Biometrika (2007), **94**, 2, *pp*. 403–414
© 2007 Biometrika Trust
Printed in Great Britain

doi:10.1093/biomet/asm027
Advance Access publication 10 May 2007

Nonparametric estimation of age-at-onset distributions from censored kin-cohort data

BY YUANJIA WANG

Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 W. 168th St., New York, New York 10032, U.S.A.
wang@stat.columbia.edu

LORRAINE N. CLARK

Department of Pathology, College of Physicians and Surgeons, Columbia University, 630 W. 168th St., New York, New York 10031, U.S.A.
lc654@columbia.edu

KAREN MARDER

Department of Neurology and Psychiatry, College of Physicians and Surgeons, Columbia University, 630 W. 168th St., New York, New York 10031, U.S.A.
kmarder@sergievsky.cpmc.columbia.edu

AND DANIEL RABINOWITZ

Department of Statistics, Columbia University, 1255 Amsterdam Ave., New York, New York 10027, U.S.A.
dan@stat.columbia.edu

SUMMARY

We present a nonparametric estimator of genotype-specific age-at-onset distributions from kin-cohort data. Standard error calculations are derived and the methodology is illustrated through an analysis of the influence of mutations of the Parkin gene on Parkinson's disease. Semiparametric efficiency considerations are briefly discussed.

Some key words: Family data; Genotype; Parkinson's disease; Parkin mutation; Proband; Relative; Survival analysis.

1. INTRODUCTION

When an association between a disease and a genotype at a particular genetic locus is found, it may be of interest to estimate the genotype-specific age-at-onset distributions and to compare the distributions between genotypes.

However, it may be that the genotypes of greatest interest are rare mutations. When a particular genotype is rare, population-based cohort studies may result in too few carriers to allow for accurate estimation of the age-at-onset distribution associated with that genotype. Prevalence case-control studies can enrich the sample with carriers, but with such designs the genotype-specific distributions are generally not completely identifiable.

An alternative study design, the kin-cohort design, has been proposed for estimating genotype-specific distributions in the presence of rare genotypes. In this design, a sample of so-called probands, composed mostly of subjects who have experienced onset, is recruited. Genotype information is obtained from the probands, and disease status information, such as age-at-onset, is obtained from the relatives of the probands. However, genotype information is usually not obtained in the relatives; see for example Struewing et al. (1997), Wacholder et al. (1998), Gail et al. (1999a), Gail et al. (1999b), Chatterjee & Wacholder (2001), Chatterjee et al. (2001), Moore et al. (2001) and Saunders & Begg (2003).

Recruitment of probands may be dependent on the probands' age-at-onset. In this situation, the probands' age-at-onset information may not accurately represent the genotype-specific distributions of age-at-onset in the population. However, for some sampling schemes it may be appropriate to assume that the age-at-onset in the relatives is conditionally independent of recruitment of the probands, given the probands' genotype. In this situation, the experience of the relatives can be representative of the relationship between age-at-onset and genotype in the general population.

The difficulty with using the relatives' age-at-onset data is that the genotype information in the relatives is generally not observed. However, relatives' genotypes can be inferred from the probands' genotypes. The central issue in the estimation of genotype-specific distributions from kin-cohort data is combining probands' genotype information with age-at-onset in the relatives.

Several approaches have been developed for estimating genotype-specific age-at-onset distributions from kin-cohort designs. Struewing et al. (1997) & Wacholder et al. (1998) proposed nonparametric estimation when there were two genotype-specific distributions to be estimated. Chatterjee & Wacholder (2001) proposed a more general nonparametric maximum likelihood estimator that allows for estimation of age-at-onset for more than two genotypes. However, the large sample properties of nonparametric maximum likelihood estimators are unknown, and consistency of the likelihood-based variance estimators for the nonparametric maximum likelihood estimators has not been established.

An alternative approach to efficient estimation from kin-cohort data was proposed in Y. Wang's 2005 Ph.D. dissertation at Columbia University. However the approach there is only applicable to uncensored outcomes. Here, we present an estimator similar in spirit to that proposed in Wang's dissertation, but which is applicable to censored age-at-onset data, and an approach to computing standard errors.

2. Estimation

Let $\mathcal{I}$ denote the event that the probands are ascertained and their relatives are included in a kin-cohort study sample. Let n denote the number of relatives in the sample, and let i index these relatives. Let p denote the number of possible genotypes, and let j index the genotypes. Let $\mathcal{G}$ denote the genotype information available in the sample of probands, and let π_{ij} denote the conditional probability, given $\mathcal{G}$, that the ith study subject is carrying the jth genotype. Let π_i denote the p-vector whose jth component is π_{ij}. The π_i are calculated using the laws of Mendelian inheritance and, when necessary, estimates of population genotype frequencies. The calculation of π_i is discussed in some detail in the illustrative data analysis in §3, and an example of π_i is listed in Table 1.

Let T_i denote the the ith subject's age-at-onset, and let $N_i(s)$ denote the indicator that T_i is less than or equal to s. Let C_i denote the observed age of the ith subject at the time of

Table 1. *The calculation of π_i for first-degree relatives of probands. Here $p = 1 - q$ is the population frequency of mutation*

Parent or child of a proband			Sibling of a proband		
	Parent or child	Parent or child		Sibling	Sibling
Proband	AA or Aa	aa	Proband	AA or Aa	aa
AA	1	0	AA	$-\frac{1}{4}p^2 + \frac{1}{2}p + \frac{3}{4}$	$\frac{1}{4}(1-p)^2$
Aa	$\frac{1}{2}(1+p)$	$\frac{1}{2}(1-p)$	Aa	$-\frac{1}{4}p^2 + \frac{3}{4}p + \frac{1}{2}$	$\frac{1}{4}p^2 - \frac{3}{4}p + \frac{1}{2}$
aa	p	q	aa	$-\frac{1}{4}p^2 + p$	$\frac{1}{4}p^2 - p + 1$

ascertainment, and let $\mathcal{C}$ denote the entire collection of censoring times. Let $R_i(s)$ denote the indicator that C_i is greater than s. It is assumed that the T_i are conditionally independent of $\mathcal{I}$ and $\mathcal{C}$, given $\mathcal{G}$; three sufficient conditions for this assumption and sampling schemes that satisfy these conditions are discussed in §4. Let F_j denote the distribution function of age-at-onset among carriers of the jth genotype, and let F denote the corresponding vector of conditional distribution functions. Finally, let G_i denote the conditional distribution of T_i given the available genotype information in the probands: $G_i = \pi_i^{\mathrm{T}} F$.

The subject-specific distributions G_i are linear combinations of components of F. If the number of distinct π_i, and thus the number of distinct G_i, is the same as the number of genotypes, and the distinct π_i are linearly independent, then the components of F may be written as unique linear combinations of the distinct G_i. In this case, the genotype-specific distributions can be unbiasedly estimated as linear combinations of the subject-specific empirical distribution functions computed separately in each of the subsets of relatives corresponding to the distinct π_i; see Table 1. When there is not a unique linear transformation, then any generalized inverse of the mapping that takes the G_i to F may be applied to the degenerate subject-specific empirical distribution functions to obtain unbiased estimators of F. However, with censored age-at-onset data, these subject-specific empirical distribution functions are not available.

In order to motivate the approach taken with censored data, it is useful to view linear combinations of estimators of the G_i as integrals of linear combinations of the empirical estimators of dG_i. Although censored subjects cannot contribute to empirical estimators of $dG_i(s)$ for s exceeding C_i, as long as there is sufficient variability in the π_i among those subjects with censoring times exceeding s that the map $dG_i(s) = \pi_i^{\mathrm{T}} dF(s)$, when restricted to those subjects, is invertible, then $dF(s)$ may be estimated by linear combinations of estimators of dG_i restricted to the subjects with censoring times exceeding s; natural estimators of $dG_i(s)$ for relatives with $C_i > s$ are given by $dN_i(s)$. The estimators of dG_i may then be integrated to obtain estimators of F.

Although there is generally an infinite set of linear combinations of the degenerate estimators of $dG_i(s)$ among subjects with $C_i > s$ that provide unbiased estimators for $dF(s)$, each corresponding to a generalized inverse, there is a linear combination that is particularly convenient:

$$\begin{aligned} d\hat{F}(s) &= \sum_{i:R_i(s)=1} M(s)\pi_i d\hat{G}_i(s) \\ &= \sum_{i:R_i(s)=1} M(s)\pi_i dN_i(s), \qquad (1) \end{aligned}$$

where

$$M(s) = \left(\sum_i \pi_i R_i(s)\pi_i^{\mathrm{T}}\right)^{-1}.$$

A computational formula for the resulting estimator is

$$\hat{F}(t) = \sum_{i:\, T_i \leqslant t} M(T_i)\pi_i R_i(T_i).$$

The condition for the existence of the estimator is that the matrix $\sum_i \pi_i R_i(s)\pi_i^{\mathrm{T}}$ be invertible for $s \in (0, t)$. This invertibility is equivalent to the invertibility of the mapping $G_i(s) = \pi_i^{\mathrm{T}} F(s)$, for $C_i > s$. When there is not sufficient variability in π_i for subjects with $R_i(s) = 1$, $dF(s)$ is not identifiable.

To show that the estimator is unbiased, note that

$$\begin{aligned} E\{d\hat{F}(s)\} &= E\big[E\{d\hat{F}(s) \mid \mathcal{I},\ \mathcal{C},\ \mathcal{G}\}\big] \\ &= E\big[E\{\sum_{i:R_i(s)=1} M(s)\pi_i dN_i(s) \mid \mathcal{I},\ \mathcal{C},\ \mathcal{G}\}\big] \\ &= E\big[E\{\sum_i M(s)\pi_i R_i(s) dN_i(s) \mid \mathcal{I},\ \mathcal{C},\ \mathcal{G}\}\big] \\ &= E\big[\sum_i M(s)\pi_i R_i(s) E\{dN_i(s) \mid \mathcal{I},\ \mathcal{C},\ \mathcal{G}\}\big] \\ &= E\sum_i M(s)\pi_i R_i(s) dG_i(s) \qquad (2) \\ &= E\sum_i M(s)\pi_i R_i(s)\pi_i^{\mathrm{T}} dF(s) \\ &= EM(s)\big\{\sum_i \pi R_i(s)\pi_i^{\mathrm{T}}\big\} dF(s) \\ &= dF(s). \end{aligned}$$

Here the second equality follows from the definition of $d\hat{F}(s)$, and the third follows from the definition of $R_i(s)$. The fourth equality follows because π_i and $R_i(s)$ are in the σ-algebra generated by $\mathcal{C}$ and $\mathcal{G}$, and the fifth because of the assumption that the relatives' age-at-onset, T_i, are conditionally independent of the ascertainment of the probands and the inclusion of their relatives, $\mathcal{I}$, and the ascertainment times, $\mathcal{C}$, given the available genotype information in the probands, $\mathcal{G}$. The sixth equality follows from the relationship $dG_i(s) = \pi_i^{\mathrm{T}} dF(s)$, and the last from the definition of $M(s)$.

The asymptotic normality of $\sqrt{n}\{\hat{F}(t) - F(t)\}$ is discussed in Appendix 1.

We now consider estimation of the covariance matrix. Since

$$dF(s) = \sum_i M(s)\pi_i R_i(s) dG_i(s),$$

it follows that $\hat{F}(t) - F(t)$ may be expressed as a sum of zero-mean subject-specific terms,

$$\sum_i \int M(s)\pi_i R_i(s) d\big\{N_i(s) - G_i(s)\big\}.$$

The covariance matrix of $\hat{F}(t) - F(t)$ is therefore

$$\sum_i E\biggl(\int_0^t M(s)\pi_i R_i(s)\{dN_i(s) - dG_i(s)\}\biggr)^{\otimes 2}, \tag{3}$$

which suggests estimating the covariance matrix by

$$\hat{\Sigma}(t) = \sum_i \biggl(\int_0^t M(s)\pi_i R_i(s)\bigl\{dN_i(s) - d\hat{G}_i(s)\bigr\}\biggr)^{\otimes 2}, \tag{4}$$

where $d\hat{G}_i(s) = \pi_i^{\mathrm{T}} d\hat{F}(s)$. The corresponding estimator for $\mathrm{cov}\bigl\{\hat{F}(t_1), \hat{F}(t_2)\bigr\}$ is

$$\sum_i \biggl(\int_0^{t_1} M(u)\pi_i R_i(u)\bigl\{dN_i(u) - d\hat{G}_i(u)\bigr\}\biggr)\biggl(\int_0^{t_2} M(v)\pi_i R_i(v)\bigl\{dN_i(v) - d\hat{G}_i(v)\bigr\}\biggr)^{\mathrm{T}}. \tag{5}$$

The estimator (4) is appropriate when there is only one relative from each family. When there are multiple members from each family, the corresponding estimator would be

$$\sum_k \biggl(\sum_\ell \int_0^t M(s)\pi_{k\ell} R_{k\ell}(s)\bigl\{dN_{k\ell}(s) - d\hat{G}_{k\ell}(s)\bigr\}\biggr)^{\otimes 2}.$$

Here k indexes family and ℓ indexes subjects within a family.

3. Data analysis

Parkinson's disease is a neurodegenerative disorder affecting approximately one to two percent of the population aged 65 or older. Mutations in several genes have been identified in Parkinson's patients; see Polymeropoulos et al. (1997), Kitada et al. (1998), Leroy et al. (1998), Valente et al. (2004) and Paisan-Ruiz et al. (2004). Among these genes, the Parkin gene is emerging as relatively important. The frequency of mutations in the Parkin gene is estimated to be 50% in familial early-onset Parkinson's patients and 18% in sporadic Parkinson's patients (Lucking et al., 2000). In late-onset Parkinson's patients the frequency of mutations in the Parkin gene is estimated at 2% (Oliveira et al., 2003). Although Parkin mutations appear to be associated with both early- and late-onset Parkinson's, the effect of the Parkin gene on age-at-onset of Parkinson's is unknown.

A kin-cohort design to study the association between Parkin mutations and Parkinson's disease symptoms was conducted at Columbia University Medical Center (Marder et al., 2003a). Altogether 487 probands with Parkinson's disease (cases) and 409 probands without disease (controls) were recruited. Recruitment of the probands was carried out without knowledge of the family history of the disease; this independence of the recruitment of the probands and the family history of Parkinson's in the relatives corresponds to the assumption that T_i is conditionally independent of $\mathcal{I}$ and $\mathcal{C}$, given $\mathcal{G}$, used in (2).

The probands were then interviewed either face-to-face or over the telephone for ascertainment of Parkinson's and other neurological disease in their first-degree relatives. To ensure the validity of the interview, the family history interviews were also administered to a second person, preferably the first-degree relative himself or herself when possible. An algorithm was created to generate a final diagnosis for each first-degree relative based on all possible interview information; see Marder et al. (2003b).

There were 224 early-onset, i.e. age-at-onset less than or equal to 50, case probands and 105 control probands analysed for Parkin mutations by sequence analysis and semiquantitative multiplex polymerase chain reaction. An analysis of a subset of 101 cases and 105 controls was previously reported (Clark et al., 2006). Twenty-seven Parkinson's probands were found to carry mutations, among whom nine carried homozygous or compound heterozygous mutations. None of the control probands carried a mutation. There were 1976 first-degree relatives of the probands included in the analysis, including 634 parents, 734 siblings and 608 children. None of the relatives was genotyped.

The population prevalence of Parkin mutations is unknown, but may be estimated to be 0·004% as described in Appendix 2. In the calculation the rate of mutations in disease-free subjects was taken to be zero. If the upper bound of 0·028 as estimated in Clark et al. (2006) is used instead, the estimate of the population frequency becomes 0·03. Both values were used in the analyses to examine the sensitivity of the method to the misspecification of mutation frequency.

It would be desirable to estimate the distribution of age-at-onset separately in homozygous carriers and heterozygous carriers. However, because of the low population frequency of Parkin mutations, only very few of the relatives are expected to be homozygous carriers. The homozygous and the heterozygous carriers were combined into a single group in the analysis described here.

The conditional probabilities that the relative of a proband carries one or more copies of a mutation, given the proband's genotype status, can be calculated using the Mendelian law and estimated Parkin mutation frequency. These results are summarized in Table 1.

The estimated conditional distributions of age-at-onset for Parkinson's disease, given a subject carrying one or more copies of mutation, versus the corresponding distribution, given a subject carrying no mutation, and standard errors of the estimates are recorded in Table 2.

Note that $\hat{F}(t)$, the integral of the estimator (1), is not monotone. This reflects the small number of failures times available in the dataset. We applied the pooled-adjacent-violators algorithm (Barlow et al., 1972, Ch.2) to provide a monotone version; see Table 3 and Fig. 1.

The sample contained 27 relatives who developed Parkinson's disease, among whom four were relatives of Parkin-carrier probands. The cumulative incidence of Parkinson's disease by age 70 increased to 6·1%, with 95% confidence interval (1·0%, 18·6%) for carriers and to 1·0%, with 95% confidence interval (0·5%, 2·3%) for noncarriers. The distribution

Table 2. *Parkinson's disease study. Estimated values and standard errors of the estimator of the genotype-specific conditional distributions*

	Carriers		Noncarriers	
Age	Estimate	Std. Err.	Estimate	Std. Err.
25	0·013	0·013	0·000	0·000
30	0·027	0·019	0·000	0·000
35	0·027	0·019	0·000	0·000
40	0·027	0·019	0·001	0·001
45	0·013	0·014	0·002	0·001
50	0·013	0·014	0·002	0·001
55	0·000	0·001	0·004	0·002
60	0·000	0·001	0·008	0·003
65	0·023	0·026	0·009	0·003
70	0·061	0·045	0·010	0·004

Std. Err., Estimated standard error

Table 3. *Parkinson's disease study. Monotone version of the estimated genotype-specific conditional distributions*

Age	Carriers Estimate	Noncarriers Estimate
25	0·011	0·000
30	0·011	0·000
35	0·011	0·000
40	0·011	0·001
45	0·011	0·002
50	0·011	0·002
55	0·011	0·004
60	0·011	0·008
65	0·023	0·009
70	0·061	0·011

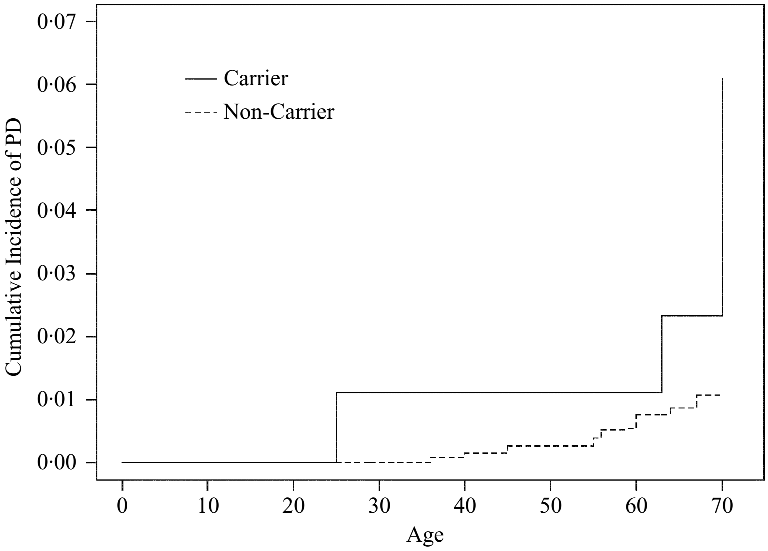

Fig. 1. Parkinson's disease study. Distribution of age-at-onset for Parkin carriers and noncarriers.

of age-at-onset up to age 70 in noncarriers is very similar to the distribution in the general population (Marder et al., 2003a). However, it should be noted that the shape of the distribution of the carriers is largely determined by the four subjects who were relatives of Parkin carrier probands and developed disease at the ages of 25, 29, 63 and 70.

To test for a difference between the distributions of age-at-onset in Parkin carriers and noncarriers, at a specific value t, one can refer

$$\frac{\hat{F}_1(t) - \hat{F}_2(t)}{\surd\{(1, -1)\hat{\Sigma}(t)(1, -1)^{\mathrm{T}}\}}$$

to critical values of the standard normal distribution.

For $t = 25$, 50 and 70, the estimates (standard errors) of $\hat{F}_1(t) - \hat{F}_2(t)$ are 0·013(0·013), 0·011(0·014) and 0·051(0·045), giving p-values for the normalized test statistic of 0·32, 0·43 and 0·26. The analysis was carried out again with the Parkin mutation frequency replaced by 0·03: the results were almost identical, suggesting that the method can be robust against the misspecification of population mutation frequencies.

To examine the relationship between familial aggregation of early-onset Parkinson's disease and Parkin gene mutations, we restricted the analyses to the 1330 relatives of 224 early-onset Parkinson's probands and compared the distribution of the noncarriers estimated from these relatives to that in the population. The Kaplan–Meier method was used to estimate the population distribution of age-at-onset of Parkinson's using the 646 relatives of the 105 control probands and compared it to the distribution of the noncarriers estimated by the proposed method: the cumulative incidence of Parkinson's up to age 70 in the population was estimated to be 1·1% (SE: 0·6%), compared with 1·3% (SE: 0·6%) ($p = 0{\cdot}77$) in the noncarrier relatives of early-onset case probands. Various familial aggregation studies reported (Marder et al., 2003a) that relatives of early-onset cases are at increased risk of being affected with Parkinson's. The current study suggests that increased risk of Parkinson's disease up to age 70 in the relatives of early-onset Parkinson patients could be due to carrying Parkin gene mutations: the relatives of early-onset case probands who do not carry Parkin mutations are not at increased risk of Parkinson's disease compared to the general population up to age 70.

4. DISCUSSION

The assumption used in (2) to ensure that the approach is unbiased is that the age-at-onset of the relatives are conditionally independent of the inclusion of the probands and ascertainment times in relatives, given probands' genotypes. Three conditions are sufficient for this assumption to hold. The first one is that the age-at-onset in relatives are conditionally independent of the sampling of probands and the inclusion of relatives, given ages at ascertainment of the relatives, probands' genotypes and probands' age-at-onset. The second one is that the age-at-onset in relatives are conditionally independent of age-at-onset in probands, given probands' genotypes. The third condition is the usual assumption of conditional independence between failure times and censoring times given covariates, in this case probands' genotypes, in survival regression analyses.

The first condition is satisfied for designs in which the sampling of probands and the inclusion of the relatives in the study dataset are based on the probands' ages-at-onset and possibly genotypes and the ages at ascertainment of their relatives, but not on the relatives' ages-at-onset. The method presented here can therefore be used for designs in which probands with certain susceptible genotypes and phenotypes are over-sampled, and relatives with late age at ascertainment to allow for possible manifestation of disease are over-sampled.

The second condition corresponds to the association between age-at-onset in relatives and age-at-onset in probands being explained completely by association between probands' and relatives' genotypes: there is no additional familial risk factor that affects age-at-onset. If there are unmeasured risk factors clustering in the family, relatives of affected probands may exhibit a stronger association between the mutation and disease than individuals in the general population; see Gail et al. (2001) and Begg (2002). Methods based on random effect models and copula models have been proposed that take into account the residual familial aggregation of additional risk factors; see for example Hsu et al. (2004) and Chatterjee et al. (2006).

If a family member dies prior to the study, the censoring time may be taken to be the subject's age at death. In this case, the approach developed here would require the death to be independent of the subject's course of the disease.

The approach developed here is not fully efficient. To consider efficiency, it is more convenient to consider estimators of the cumulative hazard rather than the cumulative distribution function. It may be conjectured that an estimator of the form

$$d\hat{\Lambda}_j(s) = \sum_i M_{ij}(s) dN_i(s)$$

achieves the semiparametric efficiency bound. The optimal weights M_{ij} for estimating the jth genotype-specific cumulative hazard are defined implicitly by the orthogonality score integral equations for $\varphi^\star$. The score equations in this situation are

$$< S^\theta - \int S^x \varphi^\star(x) dx,\, S^x >= 0,$$

for every x, and

$$S^\theta = \sum_i \int \frac{\partial}{\partial\theta} \log \tilde{\lambda}_i^\theta(s)\{dN_i(s) - \tilde{\lambda}_i(s)^\theta Y_i(s) ds\},$$

$$S^x = \sum_i \int \frac{\partial}{\partial\theta} \log \tilde{\lambda}_i^x(s)\{dN_i(s) - \tilde{\lambda}_i^x(s) Y_i(s) ds\}.$$

Here the inner product is defined by the expectation,

$$\lambda_j^\theta(s) = \lambda_j(s) + \theta \frac{1_{(s \leqslant t)}}{t},\ \lambda_j^x(s) = \lambda_j(s) + \theta\delta_x(s) - \theta 1_{(x \leqslant t)} \frac{1_{(s \leqslant t)}}{t},$$

$$\tilde{\lambda}_i^y(s) = \frac{\sum_k \pi_{ik} S_k(s) \lambda_k^y(s)}{\sum_k \pi_{ik} S_k(s)},$$

and $\delta_x(s)$ is the Dirac delta function at x.

The optimal estimator may be expressed as

$$\hat{\Lambda}_j(t) = \sum_i \int_0^t M_{ij} dN_i(s),$$

where

$$M_{ij}(s) = \Big(\sum_i \int \frac{\pi_{ij} S_j(s)}{\sum_j \pi_{ij} S_j(s) \lambda_j(s)} \xi(s) Y_i(s) ds\Big)^{-1} \frac{\pi_{ij} S_j(s)}{\sum_j \pi_{ij} S_j(s) \lambda_j(s)} \xi(s),$$

$$\xi(s) = \Big(1 + \int_0^t \varphi^\star(u) du\Big) \frac{1_{(s \leqslant t)}}{t} - \varphi^\star(s)$$

$$- \Big\{\lambda_j(s) - \tilde{\lambda}_i(s)\Big\} \Big\{\Big(1 + \int_0^t \varphi^\star(u) du\Big) \frac{s \wedge t}{t} - \int_0^s \varphi^\star(u) du\Big\}.$$

Solving the integral equation corresponds to computing a kind of least squares solution. Unfortunately there is no explicit solution.

For testing the difference of the age-at-onset between genotypes, a weighted version of T,

$$\int_0^t w(s) d\{\hat{F}_1(s) - \hat{F}_2(s)\},$$

can also be used. Computation of standard errors for such a test statistic would rely on the covariance matrix estimator in (5).

Finally, the estimator for the distribution of age-at-onset developed here is not monotone. In some situations, for example in the graphical presentation of the estimator, it may be convenient to obtain a monotone version of the estimator. A pooled-adjacent-violators algorithm or a convex minorant algorithm may be applied (Groeneboom & Wellner, 1992).

ACKNOWLEDGEMENT

This research was supported in part from U.S. National Institutes of Health.

APPENDIX 1

Large sample property of the proposed estimator

Here we derive regularity conditions sufficient for application of the multivariate Lindeberg central limit theorem to establish the pointwise joint normality of $\surd n\{\hat{F}(t) - F(t)\}$.

First note that $\surd n\{\hat{F}(t) - F(t)\}$ may be expressed as $\sum_i Y_{ni}$, where

$$Y_{ni} = \surd n \int_0^t M(s)\pi_i R_i(s)\{dN_i(s) - dG_i(s)\}.$$

The assumptions of the multivariate central limit theorem applied to the estimator are that there be a positive definite matrix Σ such that

$$\sum_{i=1}^n \text{cov}(Y_{ni}) \to \Sigma, \tag{A1}$$

and, for any $\varepsilon > 0$,

$$\sum_{i=1}^n E\|Y_{ni}\|^2 1_{\{\|Y_{ni}\|>\varepsilon\}} \to 0, \tag{A2}$$

where $\| \ \|$ denotes the Euclidean norm.

To show (A1), note that the covariance matrix in (A1) is

$$\sum_{i=1}^n \text{cov}(Y_{ni}) = \frac{1}{n}\sum_{i=1}^n E\left(\int_0^t nM(s)\pi_i R_i(s)\{dN_i(s) - dG_i(s)\}\right)^{\otimes 2}.$$

A law of large numbers argument applied to the convergence of $nM(s)$ to its expectation, for s such that $\text{pr}(C_i > s) > 0$, together with a law of large numbers argument applied to the terms

$$\left(\int_0^t nEM(s)\pi_i R_i(s)\{dN_i(s) - dG_i(s)\}\right)^{\otimes 2},$$

provide the result (A1) for independent and identically distributed C_i and π_i.

To show (A2), first note that

$$\sum_{i=1}^n E\|Y_{ni}\|^2 = \sum_{i=1}^n E\{(Y_{ni}^1)^2 + \cdots + (Y_{ni}^p)^2\},$$

so that $\sum_{i=1}^n E\|Y_{ni}\|^2$ converges to the summation of the diagonal elements of Σ. Then note that, since

$$E\sum_{i=1}^n \|Y_{ni}\|^2 1_{\{\|Y_{ni}\|>\varepsilon\}} \leqslant E\left(\sum_{i=1}^n \|Y_{ni}\|^2\right) \sup_i 1_{\{\|Y_{ni}\|>\varepsilon\}},$$

it suffices to show that

$$\sup_i \|Y_{ni}\| \to 0. \tag{A3}$$

For (A3) to hold, note that, by

$$\sup_i \|Y_{ni}\| = \sup_i \| \frac{1}{\sqrt{n}} \int_0^t nM(s)\pi_i R_i(s)\{dN_i(s) - dG_i(s)\}\|,$$

and the convergence of $nM(s)$ to $nEM(s)$, it suffices that the assumption $nEM(s) < \infty$ hold and $\max_i \pi_i R_i(s) < \infty$. The former condition is satisfied if $\text{pr}(C_i > s) > 0$ for $s \in (0, t)$. The latter condition is satisfied by the definition of π_i and $R_i(s)$.

It follows directly that sufficient conditions for $\sqrt{n}\{\hat{F}(t) - F(t)\}$ to be asymptotically normal are that C_i and π_i be independent and identically distributed, and, for all $s \in (0, t)$, $\text{pr}(C_i > s) > 0$.

Appendix 2

Estimate of the population frequency of Parkin mutations

Let A denote the event that a subject carries a Parkin mutation, let B_1 denote the event that the subject has early-onset Parkinson's, let B_2 denote the event that the subject has late-onset Parkinson's, and let B_3 denote the event that the subject does not have Parkinson's. Here $\text{pr}(A|B_1)$ may be taken as 18% (Lucking et al., 2000), and $\text{pr}(A|B_2)$ may be taken as 2% (Oliveira et al., 2003). Clark et al. (2006) showed that in a sample of 105 disease-free subjects the frequency of Parkin mutation, $\text{pr}(A|B_3)$, was zero, with the upper 95% confidence bound 0·028. The prevalence rate for Parkinson's is estimated to be 0·1% (Mayeux et al., 1995), and 10% of Parkinson's cases are estimated to be early onsets, i.e. they develop Parkinson's before the age of 50. Therefore, the prevalence rate of the early-onset Parkinson disease, $\text{pr}(B_1)$, is about 0·01%, and the population frequency of Parkin mutation can then be estimated as

$$\begin{aligned}\text{pr}(A) &= \text{pr}(A|B_1)\text{pr}(B_1) + \text{pr}(A|B_2)\text{pr}(B_2) + \text{pr}(A|B_3)\text{pr}(B_3)\\ &= (0{\cdot}18)(0{\cdot}01\%) + (0{\cdot}02)(0{\cdot}09\%) + (0)(99{\cdot}5\%)\\ &= 0{\cdot}004\%.\end{aligned}$$

If the upper confidence bound of 0·028 for mutation frequency in disease free subjects is used, the estimate of the population frequency becomes 0·03.

References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. & Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions*. New York: John Wiley.

Begg, C. (2002). On the use of familial aggregation in population based case probands for calculating penetrance. *J. Nat. Cancer Inst.* **94**, 1221–6.

Chatterjee, N. & Wacholder, S. (2001). A marginal likelihood approach for estimating penetrance from kin-cohort designs. *Biometrics* **57**, 245–52.

Chatterjee, N., Kalaylioglu, Z., Shih, J. H. & Gail, M. H. (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., Shih, J., Hartge, P., Brody, L., Tucker, M. & Wacholder, S. (2001). Association and aggregation analysis using kin-cohort designs with applications to genotype family history data from the Washington Ashkenazi study. *Genet. Epidemiol.* **21**, 123–38.

Clark, L. N., Afridi, S., Karlins, E., Wang, Y., Mejia-santana, H., Harris, J., Louis, E., Cote, J. L., Andrews, H., Fahn, S., Waters, C., Ford, B., Frucht, S., Ottman, R. & Marder, K. (2006). Case-control study of the parkin gene in early onset PD. *Arch. Neurol.* **63**, 548–52.

Gail, M., Pee, D. & Carroll, R. (1999a). Kin-cohort designs for gene characterization. *J. Nat. Cancer Inst.* **26**, 55–60.

Gail, M., Pee, D., Benichou, J. & Carroll, R. (1999b). Designing studies to estimate the penetrance of an identified autosomal dominant mutation: cohort, case-control, and genotyped-proband designs. *J. Statist. Plan. Infer.* **96**, 167–77.

Gail, M., Pee, D., Benichou, J. & Carroll, R. (2001). Effects of violations of assumptions on likelihood methods for estimating the penetrance of an autosomal dominant mutation from kin-cohort studies. *Genet. Epidemiol.* **16**, 15–39.

Groeneboom, P. & Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Boston: Birkhäuser.

Hsu, L., Chen, L., Gorfine, M. & Malone, K. (2004). Semiparametric estimation of marginal hazard function from case-control family studies. *Biometrics* **60**, 936–44.

Kitada, T., Asakawa, S., Hattori, N., Matsumine, H., Yamamura, Y., Minoshima, S., Yokochi, M., Mizuno, Y. & Shimizu, N. (1998). Mutations in the Parkin gene cause autosomal recessive juvenile Parkinsonism. *Nature* **392**, 605–8.

Leroy, E., Boyer, R., Auburger, G., Leube, B., Ulm, G., Mezey, E., Harta, G., Brownstein, M. J., Jonnalagada, S., Chernova, T., Dehejia, A., Lavedan, C., Gasser, T., Steinbach, P. J., Wilkinson, K. D. & Polymeropoulos, M. H. (1998). The ubiquitin pathway in Parkinson's disease. *Nature* **395**, 451–2.

Lucking, C. B., Durr, A., Bonifati, V., Vaughan, J., De michele, G., Gasser, T., Harhangi, B. S., Meco, G., Denefle, P., Wood, N. W., Agid, Y., Brice, A. & French Parkinson's Disease Genetics Study Group. (2000). Association between early-onset Parkinson's disease and mutations in the Parkin gene. *New Engl. J. Med.* **342**, 1560–7.

Marder, K., Levy, G., Louis, E. D., Mejia-santana, H., Cote, L., Andrews, H., Harris, J., Waters, C., Ford, B., Frucht, S., Fahn, S. & Ottman, R. (2003a). Familial aggregation of early- and late- onset Parkinson's disease. *Ann. Neurol.* **54**, 507–13.

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

Mayeux, R., Marder, K., Cote, L. J., Denaro, J., Hemenegildo, N., Mejia, H., Tang, M., Lantigua, R., Wilder, D., Gurland, B. & Hauser, A. (1995). The frequency of idiopathic Parkinson's disease by age, ethnic group, and sex in northern Manhattan, 1998–1993. *Am. J. Epidemiol.* **142**, 820–7.

Moore, D., Chatterjee, N., Pee, D. & Gail, M. (2001). Pseudo-likelihood estimates of the cumulative risk of an autosomal dominant disease from a kin-cohort study. *Genet. Epidemiol.* **20**, 210–27.

Oliveira, S. A., Scott, W. K., Martin, E. R., Nance, M. A., Watts, R. L., Hubble, J. P., Koller, W. C., Pahwa, R., Stern, M. B., Hiner, B. C., Ondo, W. G., Allen, F. H., Scott, B. L., Goetz, C. G., Small, G. W., Mastaglia, F., Stajich, J. M., Zhang, F., Booze, M. W., Winn, M. P., Middleton, L. T., Haines, J. L., Pericak-vance, M. A. & Vance, J. M. (2003). Parkin mutations and susceptibility alleles in late-onset Parkinson's disease. *Ann. Neurol.* **53**, 624–9.

Paisan-ruiz, C., Jain, S., Evans, E. W., Gilks, W. P., Simon, J., Van der brug, M., De munain, A. L., Aparicio, S., Gil, A. M., Khan, N., Johnson, J., Martinez, J. R., Nicholl, D., Carrera, I. M., Pena, A. S., De silva, R., Lees, A., Marti-masso, J. F., Perez-tur, J., Wood, N. W. & Singleton, A. B. (2004). Cloning of the gene containing mutations that cause PARK8-linked Parkinson's disease. *Neuron* **44**, 595–600.

Polymeropoulos, M. H., Lavedan, C., Leroy, E., Ide, S. E., Dehejia, A., Dutra, A., Pike, B., Root, H., Rubenstein, J., Boyer, R., Stenroos, E. S., Chandrasekharappa, S., Athanassiadou, A., Papapetropoulos, T., Johnson, W. G., Lazzarini, A. M., Duvoisin, R. C., Di iorio, G., Golbe, L. I. & Nussbaum, R. L. (1997). Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science* **276**, 2045–7.

Saunders, C. L. & Begg, C. B. (2003). Kin-cohort evaluation of relative risks of genetic variants. *Genet. Epidemiol.* **24**, 220–9.

Struewing, J. P., Hartge, P., Wacholder, S., Baker, S. M., Berlin, M., McAdams, M., Timmerman, M. M., Brody, L. C. & Tucker, M. A. (1997). The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *New Engl. J. Med.* **336**, 1401–8.

Valente, E. M., Abou-sleiman, P. M., Caputo, V., Muqit, M. M., Harvey, K., Gispert, S., Ali, Z., Del turco, D., Bentivoglio, A. R., Healy, D. G., Albanese, A., Nussbaum, R., Gonzalez-maldonado, R., Deller, T., Salvi, S., Cortelli, P., Gilks, W. P., Latchman, D. S., Harvey, R. J., Dallapiccola, B., Auburger, G. & Wood, N. W. (2004). Hereditary early-onset Parkinson's disease caused by mutations in PINK1. *Science* **304**, 1158–60.

Wacholder, S., Hartge, P., Struewing, J., Pee, D., McAdams, M., Brody, L. & Tucker, M. (1998). The kin-cohort study for estimating penetrance. *Am. J. Epidemiol.* **148**, 623–30.

[*Received August* 2005. *Revised August* 2006]