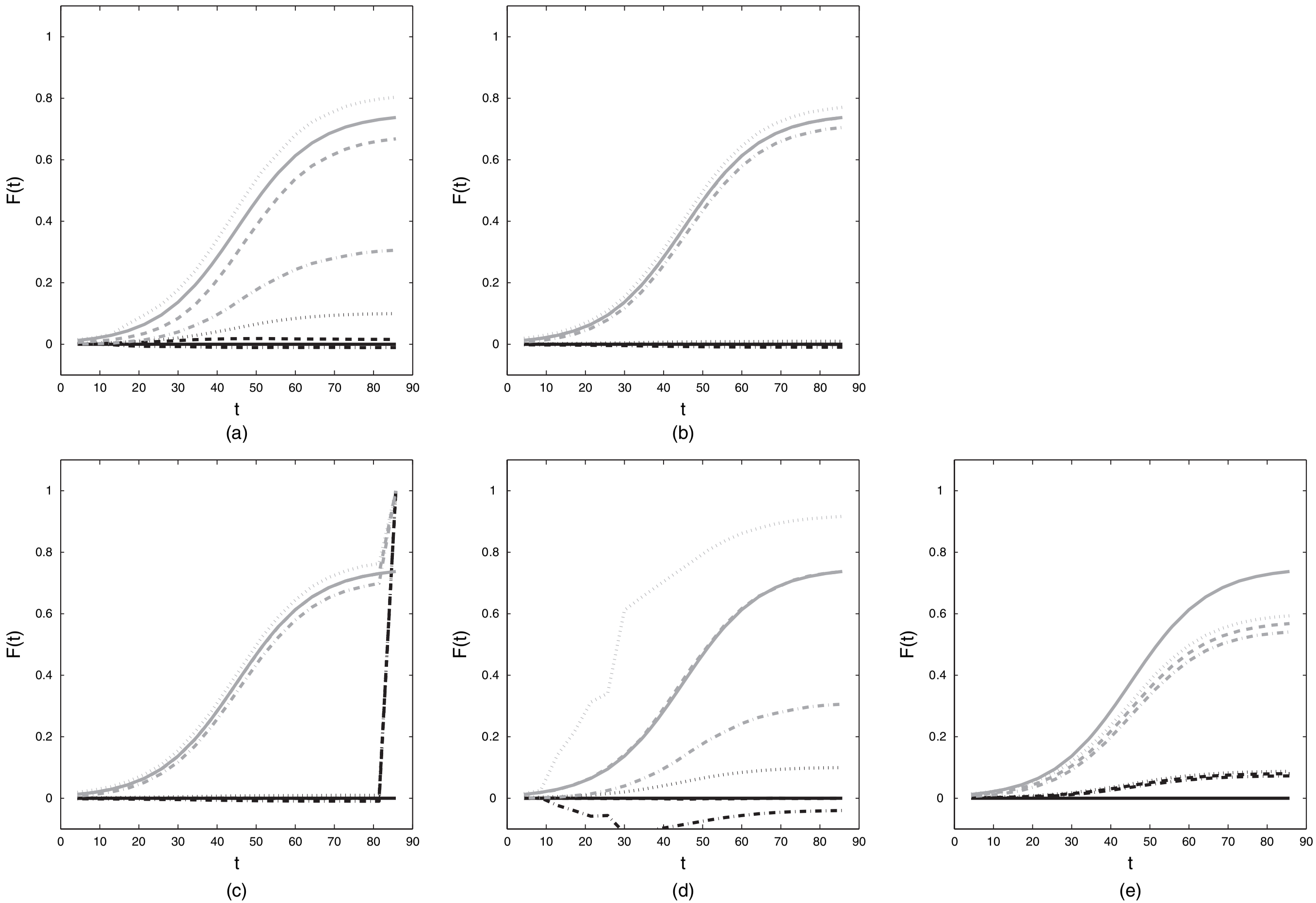
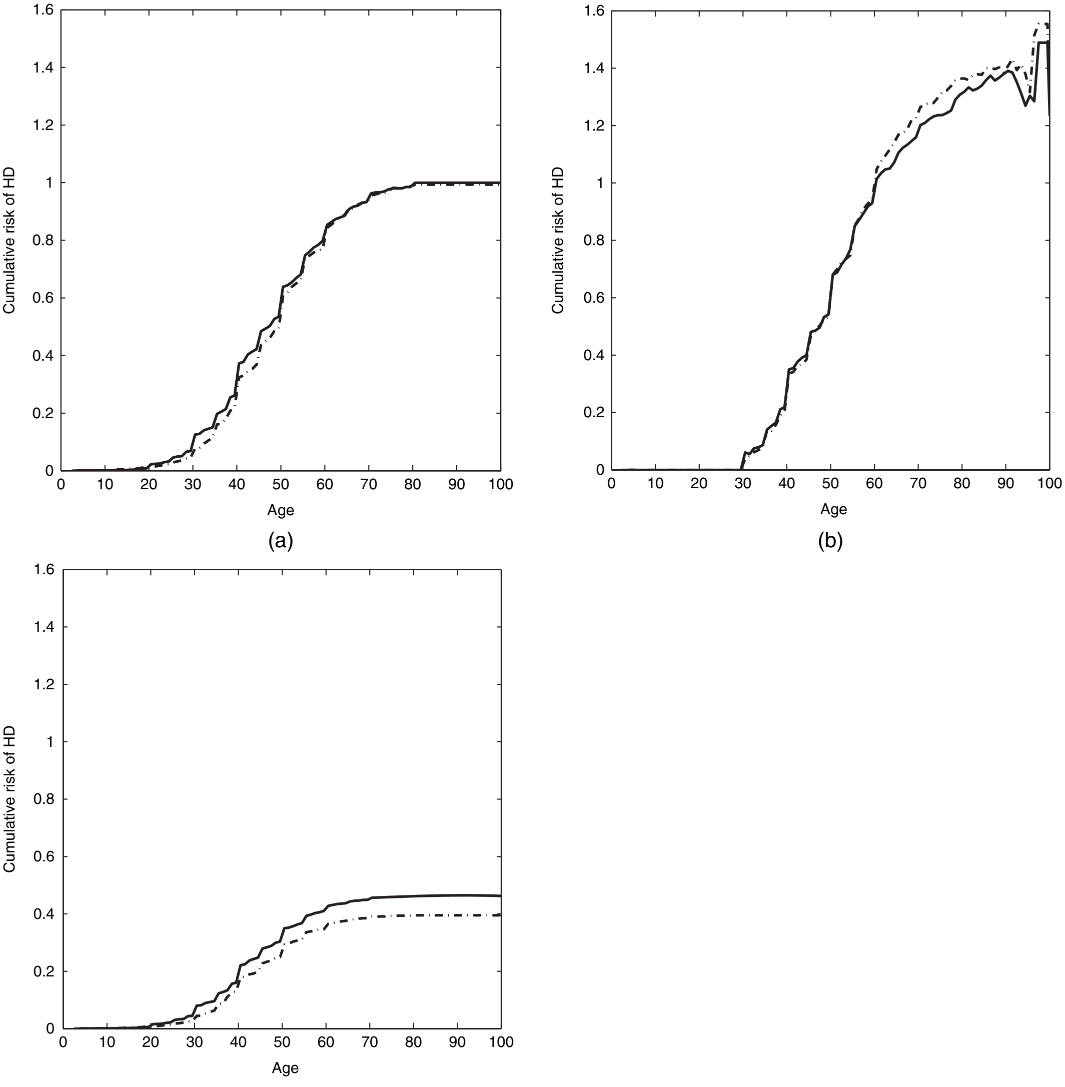


Fig. 4. Second simulation study—true cumulative distribution function (———) and the mean (------), and 95% pointwise confidence bands (· · · · · ·, -·-·-·-) of the estimated cumulative distribution functions (the grey and the black curves represent the two distinct populations; the oracle and WLS estimators are almost identical and have superior performance; IND, NPMLE1 and NPMLE2 have either wider bands or are biased): (a) oracle; (b) WLS; (c) IND; (d) NPMLE1; (e) NPMLE2



(a)

(b)

(c)

Fig. 5. COHORT study: cumulative risk of onset of HD based on (a) WLS, (b) IND and (c) NPMLE1 stratified by gender: ———, females; ·-·-·-, males

the proposed WLS estimator has desirable performance in that it shows very small bias and has the narrowest confidence band. IND breaks down towards the end of the study range, which is a consequence of the censoring process being supported on a subset of the event process. The performance of NPMLE1 is greatly hampered by the small sample size in one group. This is reflected in the large variability. Interestingly, the oracle estimator is also not well behaved. This is because the oracle estimator uses the asymptotic variance of the Kaplan–Meier estimator. However, for groups with small or moderate sample sizes, the finite sample performance of the Kaplan–Meier estimator is more relevant and it may be very different from the inference based on the asymptotic variance.

5. Additional analyses of the study data

In Section 2, we provided some initial analyses of the COHORT data by using IND, NPMLE1 and WLS. Here we provide more detailed analyses using WLS. First, we estimate the disease distribution functions stratified by gender. Fig. 5 presents the estimated cumulative distribution of age at onset of HD for males and females. We present the same three estimators as in the overall analysis and similar conclusions for these estimators can be drawn comparing WLS, NPMLE1 and IND. The proposed WLS estimator suggests that females might have a slightly elevated risk than males across a wide range of ages. We performed a permutation test of the difference between the entire distribution curves of female and male carriers as introduced in Section 3.1 and obtained a p-value of 0.083.

Furthermore, we estimate the distribution functions stratified by both gender and whether a subject reported an affected father or affected mother at the time of the family history interview (Fig. 6). We observe that female carriers with an affected father had a slightly higher risk than female carriers with an affected mother across a wide range of ages. In contrast, male carriers with an affected father had a similar risk compared with male carriers with an affected mother until age 60 years, and after age 60 years the risk in the former is slightly higher. The test comparing the difference between female carriers with an affected father with female carriers with an affected mother had a p-value of 0.096. These results are consistent with a potential anticipation effect: it is observed that a male could transmit an expanded CAG repeats sequence to his offspring, which may increase the likelihood of an earlier age at onset in the offspring (Ranen *et al.*, 1995; Wexler *et al.*, 2004). Our analysis suggests that the anticipation effect might manifest in female offspring across a wide range of ages, whereas for male offspring the anticipation effect might not manifest until about age 60 years. Further analysis on an independent sample is needed to corroborate these observations. Finally, a test comparing female or male carriers who reported an affected parent (either mother or father) with female or male carriers who did not report any affected parent at the time of interview is significant

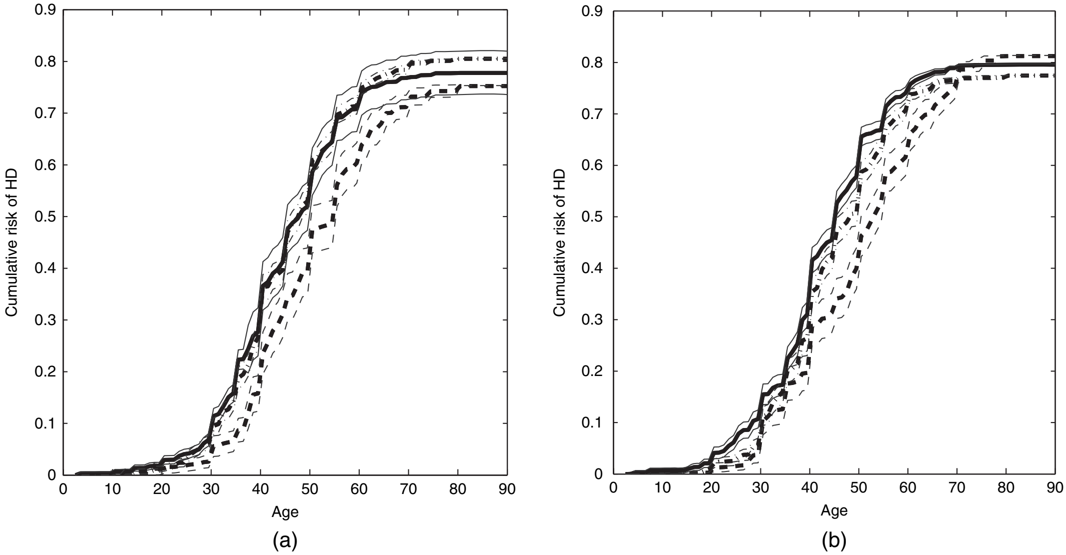


Fig. 6. COHORT study—cumulative risk of onset of HD stratified by gender and the status of reporting affected father (——), affected mother (·-·-·-) or none (------) at the time of family history interview, based on WLS (the lighter curves represent the 90% pointwise confidence band): (a) males; (b) females

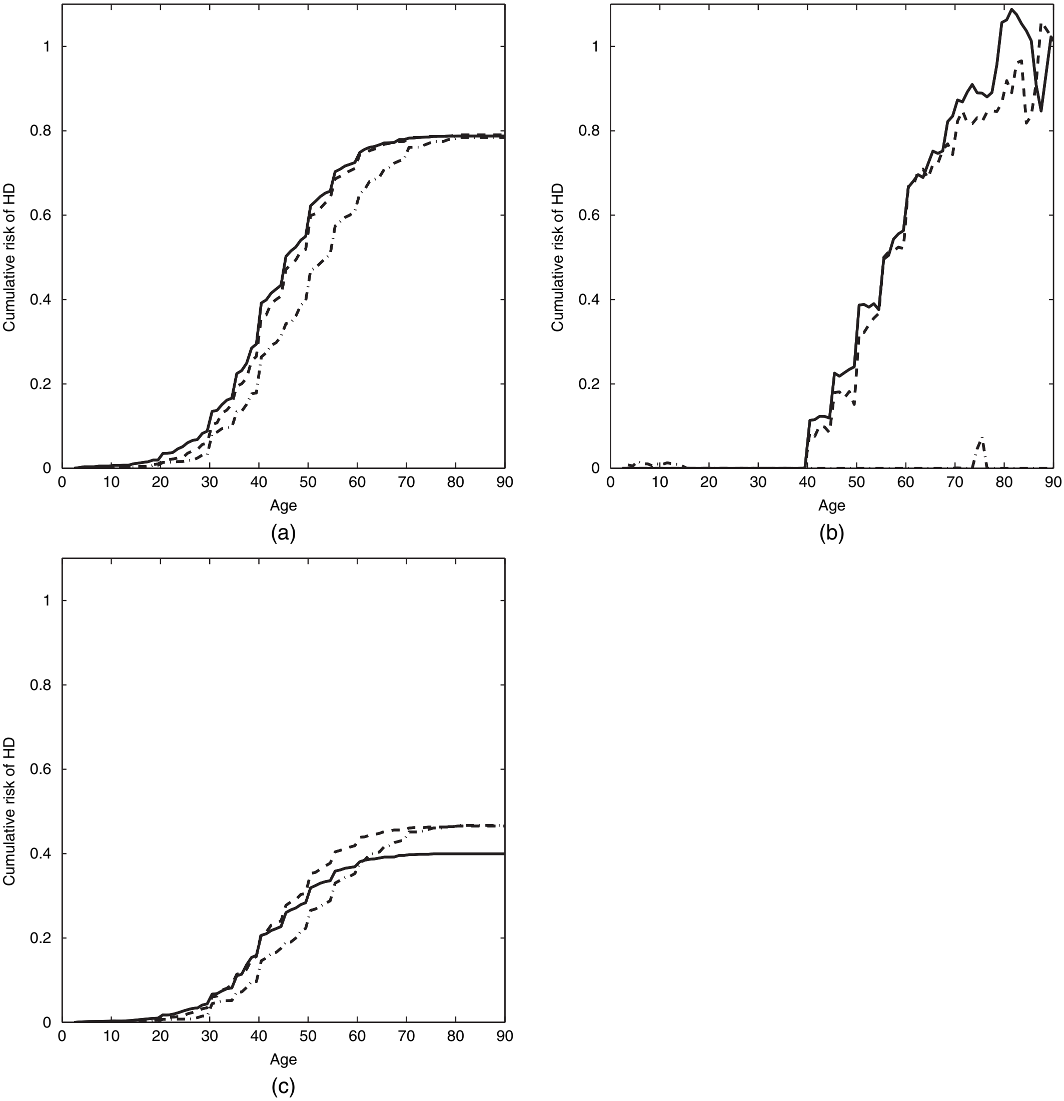


Fig. 7. COHORT study: cumulative risk of onset of HD based on (a) WLS, (b) IND and (c) NPMLE1 stratified by the status of reported affected father (———), affected mother (------) or neither of the parents (·-·-·-) at the time of family history interview

with a p-value less than 0.001 calculated on the basis of 1000 permuted samples. We further combined the male and female individuals and performed a similar analysis of the risk of HD onset based solely on parental status. The corresponding cumulative risks are given in Fig. 7.

The estimated cumulative risk curve can also be used as measures of the time-dependent positive or negative predictive values (see, for example, Heagerty and Zheng (2005)) of the HD mutation test. To see this, note that the first component of $\mathbf{F}(t)$ is the cumulative risk for carriers, i.e. $F_1(t) = \text{pr}(T \leqslant t | \text{CAG} \geqslant c)$ with $c = 36$, since here CAG $\geqslant 36$ defines a positive mutation test and CAG $< c$ defines a negative mutation test. Thus the quantity $F_1(t)$ is also referred to as the time-dependent positive predictive value in the diagnostic testing literature

(Heagerty and Zheng, 2005) and is used to summarize the performance of a test for time-to-event outcomes collected in non-standard designs (Liu *et al.*, 2012). These measures provide a numerical summary of cumulative risk by certain age associated with a positive mutation test. In addition, the estimated curves can also be used to predict a subject's risk of HD given his or her mutation test results and other demographic information. For example, from Fig. 6, a female subject who has a positive HD mutation and reports an affected father has a chance of about 65% of developing HD by age 50 years. Lastly, these measures are useful to predict the conditional probabilities of developing HD in the next few years given the current age of a subject. For example, one can estimate the conditional probability of developing HD in the next 5 years for a mutation carrier free of disease at age 50 years, i.e. $\mathrm{pr}(T<55|T\geqslant 50, \mathrm{CAG}\geqslant 36)$.

6. Discussion

We have provided a general WLS family to estimate the distribution functions of several populations when the observations are from a mixture of these populations and are subject to right censoring. Existing consistent non-parametric estimators in these problems are NPMLE1 and IND, and they are shown to be non-ideal members of this family. We have further proposed a practically optimal member of the WLS family. It is easy to see that, when there is no censoring, the proposed WLS estimator is identical to IND. However, when there is censoring we demonstrate that the estimator proposed has superior performance and computational stability compared with both IND and NPMLE1. In addition, the estimator proposed is extremely easy to implement and its asymptotic properties are also easily established. We illustrate the methods and their applications to perform risk prediction through an application to the COHORT study. Here we estimate the cumulative distribution function of onset of HD in HD mutation carriers (CAG lengths 36 or longer) instead of in each CAG repeat length group. The estimates are useful in genetic counselling settings when a subject knows only the CAG expansion status (expanded *versus* not expanded) in a family member but does not necessarily know the actual CAG repeat length. These distribution functions quantify the effect of having a family member with a positive HD mutation test on one's own risk of developing HD.

An alternative method of treating censoring is imputation, related to the self-consistent estimator (Efron, 1967). In our context, we show in Appendix A.1 that the imputation estimator is also a member of the WLS family. In addition to the WLS family, yet another estimator is a maximum likelihood estimator (MLE) through imputation (see Appendix A.2 for details). Like the imputation method, the MLE has not been reported in the literature before; hence it provides another new estimator. However, when examining it in the simulations, we find no gain in efficiency over the proposed WLS estimator. In addition, since the MLE cannot be solved explicitly, its computation requires an iteration procedure such as the Newton–Raphson method. In some occasions, the iterative computation may cause numerical instability, and the algorithm may even fail to converge. In light of these numerical performances, we suggest that the proposed WLS method in equation (2) is used.

All the estimators that we have studied are developed under the situation that the different number of mixing probability vectors, m, is fixed. When m increases with the sample size n, a completely different treatment is required and valid estimators have been developed in Ma *et al.* (2011). It is also interesting to note that all the consistent estimators in the literature, including those which we have newly developed, carry out the analysis within each of the m mixing groups and then recover the estimate on $\mathbf{F}$. The only exceptions to this approach are NPMLE2 and the MLE. NPMLE2 turns out to be not valid, whereas the practical performance of the MLE is not ideal as we discussed before. Although we have found that different choices of weights and group-specific estimators lead to differences in efficiency, it may be interesting to investigate further whether there can be more estimators that directly perform the estimation on $\mathbf{F}$ without performing the individual analysis within each mixing group.

Lastly, we have constructed estimators of cumulative distribution functions from the proband–relative pairs which are similar to that of Chatterjee *et al.* (2006). Using the full pedigree information may increase efficiency in computing the joint probability of the mutation status of all relatives in the family given the proband's genotypes. Such a joint approach is worth investigating in a future work.

Acknowledgements

This research is supported by US National Institutes of Health grant NS073671-01 and National Science Foundation grants DMS-1206693 and DMS-1000354. Samples and data from the COHORT study, which receives support from HP Therapeutics, Inc., were used in this study. The authors thank the Huntington Study Group COHORT investigators and co-ordinators who collected data and/or samples that were used in this study, as well as participants and their families who made this work possible.

Appendix A

A.1. Illustration on imputation estimators as a member of the weighted least squares family

Suppose that, with full data, we have a consistent estimating equation

$$0=\sum_{i=1}^{n}\phi\{I(S_i\leqslant t),\mathbf{q}_i,\mathbf{F}(t)\}. \tag{3}$$

Under censoring, if S_i is observed, then we can use the ith observation as it is in equation (3). If S_i is censored by C_i, then two situations can occur. If $C_i>t$, then it is certain that $S_i>t$ as well. Hence we can safely replace $I(S_i\leqslant t)$ by 0 in the ith observation in equation (3). If $C_i\leqslant t$, then S_i can be in $(C_i,t]$ or in (t,∞). Given that S_i is censored, the probability of $S_i\in(C_i,t]$ is $\{\mathbf{q}_i^{\mathrm{T}}\mathbf{F}(t)-\mathbf{q}_i^{\mathrm{T}}\mathbf{F}(C_i)\}/\{1-\mathbf{q}_i^{\mathrm{T}}\mathbf{F}(C_i)\}$, whereas the probability of $S_i\in(t,\infty)$ is $\{1-\mathbf{q}_i^{\mathrm{T}}\mathbf{F}(t)\}/\{1-\mathbf{q}_i^{\mathrm{T}}\mathbf{F}(C_i)\}$. Thus we can replace the ith term in equation (3) by the two terms

$$\frac{\mathbf{q}_i^{\mathrm{T}}\mathbf{F}(t)-\mathbf{q}_i^{\mathrm{T}}\mathbf{F}(C_i)}{1-\mathbf{q}_i^{\mathrm{T}}\mathbf{F}(C_i)}\phi\{1,\mathbf{q}_i,\mathbf{F}(t)\}+\frac{1-\mathbf{q}_i^{\mathrm{T}}\mathbf{F}(t)}{1-\mathbf{q}_i^{\mathrm{T}}\mathbf{F}(C_i)}\phi\{0,\mathbf{q}_i,\mathbf{F}(t)\}.$$

Of course, $\mathbf{q}_i^{\mathrm{T}}\mathbf{F}(\cdot)$ is unknown, but it is $H_j(\cdot)$ for $\mathbf{q}_i=\mathbf{u}_j$ and can be estimated by using any of the previously mentioned estimators. In summary, the final estimating equation is

$$\begin{aligned}0=&\sum_{i=1}^{n}\delta_i\phi\{I(S_i\leqslant t),\mathbf{q}_i,\mathbf{F}(t)\}+\sum_{i=1}^{n}(1-\delta_i)I(C_i>t)\phi\{0,\mathbf{q}_i,\mathbf{F}(t)\}\\&+\sum_{i=1}^{n}(1-\delta_i)I(C_i\leqslant t)\left[\frac{\widetilde{\mathbf{q}_i^{\mathrm{T}}\mathbf{F}}(t)-\widetilde{\mathbf{q}_i^{\mathrm{T}}\mathbf{F}}(C_i)}{1-\widetilde{\mathbf{q}_i^{\mathrm{T}}\mathbf{F}}(C_i)}\phi\{1,\mathbf{q}_i,\mathbf{F}(t)\}+\frac{1-\widetilde{\mathbf{q}_i^{\mathrm{T}}\mathbf{F}}(t)}{1-\widetilde{\mathbf{q}_i^{\mathrm{T}}\mathbf{F}}(C_i)}\phi\{0,\mathbf{q}_i,\mathbf{F}(t)\}\right],\end{aligned}$$

where we write $\tilde{H}_j(C_i)=\widetilde{\mathbf{q}_i^{\mathrm{T}}\mathbf{F}}(C_i)$ and $\tilde{H}_j(t)=\widetilde{\mathbf{q}_i^{\mathrm{T}}\mathbf{F}}(t)$ if $\mathbf{q}_i=\mathbf{u}_j$.

In fact, the only known class of consistent estimating equations of the form (3) is $\phi\{I(S_i\leqslant t),\mathbf{q}_i,\mathbf{F}(t)\}=\omega_i\mathbf{q}_i\,I(S_i\leqslant t)-\omega_i\mathbf{q}_i\mathbf{q}_i^{\mathrm{T}}\mathbf{F}(t)$ (Ma and Wang, 2012). This yields the estimating equation

$$\sum_{i=1}^{n}\left\{\delta_i\omega_i\mathbf{q}_i\,I(S_i\leqslant t)+(1-\delta_i)I(C_i\leqslant t)\frac{\widetilde{\mathbf{q}_i^{\mathrm{T}}\mathbf{F}}(t)-\widetilde{\mathbf{q}_i^{\mathrm{T}}\mathbf{F}}(C_i)}{1-\widetilde{\mathbf{q}_i^{\mathrm{T}}\mathbf{F}}(C_i)}\omega_i\mathbf{q}_i-\omega_i\mathbf{q}_i\mathbf{q}_i^{\mathrm{T}}\mathbf{F}(t)\right\}=0,$$

which can be explicitly solved to obtain

$$\hat{\mathbf{F}}(t)=\left(\sum_{i=1}^{n}\omega_i\mathbf{q}_i\mathbf{q}_i^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{n}\left\{\delta_i\omega_i\mathbf{q}_i\,I(S_i\leqslant t)+(1-\delta_i)\,I(C_i\leqslant t)\frac{\widetilde{\mathbf{q}_i^{\mathrm{T}}\mathbf{F}}(t)-\widetilde{\mathbf{q}_i^{\mathrm{T}}\mathbf{F}}(C_i)}{1-\widetilde{\mathbf{q}_i^{\mathrm{T}}\mathbf{F}}(C_i)}\omega_i\mathbf{q}_i\right\}$$

$$=\left[\sum_{j=1}^{m}\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\left\{\sum_{i=1}^{n}\omega_i I(\mathbf{q}_i=\mathbf{u}_j)\right\}\right]^{-1}\sum_{j=1}^{m}u_j\sum_{i=1}^{n}I(\mathbf{q}_i=\mathbf{u}_j)\omega_i\left\{\delta_i\,I(S_i\leqslant t)+(1-\delta_i)\,I(C_i\leqslant t)\frac{\tilde{H}_j(t)-\tilde{H}_j(C_i)}{1-\tilde{H}_j(C_i)}\right\}.$$

Denote

$$\hat{H}_j(t)=\sum_{i=1}^{n}\frac{\omega_i\,I(\mathbf{q}_i=\mathbf{u}_j)}{\sum_{i=1}^{n}\omega_i\,I(\mathbf{q}_i=\mathbf{u}_j)}\left\{\delta_i\,I(S_i\leqslant t)+(1-\delta_i)\,I(C_i\leqslant t)\frac{\tilde{H}_j(t)-\tilde{H}_j(C_i)}{1-\tilde{H}_j(C_i)}\right\}$$

$$=\sum_{i=1}^{n}\frac{\omega_i\,I(\mathbf{q}_i=\mathbf{u}_j)}{\sum_{i=1}^{n}\omega_i\,I(\mathbf{q}_i=\mathbf{u}_j)}\left\{\delta_i\,I(S_i\leqslant t)+(1-\delta_i)\,I(C_i\leqslant t)\frac{H_j(t)-H_j(C_i)}{1-H_j(C_i)}\right\}+R.$$

Then

$$R=\sum_{i=1}^{n}\frac{\omega_i\,I(\mathbf{q}_i=\mathbf{u}_j)}{\sum_{i=1}^{n}\omega_i\,I(\mathbf{q}_i=\mathbf{u}_j)}(1-\delta_i)\,I(C_i\leqslant t)\left\{\frac{\tilde{H}_j(t)-\tilde{H}_j(C_i)}{1-\tilde{H}_j(C_i)}-\frac{H_j(t)-H_j(C_i)}{1-H_j(C_i)}\right\}$$

has the property that $n^{1/2}R$ has a normal distribution with mean 0 when $n\rightarrow\infty$ as long as the $\tilde{H}_j(t)$s are consistent estimates of $H_j(t)$ and are asymptotically normal. Simple calculation shows that, in the jth group,

$$E\left\{\delta_i\,I(S_i\leqslant t)+(1-\delta_i)\,I(C_i\leqslant t)\frac{H_j(t)-H_j(C_i)}{1-H_j(C_i)}\right\}=H_j(t).$$

Hence, $\hat{H}_j(t)$ is a root-n-consistent estimator of $H_j(t)$, and the imputation estimator has the equivalent form of

$$\hat{\mathbf{F}}(t)=\left[\sum_{j=1}^{m}\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\left\{\sum_{i=1}^{n}\omega_i\,I(\mathbf{q}_i=\mathbf{u}_j)\right\}\right]^{-1}\sum_{j=1}^{m}\mathbf{u}_j\left\{\sum_{i=1}^{n}\omega_i\,I(\mathbf{q}_i=\mathbf{u}_j)\right\}\hat{H}_j(t).$$

Viewing $\Sigma_{i=1}^{n}\,\omega_i\,I(\mathbf{q}_i=\mathbf{u}_j)$ as w_j, the imputation estimator is within the WLS family (1).

A.2. Maximum likelihood estimator

When no censoring is present, treating $\mathbf{F}(t)$ as a parameter, its log-likelihood function is

$$\sum_{i=1}^{n}I(S_i\leqslant t)\log\{\mathbf{q}_i^{\mathrm{T}}\mathbf{F}(t)\}+\sum_{i=1}^{n}I(S_i>t)\log\{1-\mathbf{q}_i^{\mathrm{T}}\,\mathbf{F}(t)\}.$$

Maximizing this function with respect to $\mathbf{F}(t)$ will yield an estimating equation

$$\sum_{i=1}^{n}\phi\{I(s_i\leqslant t),\mathbf{q}_i,\mathbf{F}(t)\}=\sum_{i=1}^{n}\frac{I(S_i\leqslant t)-\mathbf{q}_i^{\mathrm{T}}\,\mathbf{F}(t)}{\mathbf{q}_i^{\mathrm{T}}\mathbf{F}(t)\{1-\mathbf{q}_i^{\mathrm{T}}\mathbf{F}(t)\}}\mathbf{q}_i=0.$$

We can then use the same imputation procedure as in Appendix A.1 to obtain a new MLE for $\mathbf{F}(t)$. This estimator is not within the WLS family.

References

Breslow, N. and Crowley, J. (1974) A large sample study of the life table and product limit estimates under random censorship. *Ann. Statist.*, **2**, 437–453.

Chatterjee, N., Kalaylioglu, Z., Shih, J. and Gail, M. (2006) Case-control and case-only designs with genotype and family history data: estimating relative risk, residual familial aggregation, and cumulative risk. *Biometrics*, **62**, 36–48.

Chatterjee, N. and Wacholder, S. (2001) A marginal likelihood approach for estimating penetrance from kin-cohort designs. *Biometrics*, **57**, 245–252.

Churchill, G. A. and Doerge, R. W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963–971.

Dorsey, E. R. and the Huntington Study Group COHORT Investigators (2012) Characterization of a large group of individuals with Huntington disease and their relatives enrolled in the COHORT study. *PLOS ONE*, **7**, article e29522.

Efron, B. (1967) The two-sample problem with censored data. In *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, vol. IV (eds L. Le Cam and J. Neyman), pp. 831–853. New York: Prentice Hall.

Ferreira, M. E., Satagopan, J., Yandell, B. S., Williams, P. H. and Osborn, T. C. (1995) Mapping loci controlling vernalization requirement and flower time in Brassica napus. *Theor. Appl. Genet.*, **90**, 727–732.

Fine, J. P., Zou, F. and Yandell, B. S. (2004) Nonparametric estimation of the effects of quantitative trait loci. *Biometrics*, **5**, 501–513.

Fleming, T., O'Fallon, J. and O'Brien, P. (1980) Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data. *Biometrics*, **36**, 607–625.

Heagerty, P. and Zheng, Y. (2005) Survival model predictive accuracy and ROC curves. *Biometrics*, **61**, 92–105.

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

Kaplan, E. L. and Meier, P. (1958) Nonparametric Estimation from incomplete observations. *J. Am. Statist. Ass.*, **53**, 457–481.

Lander, E. S. and Botstein, D. (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 743–756.

Langbehn, D. R., Brinkman, R. R., Falush, D., Paulsen, J. S. and Hayden, M. R. (2004) A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clin. Genet.*, **65**, 267–277.

Lin, M. and Wu, R. L. (2006) A joint model for nonparametric functional mapping of longitudinal trajectories and time-to-events. *BMC Bioinform.*, **7**, article 138.

Liu, D., Cai, T. and Zheng, Y. (2012) Evaluating the predictive value of biomarkers with stratified case-cohort design. *Biometrics*, **68**, 1219–1227.

Ma, Y., Hart, J. D. and Carroll, R. J. (2011) Density estimation in several populations with uncertain population membership. *J. Am. Statist. Ass.*, **106**, 1180–1192.

Ma, Y. and Wang, Y. (2012) Efficient distribution estimation for data with unobserved sub-population identifiers. *Electron. J. Statist.*, **6**, 710–737.

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.

McLachlan, G. J. and Peel, D. (2000) *Finite Mixture Models.* New York: Wiley.

Nance, M. A., Seltzer, W., Ashizawa, T., Bennett, R., McIntosh, N., Myers, R. H., Potter, N. T., Shea, D. K. and ACMG/ASHG Statement (1998) Laboratory guidelines for Huntington disease genetic testing. *Am. J. Hum. Genet.*, **62**, 1243–1247.

Pepe, M. and Fleming, T. (1989) Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics*, **45**, 497–507.

Ranen, N. G., Stine, O. C., Abbott, M. H., Sherr, M., Codori, A. M., Franz, M. L., Chao, N. I., Chung, A. S., Pleasant, N., Callahan, C., Kasch, L., Ghaffari, M., Chase, G., Kazazian, H., Brandt, J., Folstein, S. and Ross, C. (1995) Anticipation and instability of IT-15 (CAG)n repeats in parent-offspring pairs with Huntington disease. *Am. J. Hum. Genet.*, **57**, 593–602.

Rubinsztein, D. C., Leggo, J., Coles, R., Almqvist, E., Biancalana, V., Cassiman, J. J., Chotai, K., Connarty, M., Crauford, D., Curtis, A., Curtis, D., Davidson, M. J., Differ, A. M., Dode, C., Dodge, A., Frontali, M., Ranen, N. G., Stine, O. C., Sherr, M., Abbott, M. H., Franz, M. L., Graham, C. A., Harper, P. S., Hedreen, J. C. and Hayden, M. R. (1996) Phenotypic characterization of individuals with 30-40 CAG repeats in the Huntington disease (HD) gene reveals HD cases with 36 repeats and apparently normal elderly individuals with 36-39 repeats. *Am. J. Hum. Genet.*, **59**, 16–22.

Wacholder, S., Hartge, P., Struewing, J., Pee, D., McAdams, M., Brody, L. and Tucker, M. (1998) The kin-cohort study for estimating penetrance. *Am. J. Epidem.*, **148**, 623–630.

Wang, Y., Clark, L. N., Louis, E. D., Mejia-Santana, H., Harris, J., Cote, L. J., Waters, C., Andrews, D., Ford, B., Frucht, S., Fahn, S., Ottman, R., Rabinowitz, D. and Marder, K. (2008) Risk of Parkinson's 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.

Wang, Y., Garcia, T. and Ma, Y. (2012) Nonparametric estimation for censored mixture data with application to the Cooperative Huntington's Observational Research Trial. *J. Am. Statist. Ass.*, **107**, 1324–1338.

Wellner, J. A. (1982) Asymptotic optimality of the product limit estimator. *Ann. Statist.*, **10**, 595–602.

Wexler, N. S., Lorimer, J., Porter, J., Gomez, F., Moskowitz, C., Shackell, E., Marder, K., Penchaszadeh, G., Roberts, S. A., Gayán, J., Brocklebank, D., Cherny, S. S., Cardon, L. R., Gray, J., Dlouhy, S. R., Wiktorsi, S., Hodes, M. E., Conneally, P. M., Penney, J. B., Gusella, J., Cha, J. H., Irizarry, M., Rosas, D., Hersch, S., Hollingsworth, Z., MacDonald, M., Young, A. B., Andresen, J. M., Housman, D. E., De Young, M. M., Bonilla, E., Stillings, T., Negrette, A., Snodgrass, S. R., Martinez-Jaurrieta, M. D., Ramos-Arroyo, M. A., Bickham, J., Ramos, J. S., Marshall, F., Shoulson, I., Rey, G. J., Feigin, A., Arnheim, N., Acevedo-Cruz, A., Accosta, L., Alvir, J., Fischbeck, K., Thompson, L. M., Young, A., Dure, L., O'Brien, C. J., Paulsen, J., Brickman, A., Krch, D., Peery, S., Hogarth, P., Higgins, Jr, D. S., Landwehrmeyer, B. and US–Venezuela Collaborative Research Project (2004) Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. *Proc. Natn. Acad. Sci. USA*, **101**, 3498–3503.

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

Ying, Z., Jung, S. H. and Wei, L. J. (1995) Survival Analysis with median regression models. *J. Am. Statist. Ass.*, **90**, 178–184.