Published in final edited form as:

Biometrika. 2015 September 1; 102(3): 515–532. doi:10.1093/biomet/asv030.

Efficient Estimation of Nonparametric Genetic Risk Function with Censored Data

YUANJIA WANG,
Department of Biostatistics, Mailman School of Public Health, 722 W168th Street, New York 10032, U.S.A. yw2016@columbia.edu

BAOSHENG LIANG,
School of Mathematical Sciences, Beijing Normal University, Beijing 100875, China. liangbs@mail.bnu.edu.cn

XINGWEI TONG,
School of Mathematical Sciences, Beijing Normal University, Beijing 100875, China. xweitong@bnu.edu.cn

KAREN MARDER,
Department of Neurology and Psychiatry, College of Physicians and Surgeons, Columbia University, New York 10032, U.S.A. ksm1@columbia.edu

SUSAN BRESSMAN,
The Alan and Barbara Mirken Department of Neurology, Beth Israel Medical Center, New York, 10003, U.S.A. sbressma@chpnet.org

AVI ORR-URTREGER,
Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel. aviorr@tasmc.health.gov.il

NIR GILADI, and
Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel. nirg@tasmc.health.gov.il

DONGLIN ZENG
Department of Biostatistics, CB # 7420, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7420, U.S.A. dzeng@bios.unc.edu

Summary

With an increasing number of causal genes discovered for complex human disorders, it is crucial to assess the genetic risk of disease onset for individuals who are carriers of these causal mutations and compare the distribution of age-at-onset with that in non-carriers. In many genetic epidemiological studies aiming at estimating causal gene effect on disease, the age-at-onset of disease is subject to censoring. In addition, some individuals' mutation carrier or non-carrier status can be unknown due to the high cost of in-person ascertainment to collect DNA samples or death in older individuals. Instead, the probability of these individuals' mutation status can be obtained

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes a proof for model identifiability, additional tables for Simulations 1 and 2, and results from two additional simulation studies.

from various sources. When mutation status is missing, the available data take the form of censored mixture data. Recently, various methods have been proposed for risk estimation from such data, but none is efficient for estimating a nonparametric distribution. We propose a fully efficient sieve maximum likelihood estimation method, in which we estimate the logarithm of the hazard ratio between genetic mutation groups using B-splines, while applying nonparametric maximum likelihood estimation for the reference baseline hazard function. Our estimator can be calculated via an expectation-maximization algorithm which is much faster than existing methods. We show that our estimator is consistent and semiparametrically efficient and establish its asymptotic distribution. Simulation studies demonstrate superior performance of the proposed method, which is applied to the estimation of the distribution of the age-at-onset of Parkinson's disease for carriers of mutations in the leucine-rich repeat kinase 2 gene.

Keywords

Empirical process; Mixture distribution; Parkinson's disease; Semiparametric efficiency; Sieve maximum likelihood estimation

1. Introduction

Identification of causal genes for many genetic disorders has made personalized risk assessment and prediction of disease onset a real possibility. However, although interest lies in estimating the cumulative risk distributions of disease onset for individuals who are carriers of deleterious mutations or for those with a certain haplotype, investigators may encounter missing genotypes or phase information of the haplotypes in a large proportion of individuals. For instance, genotypes in family members may be missing due to the high cost of collecting blood samples from relatives, death of a relative (Wacholder et al., 1998; Marder et al., 2003; Zhang et al., 2010; Wang et al., 2012; Qin et al., 2014), or limitations in the technology to separate two homologous chromosomes in genotyping. Furthermore, disease onset information is subject to censoring due to lost to follow-up or death.

In the presence of missing genotype information, the statistical framework for estimating disease risk distribution associated with genetic mutations is essentially the analysis of censored mixture data. There is a large body of literature on inference for mixture models. See for example, Titterington et al. (1985) and Mclachlan & Basford (1988) for parametric models, and Hall & Zhou (2003) for nonparametric models. Most of these papers address non-censored outcomes. For many genetic epidemiological studies of disease risk distributions, two features distinguish them from other censored mixture models. First, each subgroup in the mixture model is biologically meaningful and corresponds to mutation carriers or non-carriers; second, the mixing probability is usually known to the investigators or can be inferred from family pedigrees and other external sources. For example, in a case-control genetic study with valid family history information on relatives (Marder et al., 2003), the probability of a relative having a certain genotype is obtained through the relationship between relatives and probands under Mendelian assumptions (Wacholder et al., 1998; Zhang et al., 2010; Wang et al., 2012; Qin et al., 2014). In haplotype studies, the probability of a certain haplotype can be inferred from unphased genotypes under Hardy–

Weinberg equilibrium (Zeng et al., 2006), from external sources such as the HapMap project, or from sequencing data (Yang et al., 2013).

One application of this paper is to a recent study on age-specific risk of Parkinson's disease associated with mutations in the leucine-rich repeat kinase 2 gene (Paisán-Ruíz et al., 2004; Healy et al., 2008). Although Parkinson's disease is traditionally considered a non-genetic disorder, recent studies have identified genetic risk factors for Parkinson's disease especially in more genetically homogeneous sub-populations such as Ashkenazi Jews (Trinh, 2013). The goal of the current study is to estimate age-specific risk of Parkinson's disease in Ashkenazi Jews for the leucine-rich repeat kinase 2 gene mutation carriers and compare it to non-carriers. Since the leucine-rich repeat kinase 2 mutations have low prevalence, it is not efficient to randomly sample individuals from the Ashkenazi population. Instead, the study used the kin-cohort design (Wacholder et al., 1998) which was initially implemented to study genetic risks of breast cancer. In our study, an initial sample of individuals with Parkinson's disease referred as probands were sequenced for the leucine-rich repeat kinase 2 mutations and provided age-at-onset information for their first-degree relatives. Most of the relatives were not genotyped due to limited resources and therefore had unknown leucine-rich repeat kinase 2 mutation status. In addition, for older relatives who were deceased, it was not possible to collect blood samples.

Several existing works consider estimating distribution functions for such mixture data in a parametric or semiparametric framework (e.g., Diao & Lin, 2005; Zhang et al., 2010). When concerns over model misspecification arise in practice (e.g., Langbehn et al., 2004), a nonparametric model and inference through nonparametric maximum likelihood estimation are natural. However, although the Kaplan–Meier estimator is nonparametric efficient for censored data, nonparametric maximum likelihood estimators are either inconsistent or inefficient for mixture data (Wang et al., 2012). To account for censoring and the mixture nature of the problem while ensuring monotonicity of the estimated distribution function on the entire support, Qin et al. (2014) proposed methods based on a binomial likelihood and a sequence of nonparametric estimates performed by reducing censored data to current status data and implementing the expectation-maximization algorithm (Larid & Ware, 1982) with the pooled-adjacent-violators algorithm. However, this method is not guaranteed to be efficient and can be computationally intensive. Other works involving a nonparametric model based on estimating equations and weighting of Kaplan–Meier survival curves include Wacholder et al. (1998) and Fine et al. (2004).

In this work, we propose a sieve maximum likelihood estimation method to estimate disease risk associated with genetic mutations in censored mixture models. Specifically, we utilize sieve estimation based on B-splines to estimate the log-hazard ratios between the carriers and non-carriers, while the nonparametric maximum likelihood estimator is used to estimate the reference baseline hazard function. The derived estimators for the disease risk distributions are guaranteed to be asymptotically efficient. Furthermore, the calculation of the sieve maximum likelihood estimators can be easily implemented via an expectation-maximization algorithm which converges much faster than existing algorithms, due to closed form solutions in the M-step. We tackle the theoretical challenge when one functional parameter is estimated using a nonparametric maximum likelihood estimator while the other

parameter is estimated using a sieve estimator. We demonstrate substantial efficiency gains of the proposed method by simulation. Finally, we apply our method to estimate the age-at-onset of Parkinson's disease for individuals with deleterious leucine-rich repeat kinase 2 mutations (Goldwurm et al., 2011).

2. Method and Inference Procedure

2·1. Data and likelihood function

Let T_i be the age-at-onset of a disease which is subject to random censoring. Let B_i denote the potentially missing mutation status, with 1 indicating the carrier group where each individual has at least one copy of the mutation, and 2 indicating the non-carrier group. As in the Parkinson's disease study described in Section 1, the probability of being a carrier takes a finite number of values. For example, a child of a heterozygote carrier parent has a probability of 0·5 of carrying this mutation under the Mendelian assumption, so if the mutation prevalence in the general population is denoted as f, we have $\text{pr}(B = 1) = 0.5(1 + f)$ for this child. For individuals with observed carrier status, $\text{pr}(B = 1)$ equals 1 for carriers and 0 for non-carriers. We denote the finite set of values for the probability $\text{pr}(B = 1)$ by $p_1, \ldots, p_m$. Our goal is to estimate the risk distribution of the age-at-onset in the mutation group and no-mutation group, that is, $F_1(t) = \text{pr}(T \le t \mid B = 1)$ and $F_2(t) = \text{pr}(T \le t \mid B = 2)$, respectively.

Due to right censoring, the observations from n individuals consist of $\{Y_i = T_i \wedge C_i, \Delta_i = I(T_i \le C_i), \text{pr}(B_i = 1)\}$, $i = 1, \ldots, n$, where C_i denotes the censoring time assumed to be independent of T_i. We introduce an indicator variable G_i to denote m distinct mixing probabilities, so $G_i = g$ indicates $\text{pr}(B_i = 1) = p_g$ ($g = 1, \ldots, m$). After grouping individuals with the same p_g value together, the likelihood function can be written as

$$\prod_{i=1}^{n}\prod_{g=1}^{m}\left[\{p_g f_1(Y_i) + (1-p_g) f_2(Y_i)\}^{\Delta_i}\{1 - p_g F_1(Y_i) - (1-p_g) F_2(Y_i)\}^{1-\Delta_i}\right]^{I(G_i=g)},$$

where f_k is the density function corresponding to F_k ($k = 1, 2$). Our interest is to estimate F_k.

In survival analysis, it is usually more convenient to re-write the observed likelihood function using hazard functions instead of distribution functions. Let $\lambda_k(t)$ be the hazard function for T in the group with $B = k$, and let $\Lambda_k(t)$ be the corresponding cumulative hazard function. Then the likelihood function can be re-expressed as

$$\prod_{i=1}^{n}\prod_{g=1}^{m}\left(\left\{p_g\lambda_1(Y_i)\,e^{-\Lambda_1(Y_i)} + (1-p_g)\,\lambda_2(Y_i)\,e^{-\Lambda_2(Y_i)}\right\}^{\Delta_i} \times \left[1 - p_g\left\{1 - e^{-\Lambda_1(Y_i)}\right\} - (1-p_g)\left\{1 - e^{-\Lambda_2(Y_i)}\right\}\right]^{1-\Delta_i}\right)^{I(G_i=g)}. \quad (1)$$

The goal is to maximize the likelihood function (1) to estimate $\Lambda_1(t)$ and $\Lambda_2(t)$ nonparametrically and thus to obtain the age-at-onset distributions, $F_1(t)$ and $F_2(t)$. In the likelihood function (1), p_g equals 1 or 0 if an individual is observed to be a carrier or non-carrier respectively.

2·2. Sieve maximum likelihood estimation

At first glance, to estimate Λ_1 or Λ_2 in (1), one may consider a nonparametric maximum likelihood estimator (Zeng & Lin, 2010), where Λ_1 or Λ_2 are treated as step functions with jumps at the observed event times. However, due to ambiguous support points for event times when the mutation group membership is not observed, the nonparametric maximum likelihood estimator may not be consistent and its bias was observed in simulations even for very large samples (Ma & Wang, 2012; Wang et al., 2012). We therefore propose a hybrid approach involving nonparametric estimator and sieve maximum likelihood estimators that leads to consistent and semiparametric efficient estimation.

Define $\beta(t) = \log\{\lambda_1(t)/\lambda_2(t)\}$, so $\Lambda_1(t) = \int_0^t exp\{\beta(s)\}\, d\Lambda_2(s)$. The likelihood in (1) can be re-expressed as

$$\prod_{i=1}^{n}\prod_{g=1}^{m}\left(\lambda_2(Y_i)^{\Delta_i}\left[p_g e^{\beta(Y_i)} exp\left\{-\int_0^{Y_i} e^{\beta(t)} d\Lambda_2(t)\right\} + (1-p_g)\ exp\{-\Lambda_2(Y_i)\}\right]^{\Delta_i} \times \left[p_g\, exp\left\{-\int_0^{Y_i} e^{\beta(t)} d\Lambda_2(t)\right\} + (1-p_g)\ exp\{-\right.\right.$$

To maximize (2), consider using a nonparametric maximum likelihood estimator to estimate the cumulative hazard function in the baseline group, say, $\Lambda_2(t)$, but adopting a sieve approximation to estimate $\beta(t)$. Specifically, we assume that Λ_2 jumps at observed Y_i's with $\Delta_i = 1$, and we use a sieve approximation for the log-hazard ratio $\beta(t)$, letting $\beta(t) = \sum_{j=1}^{K_n} \alpha_j \phi_j(t)$, where $\varphi_1, \ldots, \varphi_{K_n}$ are basis functions for the sieve approximation. The resulting estimator maximizes a partially smoothed likelihood, where the smoothing is performed on the hazard ratio function. The use of a smoothed approximation enables one to borrow information to estimate $\Lambda_1(t)$ and thus avoid specifying its ambiguous support points as required for the nonparametric maximum likelihood estimator. In our implementation, we choose B-splines as the basis functions: We let the spline knots be $0 = t_1 = \cdots = t_l < t_{l+1} < \cdots < \tau = t_{m_n+l} = t_{m_n+l+1} = \cdots = t_{m_n+2l}$, where τ is the study duration, m_n is an integer to be chosen in a data-driven fashion, and l is the order of the B-splines. There is a total of $K_n = m_n + l$ B-spline basis functions, denoted as $\{\varphi_j : j = 1, \ldots, K_n\}$.

Using the nonparametric maximum likelihood estimator for Λ_2 and the sieve estimate for $\beta(t)$, we aim to maximize (2) or its logarithm over all the parameters including the jumps of Λ_2 and the spline coefficients $a_1, \ldots, a_{K_n}$. Direct maximization is computationally intensive and inefficient since the log-likelihood is not convex and the parameters include the potentially many jumps of Λ_2. However, using the expectation-maximization algorithm with $B_1, \ldots, B_n$, the mutation status of all individuals, treated as missing data, fast numerical convergence can be obtained due to various closed-form solutions in the M-step.

Assuming that the B_i were observed, the complete data log-likelihood function for $(Y_i, \Delta_i, B_i, G_i)$, $i = 1, \ldots, n$, is

$$\sum_{i=1}^{n} I\left(B_i{=}1\right)\left[\Delta_i \log \delta\Lambda_2\left(Y_i\right) + \Delta_i \sum_{j=1}^{K_n} \alpha_j \phi_j\left(Y_i\right) - \sum_{Y_k \le Y_i} \delta\Lambda_2\left(Y_k\right) \exp\left\{\sum_{j=1}^{K_n} \alpha_j \phi_j\left(Y_k\right)\right\}\right]$$
$$+ \sum_{i=1}^{n} I\left(B_i{=}2\right)\left\{\Delta_i \log \delta\Lambda_2\left(Y_i\right) - \Lambda_2\left(Y_i\right)\right\}$$
$$+ \sum_{i=1}^{n}\sum_{g=1}^{m} I\left(G_i{=}g, \quad B_i{=}1\right) \log p_g$$
$$+ \sum_{i=1}^{n}\sum_{g=1}^{m} I\left(G_i{=}g, \quad B_i{=}2\right) \log\left(1 - p_g\right),$$

where $\delta\Lambda_2(y)$ denotes the jump of Λ_2 at y. Therefore, the expectation-maximization algorithm consists of the following E- and M-steps. In the E-step, we evaluate the conditional probability of $B_i = 1$ given the data (G_i, Y_i, Δ_i),

$$q_i = \frac{p_{G_i} \exp\left[\Delta_i \sum_{j=1}^{K_n} \alpha_j \phi_j\left(Y_i\right) - \int_0^{Y_i} \exp\left\{\sum_{j=1}^{K_n} \alpha_j \phi_j\left(t\right)\right\} d\Lambda_2\left(t\right)\right]}{p_{G_i} \exp\left[\Delta_i \sum_{j=1}^{n} \alpha_j \phi_j\left(Y_i\right) - \int_0^{Y_i} \exp\left\{\sum_{j=1}^{K_n} \alpha_j \phi_j\left(t\right)\right\} d\Lambda_2\left(t\right)\right] + \left(1 - p_{G_i}\right) \exp\left\{-\Lambda_2\left(Y_i\right)\right\}}.$$

In the M-step, we maximize

$$\sum_{i=1}^{n} q_i \left[\Delta_i \log \delta\Lambda_2\left(Y_i\right) + \Delta_i \sum_{j=1}^{K_n} \alpha_j \phi_j\left(Y_i\right) - \sum_{Y_k \le Y_i} \delta\Lambda_2\left(Y_k\right) \exp\left\{\sum_{j=1}^{K_n} \alpha_j \phi_j\left(Y_k\right)\right\}\right] \quad (3)$$
$$+ \sum_{i=1}^{n} \left(1 - q_i\right)\left\{\Delta_i \log \delta\Lambda_2\left(Y_i\right) - \Lambda_2\left(Y_i\right)\right\}.$$

By differentiating (3) with respect to the jumps of Λ_2, we obtain a closed form solution

$$\delta\Lambda_2\left(Y_i\right) = \Delta_i \Big/ \sum_{k=1}^{n} I\left(Y_k \ge Y_i\right)\left[q_k \exp\left\{\sum_{j=1}^{K_n} \alpha_j \phi_j\left(Y_i\right)\right\} + \left(1 - q_k\right)\right]. \quad (4)$$

After inserting (4) into (3) and differentiating with respect to the αs, we obtain αs that solve the estimating equation

$$\sum_{i=1}^{n} \Delta_i \left(q_i - \frac{\sum_{k=1}^{n} I\left(Y_k \ge Y_i\right) q_k \exp\left\{\sum_{j=1}^{K_n} \alpha_j \phi_j\left(Y_i\right)\right\}}{\sum_{k=1}^{n} I\left(Y_k \ge Y_i\right)\left[q_k \exp\left\{\sum_{j=1}^{K_n} \alpha_j \phi_j\left(Y_i\right)\right\} + \left(1 - q_k\right)\right]}\right)\begin{pmatrix} \phi_1\left(Y_i\right) \\ \vdots \\ \phi_{K_n}\left(Y_i\right) \end{pmatrix} = 0, \quad (5)$$

which is easily solved using the Newton–Raphson method. With updated α's, we use (4) to update the jumps of $\Lambda_2(\cdot)$. We iterate between the E- and M-steps until convergence. We denote the final estimators by $\hat{\Lambda}_{2n}(t)$ and $\hat{\beta}_n(t) = \sum_{j=1}^{K_n} \hat{\alpha}_j \phi_j(t)$. Although we choose Λ_2 as the baseline group for the nonparametric maximum likelihood estimation and use sieve estimation to obtain a time-dependent log-hazard ratio of the first group versus the second group, the procedure can also be reversed by treating Λ_1 as the baseline group. In the

subsequent arguments, for the ease of theoretical justification, we will denote this reversely estimated $\hat{\Lambda}_{1n}(t)$ as an estimator of $\Lambda_1(t)$ instead of using $\int_0^t exp\left\{\hat{\beta}_n(s)\right\} d\hat{\Lambda}_{2n}(s)$. Empirically, we find that these two estimators of Λ_1 are almost identical.

Our theoretical results show that $\hat{\Lambda}_{kn}(t)\ (k=1,2)$ converges in distribution to a Gaussian process after normalization. To estimate its asymptotic variance, following results in Zeng & Lin (2010), one approach is to compute the observed information matrix for the jump sizes of $\hat{\Lambda}_1$ and $\hat{\Lambda}_2$ and use the inverse of this matrix to estimate the asymptotic covariance of $\hat{\Lambda}_1$ and $\hat{\Lambda}_2$. However, this approach may be numerically unstable due to inversion of a potentially high-dimensional information matrix. Alternatively, bootstrapping can be used to estimate their asymptotic covariance. Our numerical experience shows that 100 bootstrap samples are usually sufficient. In our algorithm, Λ_2 is updated using the closed form in (4) and the α's are obtained via the one-step Newton–Raphson solution to (5). Therefore, the computational burden is much less than existing methods.

Finally, using the proposed nonparametric estimators for $F_k(t) \equiv 1 - \exp\{-\Lambda_k(t)\}$, that is, $\hat{F}_{kn}(t) = 1 - exp\left\{-\hat{\Lambda}_{kn}(t)\right\}$, we can construct a variety of test statistics to compare the carrier group and the non-carrier group. One test statistic is based on the Kolmogorov–Smirnov test $\mathscr{T}_n = sup_{t\in[0,\tau]}|\hat{F}_{1n}(t) - \hat{F}_{2n}(t)|$. When $\mathscr{T}_n < c_\alpha$, we reject the null hypothesis that there is no difference between the disease risk distributions of the two groups. Here, α is the significance level and c_α is the $(1 - \alpha)$-quantile of the sampling distribution of $\mathscr{T}_n$ under permutations where the variables G_i's are permuted. Other test statistics can be $\mathscr{T}_n = \int_0^\tau \omega(t)\,|\hat{F}_{1n}(t) - \hat{F}_{2n}(t)|\,dt$, where $\omega(t)$ is a user-defined weight function that may focus on a specific time range.

2·3. Generalization to cure rate survival data

The proposed method can be generalized to analyze cure rate survival data, in which some individuals are considered to be immune to the disease of interest. To this end, we introduce a binary indicator Z to denote cure status. We assume that $\text{pr}(Z = 1 \mid B = k) = r_k$ and the disease risk function among non-cured population is $pr(T \le t|Z=0, B=k) = \tilde{F}_k(t) = 1 - exp\left\{-\tilde{\Lambda}_k(t)\right\}\ (k=1,2)$. The observed data consist of $(Y_i, \Delta_i, G_i, \Delta_i Z_i)$ $(i = 1, \ldots, n)$, where Δ_i indicates either diseased or cured. That is, for non-censored individuals, we observe some individuals, usually those who have not experienced disease after a certain age, to be cured. However, the cured status for the censored individuals is unknown. Thus, if defining $\Lambda_k(t) = -log\left[r_k + (1 - r_k)\ exp\left\{-\tilde{\Lambda}_k(t)\right\}\right]$, the observed likelihood function becomes

$$\prod_{i=1}^{n}\prod_{g=1}^{m}\left[\left\{\lambda_1(Y_i)\,e^{-\Lambda_1(Y_i)}p_g + \lambda_2(Y_i)\,e^{-\Lambda_2(Y_i)}(1-p_g)\right\}^{\Delta_i(1-Z_i)}\left\{e^{-\Lambda_1(Y_i)}p_g + e^{-\Lambda_2(Y_i)}(1-p_g)\right\}^{1-\Delta_i}\right]^{I(G_i=g)} \times \prod_{i=1}^{n}\prod_{g=1}^{m}\left[\{(1-r_1)\,p_g + (1-r_2)(1-p_g)\}^{\Delta_i(1-Z_i)}\{r_1 p_g + r_2(1-p_g)\}^{\Delta_i Z_i}\right]^{I(G_i=g)}. \quad (6)$$

We can estimate the cure rates r_k by maximizing the last part of expression (6), while we estimate $\Lambda_k(t)$ by maximizing the first part using the sieve method proposed in Section 2·2. Finally, we estimate $F_k(t)$ via $\tilde{F_k}(t) = \{1 - e^{-\Lambda_k(t)}\} / (1 - r_k)$ ($k = 1, 2$).

3. Asymptotic Results

Let λ_{k0} and Λ_{k0} be the true hazard rates and the cumulative hazard functions for group k, ($k = 1, 2$) under the setting of Sections 2.1 and 2.2. Then the true log-hazard ratio is $\beta_0(t) = \log\{\lambda_{10}(t)/\lambda_{20}(t)\}$. We need the following conditions:

Condition 1

Both $\lambda_{10}(t)$ and $\lambda_{20}(t)$ are r times continuously differentiable in $[0, \tau]$, where $r \geq 2$. In addition, there exist g_1 and g_2 such that $p_{g_1}/p_{g_2} \neq (1 - p_{g_1})/(1 - p_{g_2})$.

Condition 2

The density of C has bounded and continuous rth derivative in $[0, \tau]$, and C is independent of T conditional on G.

Condition 3

The number of interior knots m_n satisfies $m_n^{3/2}/n^{1/2} = O(1)$ and $n^{1/2}/m_n^{2r} \to 0$, as n goes to infinity.

Conditions 1 and 2 are the regularity conditions for the underlying density functions of T in both groups. The second part of Condition 1 ensures that the data contain at least two distinct kinds of p_g to ensure identifiability of the underlying distributions. In Condition 3, one particular choice for the number of the interior knots is $m_n = n^v$, where $1/(4r) < v \leq 1/3$. Under these conditions, our first theorem gives the uniform consistency of $\hat{\Lambda}_{1n}$ and $\hat{\Lambda}_{2n}$ in $[0, \tau]$.

Theorem 1—Under Conditions 1, 2 and 3 and the setting of Sections 2.1 and 2.2,

$$\sup_{t\in[0,\tau]} |\hat{\Lambda}_{1n}(t) - \Lambda_{10}(t)| + \sup_{t\in[0,\tau]} |\hat{\Lambda}_{2n}(t) - \Lambda_{20}(t)| = o_p(1), \quad n \to \infty.$$

To describe the asymptotic distributions of $\hat{\Lambda}_{1n}$ and $\hat{\Lambda}_{2n}$, we first introduce the sets $\mathscr{F}_{BV} = \{f(t) : f(t) \text{ has a total variance bounded by 1 in } [0, \tau]\}$ and $\mathscr{F}_\beta = \{g(t) : g(t) \text{ has its } r \text{ th derivative bounded by 1 in } [0, \tau]\}$. We then treat both $\hat{\Lambda}_{1n}$ and $\hat{\Lambda}_{2n}$ as bounded stochastic processes in $\mathscr{F}_{BV}$ by defining $\hat{\Lambda}_{kn}(f) = \int_0^\tau f(s)\, d\hat{\Lambda}_{kn}(s)$ $(k = 1, 2)$, $f \in \mathscr{F}_{BV}$. Similarly, we treat $\hat{\beta}_n(t)$ as a stochastic process on $\mathscr{F}_\beta$ as $\hat{\beta}_n(g) = \int_0^\tau g(s)\, \hat{\beta}_n(s)\, ds$, $g \in \mathscr{F}_\beta$. The following theorem shows the weak convergence of these stochastic processes.

Theorem 2—*Consider* $\left\{\hat{\Lambda}_{1n}(t) - \Lambda_{10}(t), \hat{\Lambda}_{2n}(t) - \Lambda_{20}(t)\right\}$ *as a stochastic process in* $l^\infty(\mathscr{F}_{BV} \times \mathscr{F}_{BV})$. Then under Conditions 1, 2 and 3 and the setting of Sections 2.1 and 2.2,

$n^{1/2}\left\{\hat{\Lambda}_{1n}(t)-\Lambda_{10}(t),\hat{\Lambda}_{2n}(t)-\Lambda_{20}(t)\right\}$ converges in distribution to a mean-zero Gaussian process in $l^{\infty}(\mathscr{F}_{BV}\times\mathscr{F}_{BV})$, *as* $n\to\infty$. *Furthermore,* $\hat{\Lambda}_{1n}$ *and* $\hat{\Lambda}_{2n}$ are semiparametrically efficient in terms of the definition in Bickel et al. (1993). In addition, as a stochastic process in $l^{\infty}(\mathscr{F}_{\beta})$, $n^{1/2}\left(\hat{\beta}_n-\beta_0\right)$ converges in distribution to a mean-zero Gaussian process, as n $\to\infty$.

Remark 1

Theorem 2 establishes that $n^{1/2}\left\{\hat{\Lambda}_{1n}(t)-\Lambda_{10}(t)\right\}$ and $n^{1/2}\left(\hat{\Lambda}_{2n}(t)-\Lambda_{20}(t)\right\}$ converge distribution to some Gaussian process in $l^{\infty}([0,\ \tau])$. By the delta method, this also holds for the corresponding distribution function estimators, $\hat{F}_{1n}(t)=1-exp\left\{-\hat{\Lambda}_{1n}(t)\right\}$ and $\hat{F}_{2n}(t)=1-exp\left\{-\hat{\Lambda}_{2n}(t)\right\}$. Thus the sieve nonparametric maximum likelihood estimators $\hat{F_{1n}}$ and $\hat{F_{2n}}$ achieve the semiparametric efficiency bound and are optimal for the censored mixture data.

Here, semiparametric efficiency is defined in the sense of Bickel et al. (1993, Chapter 6). Theorem 2 shows that $\hat{F_k}$, as a function estimator in BV[0, τ], is semiparametrically efficient, which means that any bounded linear functional of $\hat{F_k}$ achieves its efficiency bound asymptotically. The weak convergence in Theorem 2 ensures that we can construct a valid confidence band based on these estimators. The proofs of Theorems 1 and 2 are in the Appendix. The main technical challenge is to handle the mixed convergence rates of the infinite-dimensional parameter estimators, since $\hat{\Lambda}_{kn}$ has a $n^{1/2}$-convergence rate while $\hat{\beta}_n(t)$ has a slower convergence rate. In the proof of Theorem 2, with the derived rates for $\hat{\Lambda}_{kn}$ and $\hat{\beta}_n(t)$ under some suitable norms, the master *Z*-theorem in Section 3·3 of van der Vaart & Wellner (1996) is implemented to derive the asymptotic distributions of the estimators. These theorems hold for the estimators using the cure rate survival data due to the similar likelihood function in the estimation. Although the proposed method is fully efficient based on the assumption of independent T_i given the mutation status B_i, it can be easily generalized to correlated family data by maximizing

$$\prod_{i=1}^{n}\prod_{j=1}^{n_i}\prod_{g=1}^{m}\left[\{p_g f_1(Y_{ij})+(1-p_g)f_2(Y_{ij})\}^{\Delta_{ij}}\{1-p_g F_1(Y_{ij})-(1-p_g)F_2(Y_{ij})\}^{1-\Delta_{ij}}\right]^{I(G_{ij}=g)},$$

where *i* indicates the family and *j* indicates an individual in the family. In this case, the proposed inference procedure including the expectation-maximization algorithm and bootstrap over independent families is still valid, and Theorems 1 and 2 hold except that the derived estimators may not achieve the semiparametric efficiency bound due to the maximization of a marginal likelihood.

4. Simulation Studies

Extensive simulation studies were conducted to compare the small sample performances of the proposed and existing methods. Our first simulation study used the same distribution

functions as in Qin et al. (2014). Specifically, for the carriers, $F_1(t) = \{1 - \exp(-t)\}/\{1 - \exp(-10)\}$, while for the non-carriers, $F_2(t) = \{1 - \exp(-t/2{\cdot}8)\}/\{1 - \exp(-10/2{\cdot}8)\}$ for $0 \le t \le 10$. The mutation probability p_i was randomly chosen from either the set *Case I*: (1, 0·6, 0·2, 0·16) or *Case II*: (0·75, 0·6, 0·5, 0·16). The censoring time followed a uniform distribution to yield a censoring rate of 20% or 40%. In the second simulation study, we imitated the results from the Parkinson's disease study described in Section 5: we generated survival times for carriers and non-carriers using distributions similar to the estimated distributions in the actual data, F_1 = Weibull (5·0, 102), F_2 = Weibull (5·0, 125). Furthermore, the sample size was $n = 2275$ and the mutation probability p_i was taken from (0, 0·02, 0·51, 1), as in the real example. The censoring times were generated from a uniform distribution to achieve a censoring rate of 40% or 80%.

When implementing our method, we used the cubic B-spline functions to estimate $\beta(t)$. The number of knots was set at $m_n = \lfloor n^{1/3} \rfloor - 1$ and the location each interior knot was selected to evenly distributed at the quantiles of the observed failure times. Some neighboring knots were combined if the data were found to be too sparse to stably estimate the coefficient of a particular basis function. We also experimented with the number of interior knots as $m_n/2$ or $2m_n$, and the estimates for $\Lambda_1(t)$ and $\Lambda_2(t)$ varied very little. To avoid local maximization in the expectation-maximization algorithm, we used different initial estimators including the estimates from a published method such as Qin et al. (2014). Empirically, our algorithm converged to the same results. We used 500 bootstrap samples for variance estimation. Furthermore, we compared our method with the estimator in Qin et al. (2014), which sequentially censored the observed event times to construct a binomial likelihood and applied the pooled-adjacent-violators algorithm for estimation.

The simulation results from 500 replicates for the first scenario are given in Table 1. We present the average estimated values of the cumulative distribution functions F_1 and F_2 at various quartiles. Table 1 suggests that both the sieve estimator and the method of Qin et al. (2014) have small bias, the variance estimate based on bootstrap agrees adequately with the empirical variability, and the coverage probabilities are close to the nominal level. The sieve estimator is more efficient than the method of Qin et al. (2014) in all simulation settings, and the efficiency gain, which can be as large as 60%, is more evident for the upper quartiles and for the higher censoring rate. A similar advantage of the sieve estimators is seen in Table 2 for the second simulation scenario. Our method performs well even under 80% censoring. The efficiency gain is up to 15%. In the Supplementary Material, we report root integrated mean squared errors and the average of the point-wise variance for the estimators of Λ's. Our estimators for Λ's have smaller estimation errors than those of Qin et al. (2014), especially for the estimation of Λ_1.

We performed two additional simulations with crossed distributions. The results are reported in the Supplementary Material simulations 3 and 4. The findings are similar. Finally, we also conducted simulation studies to evaluate the permutation test for the Kolmogorov–Smirnov statistic comparing the two distributions. The data generation was similar to the second simulation study, except that $F_1 = F_2$ = Weibull (5·0, 102). The empirical type I error rate is 4·6% with censoring rate 40% and 5·0% with censoring rate 80%. Both are close to the nominal significance level of 5% so the proposed permutation test appears to be valid.

5. Application

Since mutations in the leucine-rich repeat kinase 2 gene were found to be a potential cause of idiopathic Parkinson's disease (Paisán-Ruíz et al., 2004), there has been great interest in estimating the cumulative risk of Parkinson's disease for the leucine-rich repeat kinase 2 mutation carriers, especially in Ashkenazi Jews, who have an increased mutation rate (Alcalay et al., 2013). Although such risk estimates are important for genetic counseling (Goldwurm et al., 2011), results on the risk for leucine-rich repeat kinase 2 carriers in the clinical literature have been inconsistent and estimates vary widely (Goldwurm et al., 2011).

To address these concerns, we aim to estimate the age-specific cumulative risk of Parkinson's disease in the leucine-rich repeat kinase 2 carriers and non-carriers. Due to the low prevalence of leucine-rich repeat kinase 2 mutations, a kin-cohort design was used (Marder et al., 2014). To avoid bias in the ascertainment of the initial samples, our analysis units are the first-degree family members excluding the initial probands (e.g., Wacholder et al., 1998; Wang et al., 2012). Our initial probands were recruited from the Michael J. Fox foundation Ashkenazi Jewish leucine-rich repeat kinase 2 consortium; the details of the sample were reported elsewhere (Alcalay et al., 2013). All probands were screened for G2019S mutations in leucine-rich repeat kinase 2 gene and common mutations in the glucocerebrosidase gene. To isolate the effect of the leucine-rich repeat kinase 2 mutations on Parkinson's disease risk, we excluded participants with other known genetic risk factors such as glucocerebrosidase mutations. A validated family history instrument (Marder et al., 2003) was applied to the probands or the first-degree relatives themselves if relatives were seen by a neurologist.

The data included information from 2275 first-degree relatives of the probands in the Ashkenazi Jewish leucine-rich repeat kinase 2 consortium. There were four groups of mutation probabilities, $p_g \in \{0, 0{\cdot}02, 0{\cdot}51, 1\}$, with frequencies 1·6%, 70·9%, 25·4% and 2·1%, respectively. There were only 3·7% of relatives with observed genotypes, that is, their corresponding p_g is either 1 or 0. The first-degree relatives including parents, siblings or children of non-carrier probands have $p_g = 0{\cdot}02$ under a 2% population prevalence of leucine-rich repeat kinase 2 in the Ashkenazi Jewish population (Orr-Urtreger et al., 2007) and the Mendelian assumption. Similarly, the first-degree relatives of heterozygote carrier probands have $p_g = 0{\cdot}51$ under the Mendelian assumption. The censoring rate was close to 95%. Due to the high censoring rate, we analyzed the data under the cure rate model (6). Individuals who did not develop Parkinson's disease by age 95 were considered immune to the disease since the largest documented age at onset is 94 years of age (Driver, 2009). In the implementation of the proposed sieve maximum likelihood approach, we used the Bayesian information criterion to choose the number of interior knots and the degree of the B-spline basis. The choices that minimizes this criterion was two interior knots and a degree of two. We used bootstrap resampling of families to construct pointwise confidence intervals to ensure valid inference.

In the practice of genetic counseling, it is more useful to provide the population cumulative risks, that is, $F_k(t)$ in model (6), regardless of the cure survival status. Thus we report the estimates of $F_k(t)$ in Table 3. This shows that the cumulative risk of Parkinson's disease by

age 80 for carriers can be as high as 27·4% with 95% confidence interval 17·6%–39·1%, while it is 10·4% with 95% confidence interval 7·8%–13·2% for non-carriers. The risk of Parkinson's disease in non-carriers is quite high compared to general non-Ashkenazi Jews, whose risk is normally 1%, indicating that they may have other risk mutations for Parkinson's disease. The estimated lifetime cumulative risk is consistent with some previous findings in Ashkenazi Jews for leucine-rich repeat kinase 2 mutation carriers (Wang et al., 2008), but it contrasts with some other studies, which estimate risk of Parkinson's disease to be 100% in leucine-rich repeat kinase 2 carriers (Lesage, 2005). Methodological issues including assigning individuals with unobserved leucine-rich repeat kinase 2 genotypes to carrier or non-carrier groups based on their Parkinson's disease status may have contributed to this large difference with those studies. Figure 1 presents the estimated cumulative Parkinson's disease distributions in the two mutation groups and their pointwise confidence intervals. The carrier group has a dramatic increase of the risk of Parkinson's disease after age 60 as compared to a slower increase in the disease risk in the non-carrier group.

To compare the distributions, we used the Kolmogorov–Smirnov test to examine the maximal difference between the two groups. We computed the *p*-value for this test based on 1,000 permutations, where for each permutation, the grouping variable G_i was perturbed. The resulting *p*-value is less than 0·001. It may be of practical interest to examine some classes of parametric models for the genetic risk functions. For example, within the class of Weibull distributions, we find the estimated distribution for the carriers is adequately approximated by a Weibull distribution with shape and scale parameters 5 and 102, while the estimated distribution for the non-carriers is close to a Weibull with shape and scale parameters 5 and 125.

The cure rates in carriers and non-carriers were estimated to be 0·3% with 95% confidence interval 0%– 19·8% and 26·6% with 95% confidence interval 17·9%–34·6%, respectively. There is a of significant difference 26·3% between the two rates with 95% confidence interval 3·6%–34·3%. In the non-cured population, the cumulative risk of Parkinson's disease for carriers by age 80 was 27·5%, that is, $F_1(t)$ as defined in Section 2·3 was 27·5% at age 80, compared to 14·2% for the non-carrier group. The low cure rate in the carrier group suggests a high risk of Parkinson's disease had a subject lived long enough. This observation is consistent with the existing clinical literature. For example, Latourelle et al. (2008) reported a high lifetime risk of Parkinson's disease, where the median risk of disease was about 70% and the upper limit of the 95% confidence interval was about 80%.

6. Discussion

One interesting theoretical is to tackle the different convergence rates of the nonparametric maximum likelihood and the sieve estimators based on B-splines. Alternatively, sieve estimation can also be applied to Λ_2, as done by Cheng & Wang (2011) for a semiparametric additive transformation model with current status data. However, one advantage of using the nonparametric maximum likelihood estimator for Λ_2 is that there is no need to determine the number of sieves. In addition, our nonparametric maximum likelihood estimator has an explicit solution in the M-step of the expectation-maximization algorithm, which leads to computational gain.

Using the re-parametrized likelihood function (2), the proposed method can be readily generalized to regression problems where other environmental covariates are included through a proportional hazards model in both groups (Diao & Lin, 2005). Lastly, to efficiently analyze family data, an alternative method using frailty models may be considered to account for within-family dependence through shared frailties.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work is supported by Michael J. Fox foundation, U.S. National Institute of Health grants, National Natural Science Foundation of China, and China Scholarship Council. We thank the editor, an associate editor and referees for helpful comments to improve the paper.

Appendix

Before proving Theorem 1 and Theorem 2, we first show that the information operator for Λ_2 and β is invertible. For $G = g$, we define

$$\begin{aligned} A_1(\Delta, G, Y) &= \Xi_1\left[(1-p_g)\,e^{-\Lambda_2(Y)} + p_g e^{\beta(Y)} exp\left\{-\int_0^Y e^{\beta(t)} d\Lambda_2(t)\right\}\right], \\ A_2(\Delta, G, Y) &= (1-p_g)\,e^{-\Lambda_2(Y)}(\Xi_1+\Xi_2), \\ A_3(\Delta, G, Y) &= p_g\left\{e^{\beta(Y)}\Xi_1+\Xi_2\right\} exp\left\{-\int_0^Y e^{\beta(t)} d\Lambda_2(t)\right\}, \\ B_1(\Delta, G, Y) &= p_g e^{\beta(Y)} exp\left\{-\int_0^Y e^{\beta(t)} d\Lambda_2(t)\right\}\Xi_1, \end{aligned}$$

where $\Xi_1 = \Delta\left(p_g e^{\beta(Y)} exp\left[-\int_0^Y exp\{\beta(t)\}\, d\Lambda_2(t)\right] + (1-p_g)\ exp\{-\Lambda_2(Y)\}\right)^{-1}$, and $\Xi_2 = (1-\Delta)\left(p_g\, exp\left[-\int_0^Y exp\{\beta(t)\}\, d\Lambda_2(t)\right] + (1-p_g)\ exp\{-\Lambda_2(Y)\}\right)^{-1}$. The log-likelihood function for a single subject is

$$l(\Lambda_2, \beta) = \sum_{g=1}^{m} I(G=g)\left(\Delta\ log\ \left[p_g \lambda_2(Y)\, e^{\beta(Y)}\ exp\left\{-\int_0^Y e^{\beta(t)} d\Lambda_2(t)\right\} + (1-p_g)\,\lambda_2(Y)\, e^{-\Lambda_2(Y)}\right] + (1-\Delta)\ log\ \left[p_g\ exp\left\{-\int_0^Y e^{\beta(t}\right.\right.$$

By differentiating $l(\Lambda_2, \beta)$ with respect to Λ_2 and β along sub-models $d\Lambda_2(1 + \varepsilon h_1)$ and $\beta + \varepsilon h_2$ respectively, we obtain the following score operators

$$\begin{aligned} & l_{\Lambda_2}(\Lambda_2, \beta)(h_1) \\ &\quad = A_1 h_1(Y) \\ &\quad - \int_0^Y h_1(t)\left[A_2 + A_3\, exp\{\beta(t)\}\right] d\Lambda_2(t), l_\beta(\Lambda_2, \beta)(h_2) \\ &\quad = B_1 h_2(Y) \\ &\quad - A_3 \int_0^Y h_2(t)\ exp\{\beta(t)\}\, d\Lambda_2(t) \end{aligned}$$

Thus, if we define $\langle f_1, f_2 \rangle = E(f_1 f_2)$, for any $L_2(P)$-integrable functions $\{w_1(\Delta, G, Y), w_2(\Delta, G, Y)\}$, we have

$$\left\langle \left\{ \begin{array}{c} l_{\Lambda_2}\left(\tilde{h}_1\right) \\ l_{\beta}\left(\tilde{h}_2\right) \end{array} \right\}, \left(\begin{array}{c} w_1 \\ w_2 \end{array} \right) \right\rangle$$
$$= \left\langle \left(\begin{array}{c} h_1 \\ h_2 \end{array} \right), \left\{ \begin{array}{c} E\left(A_1 w_1 | Y\right) \\ E\left(B_1 w_2 | Y\right) \end{array} \right\} \right\rangle$$
$$+ \int_0^Y \left\langle \left(\begin{array}{c} h_1 \\ h_2 \end{array} \right), \left[\begin{array}{c} E\left\{\left(A_2 + A_3\right) w_1 | Y = t\right\} e^{\beta(t)} \\ E\left\{A_3 w_2 | Y = t\right\} e^{\beta(t)} \end{array} \right] \right\rangle d\Lambda_2(t).$$

Thus,

$$l^*_{\Lambda_2} l_{\Lambda_2}(h_1) = E\left(A_1^2 | Y\right) h_1(Y) - \int_0^Y E\left[A_1\left\{A_2 + A_3 e^{\beta(t)}\right\} | Y = t\right] h_1(t)\, d\Lambda_2(t) - \int_0^Y E\left\{\left(A_2 + A_3\right) A_1 | Y = t\right\} h_1(t)\, d\Lambda_2(t) + \int_0^Y \int_0^t$$
$$l^*_{\beta} l_{\beta}(h_2) = E\left(B_1^2 | Y\right) h_2(Y) - \int_0^Y E\left\{B_1 A_3 e^{\beta(t)} | Y = t\right\} h_2(t)\, d\Lambda_2(t) - \int_0^Y E\left\{A_3 \sum_{g=1}^{m} I(G = g)\, B_1 | Y = t\right\} h_2(t)\, d\Lambda_2(t) + \int_0^Y \left[\int_0^$$

where $\left(l^*_{\Lambda_2}, l^*_{\beta}\right)$ is the dual operator of (l_{Λ_2}, l_{β}). Therefore, the information operator $\mathscr{I}(\Lambda_2, \beta) = (l_{\Lambda_2}, l_{\beta})^* (l_{\Lambda_2}, l_{\beta})$ can be expressed as a Fredholm operator of the first kind, which is the summation of an invertible operator and an integral operator when $\Lambda_2 = \Lambda_{20}$ and $\beta = \beta_0$. As a result, to show that $\mathscr{I}(\Lambda_{20}, \beta_0)$ is invertible, following Rudin (1973), it suffices to show that $\mathscr{I}(\Lambda_{20}, \beta_0)$ is one-to-one. That is, we need to prove that for any h_1 and h_2 if $\mathscr{I}(\Lambda_{20}, \beta_0)(h_1, h_2) = 0$, which is equivalent to $l_{\Lambda 20}(h_1) + l_{\beta_0}(h_2) = 0$, then $h_1 \equiv 0$ and $h_2 \equiv 0$. Suppose that $l_{\Lambda 20}(h_1) + l_{\beta_0}(h_2) = 0$, let $\Delta = 1$ and $G = g$ and integrate Y from 0 to any $t \in [0, \tau]$, we then obtain $\int_0^t \left[p_g \left\{h_1(s) + h_2(s)\right\} e^{\beta_0(s)} + (1 - p_g) h_2(s)\right] d\Lambda_{20}(s) = 0$. Thus, $p_g\{h_1(t) + h_2(t)\} e^{\beta_0(t)} + (1 - p_g) h_2(t) = 0$. From Condition 1, we immediately conclude that $h_1 = h_2 \equiv 0$. Therefore, $\mathscr{I}(\Lambda_{20}, \beta_0)(h_1, h_2)$ is continuously invertible.

Furthermore, we consider a different Banach space $\mathscr{H}^* = \{(h_1, h_2) : h_1 \in L_2[0, \tau], h_2 \in L_2[0, \tau]\}$. Then the above arguments still hold. Hence, the invertibility of $\mathscr{I}(\Lambda_{20}, \beta_0)$ implies $\|\mathscr{I}(\Lambda_{20}, \beta_0)\|^2_{L_2} \geq c\left(\|h_1\|^2_{L_2} + \|h_2\|^2_{L_2}\right)$ where c is a constant. Furthermore, if $\|\Lambda_2 - \Lambda_{20}\|_{\infty} + \|\beta - \beta_0\|_{\infty} < \varepsilon_0$ for a small ε_0, the continuity of $\mathscr{I}$ in this space gives

$$\|\mathscr{I}(\Lambda_2, \beta)(h_1, h_2)\|^2_{L_2} \geq \frac{c}{2}\left(\|h_1\|^2_{L_2} + \|h_2\|^2_{L_2}\right).$$

We will use this fact in the following consistency proof.

Proof of Theorem 1

We define a sieve space

$S_n = \left\{ (\Lambda_2, \beta) : \Lambda_2 \right.$ is the step function with jump at the observed events, $\beta(t) = \sum_{j=1}^{K_n} \alpha_j \phi_j(t)$, where the ϕ_j are

First, we show that there exists a local maximum of the observed data likelihood function over S_n such that the proposed estimators $\left(\hat{\Lambda}_2, \hat{\beta}\right)$ converge to the true parameters in probability under the norm $\| \cdot \|_\infty$.

By Schumaker (2007) and Condition 1 there exists a function $\hat{\beta}_0(t) = \sum_{j=1}^{K_n} \alpha_{j0} \phi_j(t)$ such that $\|\hat{\beta}_0 - \beta_0\|_\infty = O\left(m_n^{-r}\right)$. Then we consider the neighborhood of $\hat{\beta}_0$ in the following sieve space $\mathcal{N}_{\epsilon_n} = \left\{ \beta : \beta(t) = \sum_{j=1}^{K_n} \alpha_j \phi_j(t) \quad \text{with} \quad \left(\sum_{j=1}^{K_n} |\alpha_j - \alpha_{j0}|^2 \right)^{1/2} \leq \epsilon_n \right\}$ where ε_n is to be chosen later. For each $\beta \in \mathcal{N}_{\epsilon_n}$, we define $\hat{\Lambda}_{2,\beta} = argmax_{\Lambda 2} P_n l(\Lambda_2, \beta)$, where Λ_2 is a step function with jumps at the observed failure events. If we chose ε_n such that $m_n^{3/2} \epsilon_n \to 0$, then for $\beta \in \mathcal{N}_{\epsilon_n}$,

$$\|\beta - \hat{\beta}_0\|_{BV} \leq \sum_{j=1}^{K_n} |\alpha_j - \alpha_{j0}| \|\phi_j'\|_\infty \leq O(m_n) \left\{ \epsilon_n^2 (m_n + l) \right\}^{1/2} \to 0.$$

Therefore, β has bounded total variation. Define

$$\hat{\Lambda}_{20}(t) = \sum_{j=1}^{n} \int_0^t \frac{I(Y_j \geq s)}{\sum_{k=1}^{n} I(Y_k \geq s) [q_k \exp\{\beta_0(s)\} + (1 - q_k)]} dN_j(s),$$

it is easy to see that $\|\hat{\Lambda}_{20} - \Lambda_{20}\|_{BV} = O_p\left(n^{-1/2}\right)$. Therefore, $P_n l\left(\hat{\Lambda}_{2,\beta}, \beta\right) \geq P_n l\left(\hat{\Lambda}_{20}, \beta\right)$, where P_n denotes the empirical measure. Note that $P_n l\left(\hat{\Lambda}_{2,\beta}, \beta\right) - p_n l\left(\hat{\Lambda}_{20}, \beta\right)$ equals

$$n^{-1} \sum_{i=1}^{n} \sum_{g=1}^{m} I(G_i = g) \left(\Delta_i \log \left[\frac{\delta \hat{\Lambda}_{2,\beta}(Y_i)}{\delta \hat{\Lambda}_{20}(Y_i)} \frac{p_{ig} e^{\beta(Y_i)} \exp\left\{ -\int_0^{Y_i} e^{\beta(t)} d\hat{\Lambda}_{2,\beta}(t) \right\} + (1 - p_{ig}) e^{-\hat{\Lambda}_{2,\beta}(Y_i)}}{p_{ig} e^{\beta(Y_i)} \exp\left\{ -\int_0^{Y_i} e^{\beta(t)} d\hat{\Lambda}_{20}(t) \right\} + (1 + p_{ig}) e^{-\hat{\Lambda}_{20}(Y_i)}} \right] + (1 - \Delta_i) \log \left[\frac{p_{ig} \exp\left\{ -\int_0^{Y_i} \right.}{p_{ig} \exp\left\{ -\int_0^{Y_i} \right.} \right.$$

It is easy to show that $\delta \hat{\Lambda}_{20}(t) = O_p(1/n)$, so we conclude that there exist constants c_1 and c_2 independent of β such that

$$
\begin{aligned}
0 \leq & \; P_n l\left(\hat{\Lambda}_{2,\beta},\beta\right) - P_n l\left(\hat{\Lambda}_{20},\beta\right) \\
\leq & \; n^{-1}\sum_{i=1}^{n} c_1 \, log\left\{n\delta\hat{\Lambda}_{2,\beta}\left(Y_i\right)\right\}\Delta_i + c_2 \, log\left[p_i \, exp\left\{-\int_0^{\tau} e^{\beta(s)} d\hat{\Lambda}_{2,\beta}\left(s\right)\right\} + \left(1-p_i\right) e^{-\hat{\Lambda}_{2,\beta}(\tau)}\right] \\
\leq & \; n^{-1}\sum_{i=1}^{n} c_1 \, log\left\{n\delta\hat{\Lambda}_{2,\beta}\left(Y_i\right)\right\}\Delta_i + c_2 \, log\left\{p_i e^{-c\hat{\Lambda}_{2,\beta}(\tau)} + \left(1-p_i\right) e^{-c\hat{\Lambda}_{2,\beta}(\tau)}\right\} \\
\leq & \; n^{-1}\sum_{i=1}^{n} c_1 \, log\left\{n\delta\hat{\Lambda}_{2,\beta}\left(Y_i\right)\right\}\Delta_i - c_2\hat{\Lambda}_{2,\beta}\left(\tau\right) + O_p\left(1\right).
\end{aligned}
$$

Hence, $n^{-1}\sum_{i=1}^{n}\Delta_i \, log\left\{n\delta\hat{\Lambda}_{2,\beta}\left(Y_i\right)\right\} - c_1\hat{\Lambda}_{2,\beta}\left(\tau\right)$ is bounded from below in probability. since $n^{-1}\sum_{i=1}^{n}\Delta_i \, log\left\{n\delta\hat{\Lambda}_{2,\beta}\left(Y_i\right)\right\}$ is less than $log\left\{\sum_{i=1}^{n}\Delta_i\delta\hat{\Lambda}_{2,\beta}\left(Y_i\right)\right\} = log\,\hat{\Lambda}_2\left(\tau\right), \overline{lim}_{n\to\infty}\left\{sup_{\beta\in\mathcal{N}_{\epsilon_n}}\hat{\Lambda}_{2,\beta}\left(\tau\right)\right\}$ is finite with probability tending to one. As a result, $\left\{\hat{\Lambda}_{2,\beta}{:}\beta\in\mathcal{N}_{\epsilon_n}\right\}$ consists of bounded and increasing functions.

From the fact that $P_n l_{\Lambda_2}\left(\hat{\Lambda}_{2,\beta},\beta\right)\left(h_1\right)=0$, we obtain $\left(P_n-P\right) l_{\Lambda_2}\left(\hat{\Lambda}_{2,\beta},\beta\right)\left(h_1\right) = P l_{\Lambda_2}\left(\hat{\Lambda}_{2,\beta},\beta\right)\left(h_1\right)$. The left-side of this equation is $O_p(n^{-1/2})$, because l_{Λ_2} is Donsker due to the fact that both $\Lambda_{2,\beta}$ and β belong to BV[0, τ]. We apply the Taylor expansion at the true (Λ_{20}, β_0) to the right-hand side, then we have

$$
O_p\left(n^{-1/2}\right) = -\left\langle \dot{\mathscr{I}}_1\left(\Lambda_{20},\beta_0\right)\left(h_1\right), d\hat{\Lambda}_{2,\beta} - d\Lambda_{20}\right\rangle_{L_2} + o\left(\left\|\hat{\Lambda}_{2,\beta}-\Lambda_{20}\right\|_{BV}\right) + O_p\left(\left\|\beta-\beta_0\right\|_{L_2}\right),
$$

where $\mathscr{I}_1$ is the operator in $\mathscr{I}$ corresponding to Λ_2. Using the invertibility of $\mathscr{I}_1$, we have $\left\|\hat{\Lambda}_{2,\beta}-\Lambda_{20}\right\|_{BV} = A_n\left(n^{-1/2}+\left\|\beta-\beta_0\right\|_{L_2}\right)$, where $sup_{\beta\in\mathcal{N}_{\epsilon_n}}|A_n|$ is a bounded random variable.

We now consider $B_n \equiv P_n\left\{l\left(\hat{\Lambda}_{2,\beta},\beta\right) - l\left(\hat{\Lambda}_{20},\hat{\beta}_0\right)\right\}$. First, $B_n = \left(P_n-P\right)\left\{l\left(\hat{\Lambda}_{2,\beta},\beta\right) - l\left(\hat{\Lambda}_{20},\hat{\beta}_0\right)\right\} + P\left\{l\left(\hat{\Lambda}_{2,\beta},\beta\right) - l\left(\hat{\Lambda}_{20},\hat{\beta}_0\right)\right\}$. The first term on the right hand side is equal to $c_n n^{-1/2}$, where $sup_{\beta\in\mathcal{N}_{\epsilon_n}}|c_n|\to 0$. For the second term, we apply the expansion at the true values and obtain

$$
-\left\langle \ddot{\mathscr{I}}\left(\Lambda_2^*,\beta^*\right)\left(d\hat{\Lambda}_{2,\beta}/d\hat{\Lambda}_{20}-\lambda_{20},\beta-\beta_0\right), \left(d\hat{\Lambda}_{2,\beta}/d\hat{\Lambda}_{20}-\lambda_{20},\beta-\beta_0\right)\right\rangle_{L_2} + o\left(\left\|\hat{\Lambda}_{20}-\Lambda_{20}\right\|_{\infty}^2 + \left\|\hat{\beta}_0-\beta_0\right\|_{\infty}^2\right),
$$

where $\left(\Lambda_2^*,\beta^*\right)$ is between $\left(\hat{\Lambda}_{2,\beta},\beta\right)$ and (Λ_{20}, β_0). Thus, we obtain $B_n \leq c_n n^{-1/2} - c_1/2\left\|\beta-\beta_0\right\|_{L_2}^2 + b_n\left(n^{-1}+m_n^{-2r}\right)$, where $sup_{\beta\in\mathcal{N}_{\epsilon_n}}|b_n|\to 0$. Therefore, if $\beta\in\partial\mathcal{N}_{\epsilon_n}$, the result from Boor (1978) gives $\left\|\beta-\beta_0\right\|_{L_2}^2 \geq c_2\epsilon_n^2$, so that $B_n \leq \left\{|c_n|n^{-1/2} + b_n\left(n^{-1}+m_n^{-2r}\right)\right\} - c_1c_2\epsilon_n^2/2$. Hence, if we choose

$\epsilon_n^2 = 4(c_1 c_2)^{-1}\left\{|c_n| n^{-1/2} + b_n\left(n^{-1} + m_n^{-2r}\right)\right\}$, then $B_n < 0$, noting that such ε_n still satisfies $m_n^{3/2}\epsilon_n \to 0$ due to $r \geq 2$ and Condition 3. That is, there exists a local maximum $\hat{\beta}_n$ within this neighborhood. Consequently, $\|\hat{\beta}_n - \beta_0\|_{BV} \to 0$ and $\|\hat{\beta}_n - \beta_0\|_{L_2}^2 \leq \|\hat{\beta}_n - \hat{\beta}_0\|_{L_2}^2 + O\left(m_n^{-2r}\right) \leq \epsilon_n^2 + m_n^{-2r} = o_p\left(n^{-1/2}\right)$. From the result that $\|\hat{\Lambda}_{2,\beta} - \Lambda_{20}\|_{BV} = A_n\left(n^{-1/2} + \|\beta - \beta_0\|_{L_2}\right)$, the corresponding $\hat{\Lambda}_{2,\beta}$ satisfies $\|\hat{\Lambda}_{2,\hat{\beta}_n} - \Lambda_{20}\|_{BV} = O_p\left(n^{-1/2}\right) + \|\hat{\beta}_n - \beta_0\|_{L_2} = o_p\left(n^{-1/4}\right)$. It implies $sup_{t\in[0,\tau]}|\hat{\Lambda}_{2n}(t) - \Lambda_{20}(t)| + sup_{t\in[0,\tau]}|\hat{\beta}_n(t) - \beta_0(t)| = o_p(1)$. By reversing the labels, the same argument implies $sup_{t\in[0,\tau]}|\hat{\Lambda}_{1n}(t) - \Lambda_{10}(t)| = o_p(1)$. The proof of Theorem 1 is completed.

Proof of Theorem 2

For any $h_1 \in \mathrm{BV}[0, \tau]$ and any h_2 with bounded rth derivative in $[0, \tau]$, we have $P_n l_{\Lambda_2}\left(\hat{\Lambda}_{2n}, \hat{\beta}_n\right)(h_1) = 0$ and $P_n l_\beta\left(\hat{\Lambda}_{2n}, \hat{\beta}_n\right)(h_{2n}) = 0$. Here, h_{2n} is the projection of h_2 on S_n, and $\|h_{2n} - h_2\|_\infty = O\left(m_n^{-r}\right)$. This gives

$$G_n\left\{l_{\Lambda_2}\left(\hat{\Lambda}_{2n}, \hat{\beta}_n\right)(h_1) + l_\beta\left(\hat{\Lambda}_{2n}, \hat{\beta}_n\right)(h_{2n})\right\} = -n^{1/2} P\left\{l_{\Lambda_2}\left(\hat{\Lambda}_{2n}, \hat{\beta}_n\right)(h_1) + l_\beta\left(\hat{\Lambda}_{2n}, \hat{\beta}_n\right)(h_{2n})\right\}, \quad \text{(A1)}$$

where $G_n = n^{1/2}(P_n - P)$. It is straightforward to verify $\left\{l_\beta\left(\hat{\Lambda}_{2n}, \hat{\beta}_n\right)(h_1) + l_{\Lambda_2}\left(\hat{\Lambda}_{2n}, \hat{\beta}_n\right)(h_2) : \|h_1\|_{BV} \leq 1, \quad \|h_2\|_\infty \leq 1\right\}$ is a Donsker class. Thus, the left-hand side of equation (A1) is equal to $G_n\left\{l_{\Lambda_2}\left(\hat{\Lambda}_{2n}, \hat{\beta}_n\right)(h_1) + l_\beta(\Lambda_{20}, \beta_0)(h_2)\right\} + o_p(1)$ where $o_p(1)$ here and in the sequel refers to some random element that converges in probability to zero uniformly in (h_1, h_2).

By the Taylor expansion, the right-hand side of equation (A1) equals

$$\begin{aligned} &-\{1 + o_p(1)\}\left\{\int h_1^* d\left(\hat{\Lambda}_{2n} - \Lambda_{20}\right) + \int h_2^*\left(\hat{\beta}_n - \beta_0\right) dt\right\} \\ &\quad + n^{1/2} O\left(\|\hat{\Lambda}_{2n} - \Lambda_{20}\|_{BV}^2 + \|\hat{\beta}_n - \beta\|_{L_2}^2\right) \\ &\quad = -n^{1/2}\{1 + o_p(1)\}\left\{\int h_1^* d\left(\hat{\Lambda}_{2n} - \Lambda_{20}\right) + \int h_2^*\left(\hat{\beta}_n - \beta_0\right) dt\right\} + o_p(1), \end{aligned}$$

where $(h_1^*, h_2^*) = \mathscr{I}(\Lambda_{20}, \beta_0)(h_1, h_2)$. This yields that

$$G_n\left\{l_{\Lambda_2}(\Lambda_{20}, \beta_0)(h_1^{**}) + l_\beta(\Lambda_{20}, \beta_0)(h_2^{**})\right\} + o_p(1) = -n^{1/2}\left\{\int h_1 d\left(\hat{\Lambda}_{2n} - \Lambda_{20}\right) + \int h_2\left(\hat{\beta}_n - \beta_0\right) dt\right\}$$

where $(h_1^{**}, h_2^{**}) = \mathscr{I}^{-1}(\Lambda_{20}, \beta_0)(h_1, h_2)$. That is, $n^{1/2}\left\{\hat{\Lambda}_{2n}(t) - \Lambda_{20}(t), \hat{\beta}_n(t) - \beta_0(t)\right\}$ converges in distribution to mean-zero Gaussian process in $l^\infty(\mathscr{F}_{BV} \times \mathscr{F}_\beta)$. Finally, since $\hat{\Lambda}_{1n}$ is obtained using the same estimation as $\hat{\Lambda}_{2n}$ by reversing group labels, a similar

asymptotically linear expansion holds for $n^{1/2}\int h_1 d\left(\hat{\Lambda}_{1n}-\Lambda_{10}\right)$. Hence, we conclude that $n^{1/2}\left\{\hat{\Lambda}_{1n}(t)-\Lambda_{10}(t),\hat{\Lambda}_{2n}(t)-\Lambda_{20}(t)\right\}$ converges in distribution to a mean-zero Gaussian process in $l^{\infty}(\mathscr{F}_{BV}\times\mathscr{F}_{BV})$.

From the asymptotic linear expansion of $n^{1/2}\left\{\hat{\Lambda}_{kn}(t)-\Lambda_{k0}(t)\right\}$, we note that for any fixed *t*, the influence function of $\hat{\Lambda}_{kn}$ is on the tangent space of the score functions. Therefore, the estimators are semi-parametrically efficient in metric space $l^{\infty}(\mathscr{F}_{BV}\times\mathscr{F}_{BV})$ according to Theorem 18.8 in Kosorok (2008). We have completed the proof of Theorem 2.

References

Alcalay RN, Mirelman A, Saunders-Pullman R, Tang M, Mejia Santana H, Raymond D, Roos E, Orbe-Reilly M, Gurevich T, Bar Shira A, et al. Parkinson's disease phenotype in Ashkenazi jews with and without LRRK2 G2019S mutations. Mov. Disord. 2013; 28:1966–1971. [PubMed: 24243757]

Bickel, PJ.; Klaassen, CAJ.; Ritov, Y.; Wellner, JA. Efficient and Adaptive Estimation for Semiparametric Models. Springer; New York: 1993.

Cai T, Hyndman RJ, Wand W. Mixed model-based hazard estimation. Journal of the Computational and Graphical Statistics. 2002; 11:784–798.

Cheng G, Wang X. Semiparametric additive transformation model under current status data. Electron. J. Statist. 2011; 5:1735–1764.

de Boor, C. A Practical Guide to Splines. Springer; Wroclaw: 1978.

Diao G, Lin D. Semiparametric methods for mapping quantitative trait loci with censored data. Biometrics. 2005; 61:789–798. [PubMed: 16135030]

Driver JA, Logroscino G, Gaziano JM, Kurth T. Incidence and remaining lifetime risk of Parkinson disease in advanced age. Neurology. 2009; 72:432–438. [PubMed: 19188574]

Fine JP, Zou F, Yandell BS. Nonparametric estimation of mixture models, with application to quantitative trait loci. Biostatistics. 2004; 5:501–513. [PubMed: 15475415]

Goldwurm S, Tunesi S, Tesei S, Zini M, Sironi F, Primignani P, Magnani C, Pezzoli G. Kin-cohort analysis of LRRK2-G2019S penetrance in Parkinson's disease. Mov. Disord. 2011; 26:2144–2145. [PubMed: 21714003]

Hall P, Zhou XH. Nonparametric estimation of component distributions in a multivariate mixture. Ann. Statist. 2003; 31:201–224.

Healy DG, Falchi M, O'Sullivan SS, Bonifati V, Durr a. Bressman S, Brice a. Aasly J, Zabetian CP, Goldwurm S, et al. Phenotype, genotype, and worldwide genetic penetrance of LRRK2-associated Parkinson's disease: a case-control study. Lancet Neurol. 2008; 7:583–590. [PubMed: 18539534]

Hentati F, Trinh J, Thompson C, Nosova E, Farrer MJ, Aasly JO. LRRK2 Parkinsonism in Tunisia and Norway: A comparative analysis of disease penetrance. Neurology. 2014; 83:568–569. [PubMed: 25008396]

Kachergus J, Mata IF, Hulihan M, Taylor JP, Lincoln S, Aasly J, Gibson JM, Ross OA, Lynch T, Wiley J, et al. Identification of a novel LRRK2 mutation linked to autosomal dominant Parkinsonism: evidence of a common founder across European populations. American Journal of Human Genetics. 2005; 76:672–680. [PubMed: 15726496]

Kosorok, M. Introduction to Empirical Processes and Semiparametric Inference. Springer; New York: 2008.

Langbehn DR, Brinkman RR, Falush D, Paulsen JS, Hayden MR. A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. Clinical Genetics. 2004; 65:267–277. [PubMed: 15025718]

Larid NM, Ware J. Random-effect models for longitudinal data. Biometrics. 1982; 38:963–974. [PubMed: 7168798]

Latourelle JC, Sun M, Lew MF, Suchowersky O, Klein C, Golbe LI, Mark MH, Growdon JH, Wooten GF, Watts RL, et al. The Gly2019Ser mutation in LRRK2 is not fully penetrant in familial Parkinson's disease: the GenePD study. BMC medicine. 2008; 6:32. [PubMed: 18986508]

Lesage S, Leutenegger AL, Ibanez P, Janin S, Lohmann E, Durr a. Brice a. French Parkinson's Disease Genetics Study Group. LRRK2 haplotype analyses in European and North African families with Parkinson disease: a common founder for the G2019S mutation dating from the 13th century. American Journal ofHuman Genetics. 2005; 77:330.

Ma Y, Wang Y. Efficient distribution dstimation for data with unobserved sub-population identifiers. Electronic Journal ofStatistics. 2012; 6:710–737.

Marder K, Levy G, Louis ED, Mejia-Santana H, Cote L, Andrews H, Harris J, Waters C, Ford B, Frucht S, Fahn S, Ottman R. Accuracy of family history data on Parkinson's disease. Neurology. 2003; 61:18–23. [PubMed: 12847150]

Marder K, Tang M, Alcalay R, Mejia-Santana H, Raymond D, Mirelman a. Saunders-Pullman R, Clark L, Ozelius L, Orr-Urtreger A, et al. Age specific penetrance of the LRRK2 G2019S mutation in the Michael J Fox Ashkenazi Jewish (AJ) LRRK2 consortium. Neurology. 2014; 82(10 Supplement):S17–002.

Mclachlan, GJ.; Basford, KE. Mixture Models, Inference and Applications to Clustering. Dekker; New York: 1988.

Orr-Urtreger A, Shifrin C, Rozovski U, Rosner S, Bercovich D, Gurevich T, Yagev-More H, Bar-Shira A, Giladi N. The LRRK2 G2019S mutation in Ashkenazi Jews with Parkinson's disease: Is there a gender effect? Neurology. 2007; 69:1595–1602. [PubMed: 17938369]

Paisán-Ruíz C, Jain S, Evans EW, Gilks W. p. Simón J, van der Brug M, de Munain AL, Aparicio S, Gil AM, Khan N, et al. Cloning of the gene containing mutations that cause PARK8-linked Parkinson's disease. Neuron. 2004; 44:595–600. [PubMed: 15541308]

Qin J, Garcia T, Ma Y, Tang M, Marder K, Wang Y. Combining isotonic regression and EM algorithm to predict genetic risk under monotonicity constraint. The Annals of Applied Statistics. 2014; 8:1182–1208. [PubMed: 25404955]

Rudin, W. Functional Analysis. McGraw-Hill; New York: 1973.

Schumaker, L. Spline Functions: Basic Theory. Cambridge University Press; Cambridge: 2007.

Teodorescu B, Van Keilegom I, Cao R. Generalized time-dependent conditional linear models under left truncation and right censoring. Annals of the Institute of Statistical Mathematics. 2010; 62:465–485.

Titterington, DM.; Smith, AFM.; Markov, UE. Statistical Analysis of Finite Mixture Distributions. Wiley; Chichester: 1985.

Trinh J, Farrer M. Advances in the genetics of Parkinson disease. Nature Reviews Neurology. 2013; 9:445–454. [PubMed: 23857047]

Trinh J, Amouri R, Duda JE, Morley JF, Read M, Donald a. Farrer MJ. A comparative study of Parkinson's disease and leucine-rich repeat kinase 2 p. G2019S Parkinsonism. Neurobiology of Aging. 2014; 35:1125–1131. [PubMed: 24355527]

van der Vaart, A.; Wellner, J. Weak Convergence and Empirical Processes. Springer; New York: 1996.

Wacholder S, Hartge P, Struewing J, Pee D, McAdams M, Brody L, Tucker M. The Kin-Cohort Study for Estimating Penetrance. American Journal of Epidemiology. 1998; 148:623–630. [PubMed: 9778168]

Wang Y, Clark LN, Louis ED, Mejia-Santana H, Harris J, Cote LJ, Waters C, Andrews D, Ford B, Frucht S. Risk of Parkinson's disease in carriers of Parkin mutations: estimation using the kin-cohort method. Arch. Neurol. 2008; 65:467–474. [PubMed: 18413468]

Wang Y, Garcia T, Ma Y. Nonparametric estimation for censored mixture data with application to the Cooperative Huntington's Observational Research Trial. J. Amer. Statist. Assoc. 2012; 107:1324–1338.

Wang Q, Tong X, Sun L. Exploring the varying covariate effects in proportional odds models with censored data. Journal of Multivariate Analysis. 2012; 109:168–189.

Yang W, Hormozdiari F, Wang Z, He D, Pasaniuc B, Eskin E. Leveraging reads that span multiple single nucleaotide polymorphisms for haplotype inference from sequencing data. Bioinformatics. 2013; 29:2245–2252. [PubMed: 23825370]

Zeng D, Lin D. A general asymptotic theory for maximum likelihood estimation in semiparametric regression models with censored data. Statistica Sinica. 2010; 20:871–910. [PubMed: 20577580]

Zeng D, Lin D, Avery CL, North KE. Efficient Semiparametric Estimation of Haplotype-disease Associations in Case-cohort and Nested Case-control Studies. Biostatistics. 2006; 7:486–502. [PubMed: 16500923]

Zhang H, Olschwang S, Yu K. Statistical inference on the penetrances of rare genetic mutations based on a case-family design. Biostatistics. 2010; 11:519–532. [PubMed: 20179148]

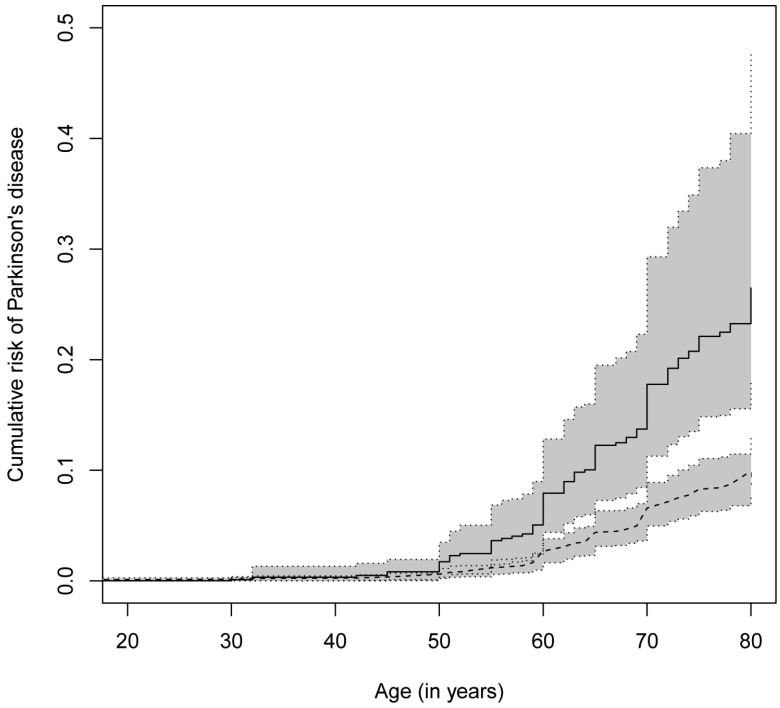

Fig. 1.

Estimated cumulative risk functions for Parkinson's disease onset in the leucine-rich repeat kinase 2 carriers and non-carriers. The solid curve is the estimated distribution function for carriers and the dashed curve is for non-carriers. The dotted curves are their pointwise 95% confidence intervals. The shaded regions indicate area covered in the pointwise confidence interval.

Table 1

Summary results for the estimated distribution functions in the first simulation scenario ($\times 10^{-2}$)

			Case I								Case II						
			Proposed				EM-PAVA				Proposed				EM-PAVA		
n	c%		Bias	SD	SE	CP	Bias	SD	Ratio	Bias	SD	SE	CP	Bias	SD	Ratio
100	20%	$F_1(Q_{0.25})$	–0·8	7·2	7·1	93	–0·4	7·4	105·5	–1·3	8·9	7·9	90	–0·1	9·5	112·2
		$F_1(Q_{0.50})$	–0·8	9·0	8·9	93	–0·2	9·2	104·6	–2·7	10·9	10·4	92	0·0	11·6	114·0
		$F_1(Q_{0.75})$	–0·8	8·0	8·1	94	–0·1	8·4	108·5	–3·2	10·4	10·1	92	0·6	11·8	130·0
		$F_2(Q_{0.25})$	0·6	7·9	8·1	94	0·4	8·1	104·7	2·8	10·9	10·9	93	0·4	11·0	101·0
		$F_2(Q_{0.50})$	0·4	9·0	8·8	94	0·3	9·2	103·5	3·8	10·6	10·9	94	0·3	12·1	128·8
		$F_2(Q_{0.75})$	–0·5	7·7	7·3	94	–0·3	8·0	108·7	2·8	8·4	7·8	89	1·9	9·5	130·0
	40%	$F_1(Q_{0.25})$	–0·6	7·8	7·2	92	–0·2	7·9	103·1	–1·1	8·2	7·7	92	0·0	8·6	110·6
		$F_1(Q_{0.50})$	–0·7	9·0	9·1	94	–0·1	9·3	106·7	–2·9	10·6	10·3	92	–0·4	11·8	123·4
		$F_1(Q_{0.75})$	–1·1	9·2	8·8	92	–0·2	9·6	109·7	–4·8	11·3	10·4	90	–1·0	13·0	130·8
		$F_2(Q_{0.25})$	0·0	8·4	8·3	92	–0·1	8·6	102·7	2·4	11·2	10·7	91	0·0	12·1	115·4
		$F_2(Q_{0.50})$	0·3	10·0	9·7	94	0·3	10·1	102·1	4·1	12·2	11·7	91	1·3	13·7	125·2
		$F_2(Q_{0.75})$	–2·4	12·0	10·2	88	2·2	14·2	140·8	0·6	11·3	10·7	88	7·0	14·3	160·1
300	20%	$F_1(Q_{0.25})$	–0·5	4·3	4·3	94	–0·4	4·4	102·3	–0·8	5·3	5·3	94	–0·3	5·4	105·7
		$F_1(Q_{0.50})$	–0·1	5·3	5·2	94	0·0	5·3	103·3	–1·7	6·6	6·8	95	–0·7	6·8	108·6
		$F_1(Q_{0.75})$	–0·4	4·8	4·7	95	–0·2	5·0	110·7	–1·9	6·5	6·7	94	–0·2	7·2	122·8
		$F_2(Q_{0.25})$	–0·1	4·7	4·8	95	–0·1	4·7	103·7	1·5	6·9	6·9	95	0·6	6·9	100·2
		$F_2(Q_{0.50})$	0·0	5·0	5·1	96	0·0	5·0	103·6	1·9	7·2	7·0	95	0·4	7·7	113·1
		$F_2(Q_{0.75})$	0·0	4·5	4·4	95	0·0	4·5	103·0	2·4	5·4	5·2	91	1·4	5·7	112·1
	40%	$F_1(Q_{0.25})$	0·1	4·5	4·4	95	0·2	4·5	100·7	–0·6	5·2	5·2	94	0·1	5·4	106·0
		$F_1(Q_{0.50})$	0·3	5·4	5·4	96	0·3	5·4	101·5	–1·6	6·8	7·0	95	–0·5	7·1	108·3
		$F_1(Q_{0.75})$	0·1	5·0	5·1	96	0·1	5·1	107·0	–2·4	6·9	7·1	94	–0·7	7·6	121·7
		$F_2(Q_{0.25})$	–0·4	4·9	5·0	95	–0·4	4·9	101·0	1·7	6·9	7·2	95	0·6	6·9	100·5
		$F_2(Q_{0.50})$	–0·3	5·7	5·8	95	–0·3	5·9	104·1	2·6	7·5	7·7	94	0·7	8·0	114·2
		$F_2(Q_{0.75})$	–0·7	7·5	7·1	92	0·5	8·3	124·6	1·3	8·4	7·8	91	3·0	10·3	150·7

EM-PAVA denotes the method of Qin et al. (2014). $Q0{\cdot}25$, $Q0{\cdot}5$ and $Q0{\cdot}75$ denote the first to third quartiles of $F1$ and $F2$, respectively, and $c\%$ denotes the censoring rate. Bias is the average estimation bias over 500 replications; SD is the empirical standard deviation; SE is the average of the estimated standard errors from bootstraps; CP is the actual coverage probability corresponding to nominal 95% confidence intervals; and Ratio gives the relative efficiency ratio between the proposed method and the method of Qin et al. (2014).

Table 2

Summary results for the estimated distribution functions in the second simulation scenario ($\times 10^{-2}$)

		Proposed				EM-PAVA		
Censoring		**Bias**	**SD**	**SE**	**CP**	**Bias**	**SD**	**Ratio**
40%	$F_1(Q_{0\cdot25})$	–0·1	3·1	3·2	95	0·0	3·3	110·3
	$F_1(Q_{0\cdot50})$	–0·1	3·8	4·1	96	0·0	3·8	103·3
	$F_1(Q_{0\cdot75})$	–0·4	3·7	4·1	96	0·0	4·0	115·1
	$F_2(Q_{0\cdot25})$	0·1	1·3	1·3	94	0·1	1·3	103·2
	$F_2(Q_{0\cdot50})$	0·1	1·6	1·5	94	0·0	1·6	100·5
	$F_2(Q_{0\cdot75})$	0·2	1·4	1·3	93	0·0	1·4	103·9
80%	$F_1(Q_{0\cdot25})$	–0·4	4·2	4·1	94	–0·3	4·3	102·5
	$F_1(Q_{0\cdot50})$	–0·7	5·4	5·8	94	–0·4	5·6	104·7
	$F_1(Q_{0\cdot75})$	–1·1	6·0	6·5	95	–0·3	6·4	112·5
	$F_2(Q_{0\cdot25})$	0·1	1·8	1·8	95	0·0	1·8	100·8
	$F_2(Q_{0\cdot50})$	0·2	2·5	2·6	95	0·0	2·5	101·4
	$F_2(Q_{0\cdot75})$	–0·2	4·0	3·7	93	1·0	4·2	107·2

For footnotes see Table 1.

Table 3

Estimated cumulative risk of Parkinson's disease onset in leucine-rich repeat kinase 2 carriers and non-carriers in the Ashkenazi Jewish leucine-rich repeat kinase 2 Consortium study ($\times 10^{-2}$)

	Carrier $F_1(\cdot)$			Non-Carrier $F_2(\cdot)$		
Age	**Est.**	**SE**	**95% CI**	**Est.**	**SE**	**95% CI**
20	0·0	0·0	(0·0, 0·1)	0·1	0·1	(0·0, 0·3)
30	0·1	0·1	(0·0, 0·3)	0·1	0·1	(0·0, 0·3)
40	0·3	0·4	(0·0, 1·4)	0·2	0·1	(0·0, 0·4)
50	1·8	0·8	(0·5, 3·4)	0·6	0·2	(0·3, 1·1)
60	8·1	1·9	(4·8, 12·5)	2·8	0·6	(1·6, 4·1)
70	18·3	3·9	(11·2, 26·2)	6·8	1·1	(4·9, 9·0)
80	27·4	5·7	(17·6, 39·1)	10·4	1·4	(7·8, 13·2)

95% CI, 95% confidence interval for estimated value.