

# Pseudo-Likelihood Estimates of the Cumulative Risk of an Autosomal Dominant Disease From a Kin-Cohort Study

Dirk F. Moore,<sup>1\*</sup> Nilanjan Chatterjee,<sup>2</sup> David Pee,<sup>3</sup> and Mitchell H. Gail<sup>2</sup>

<sup>1</sup>*Department of Statistics, Temple University, Philadelphia, Pennsylvania*

<sup>2</sup>*Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland*

<sup>3</sup>*Information Management Services, Rockville, Maryland*

Wacholder et al. [1998: *Am J Epidemiol* 148:623–629] and Struewing et al. [1997: *N Engl J Med* 336:1401–1408] have recently proposed a design called the kin-cohort design to estimate the probability of developing disease (penetrance) associated with an autosomal dominant gene. In this design, volunteers (proband) agree to be genotyped and one also determines the disease history (phenotype) of first-degree relatives of the proband. They used this design to estimate that the chance of developing breast cancer by age 70 in Ashkenazi Jewish women who carried mutations of the genes BRCA1 or BRCA2 was 0.56, a figure that was lower than previously estimated from highly affected families. The method that they used to estimate the cumulative risk of breast cancer, while asymptotically correct, does not necessarily produce monotone estimates in small samples. To obtain monotone, weakly parametric estimates, we consider separate piecewise exponential models for carriers and non-carriers. As the number of intervals on which constant hazards are assumed increases, however, the maximum likelihood score equations become unstable and difficult to solve. We, therefore, developed alternative pseudo-likelihood procedures that are readily solvable for piecewise exponential models with many intervals. We study these techniques through simulations and a re-analysis of a portion of the data used by Struewing et al. [1997] and discuss possible extensions. *Genet. Epidemiol.* 20:210–227, 2001. © 2001 Wiley-Liss, Inc.

**Key words:** age of onset; breast cancer; BRCA; penetrance; proband; survival analysis

\*Correspondence to: Dirk F. Moore, Ph.D., Department of Statistics, Speakman Hall 006-00, Temple University, Philadelphia, PA 19122. E-mail: dirk@sbm.temple.edu

Received for publication 3 February 2000; revision accepted 18 July 2000

© 2001 Wiley-Liss, Inc.

## INTRODUCTION

Wacholder et al. [1998] proposed a design called the kin-cohort design to estimate the probability of developing disease (penetrance) associated with an autosomal dominant gene. In this design, volunteers (probands) agree to be genotyped and one also determines the disease history (phenotype) of first-degree relatives of the proband. This design was used by Struewing et al. [1997] to estimate that the chance of developing breast cancer by age 70 in Ashkenazi Jewish women who carried mutations of the genes BRCA1 or BRCA2 was 0.56, a figure that was lower than previously estimated from highly affected families. Gail et al. [1999a] used the term genotyped-proband design, instead of kin-cohort design, to emphasize that the proband was genotyped, and they stressed the importance of obtaining representative samples of probands, conditional on their phenotypes, in order to obtain population-based estimates of penetrance.

To estimate the disease survival distribution (one minus the cumulative incidence function) for mutation carriers and non-carriers, Wacholder et al. [1998] used the fact that the survival distribution for first-degree relatives of probands who carried a mutation was a mixture of survival distributions for carriers and non-carriers, with mixing proportions about 50:50 for rare mutations. Likewise, the survival distributions for first-degree relatives of non-carrier probands was approximately a 0:100 mixture of carrier and non-carrier distributions. The actual mixing proportions are functions of the allele frequency  $q = P(A)$ , where  $A$  is the mutant allele. Wacholder et al. [1998] and Struewing et al. [1997] obtained Kaplan-Meier estimates of the survival distributions for first-degree relatives of carrier and non-carrier probands, respectively, and then solved two linear equations that describe the mixing to estimate the survival distributions for carriers and non-carriers. These estimates are consistent, provided consistent estimates of  $q$  are available, but the estimates of survival distributions are not necessarily monotone in small samples. Wacholder et al. [1998] and Gail et al. [1999a,b] discuss the advantage of the kin-cohort design for estimates of penetrance in comparison with other population-based designs, such as cohort- and population-based case-control designs.

Gail et al. [1999a] showed how to obtain parametric maximum likelihood estimates (mle's) of  $q$  and of the survival distributions from kin-cohort data for carriers and non-carriers for improper Weibull models that included a shape parameter, a scale parameter, and a parameter describing the probability that disease will ever develop. An advantage of this parametric approach is that it will yield monotonically increasing estimates of the cumulative incidence. In this paper, we develop methods for relaxing the parametric assumption by considering separate piecewise exponential models for carriers and non-carriers. As the number of intervals on which constant hazards are assumed increases, these models become weakly parametric. The maximum likelihood score equations become unstable and difficult to solve, however. We, therefore, have developed alternative pseudo-likelihood procedures that are readily solvable for piecewise exponential models with many intervals and extend to a fully non-parametric estimator for the survival curves.

In this paper, we present notation and methods, and we evaluate the relative efficiency of the pseudo-likelihood approaches, compared to maximum likelihood for dichotomous outcomes and for time to response outcomes with hazards assumed

to be piecewise constant in six intervals. We use exact calculation for dichotomous outcomes and simulations for time-to-response data. We re-analyze a portion of the data used by Struewing et al. [1997] using the non-parametric pseudo-likelihood approach and conclude with a discussion.

## NOTATION AND METHODS

### General Methods

Let  $Y_0$  denote the phenotype of the proband and  $Y_1^T = (Y_{11}, Y_{12}, \dots, Y_{1m})$  the array of phenotypes of relatives. In this paper, we confine attention to first-degree relatives, but many of the formulas apply more generally. For dichotomous outcomes,  $Y_0 = 1$  or  $0$  according as the proband is diseased or not. For quantitative data,  $Y_0$  is a measurement, such as blood pressure, and for survival data,  $Y_0 = (T, \delta)$  is a pair describing the age,  $T$ , at end of follow-up and the disease status,  $\delta = 1$  or  $0$  according as the proband is diseased or not. Follow-up,  $T$ , ends at the earliest of the date of disease onset or censoring. The components  $Y_{1j}$  are defined similarly. In this study, we concentrate on survival data but also consider dichotomous outcomes.

We assume an autosomal dominant disease model with mutant allele  $A$  and wild type allele  $a$ . We assume Hardy-Weinberg equilibrium, under which a randomly selected subject has genotypes  $AA$ ,  $Aa$ , or  $aa$  with probabilities  $q^2$ ,  $2q(1 - q)$ , and  $(1 - q)^2$ , respectively, where, as before,  $q = P(A)$ . Under an autosomal dominant model, the probability of disease depends only on whether a subject is a mutation carrier ( $AA$  or  $Aa$ ) or a non-carrier ( $aa$ ). Therefore, it is convenient to use the carrier frequency  $\pi = P(AA \text{ or } Aa) = 1 - (1 - q)^2$  instead of  $q$  in our calculations. Moreover, one can characterize the proband's genotype by  $g_0 = 1$  or  $0$  according as the proband is a carrier or not, and the  $m \times 1$  genotype indicator for relatives,  $g_1$ , has components  $g_{1j}$  that are defined similarly. Assuming Hardy-Weinberg equilibrium, one can use standard Mendelian calculations [e.g., Li, 1976] to obtain the conditional mass function  $p(g_1 | g_0; \pi)$ . The assumption of Hardy-Weinberg equilibrium is needed to calculate the probabilities of genotypes of pedigree founders. Gail et al. [1999a] describe simple methods of enumeration to calculate  $p(g_1 | g_0; \pi)$  for small pedigrees of the type we consider in this paper. When we need to index the  $i$ 'th family,  $i = 1, \dots, I$ , we use the notation  $y_{0i}$ ,  $y_{1i}$ ,  $y_{1ij}$ ,  $g_{0i}$ ,  $g_{1i}$ , and  $g_{1ij}$ .

We are principally interested in estimating the conditional densities (or mass functions) of phenotype given genotype, namely  $f(y_0 | g_0; \varphi)$ . For example, for dichotomous data,  $f(y_0 | g_0 = 1; \varphi_0, \varphi_1) = \varphi_1^{y_0} (1 - \varphi_1)^{1 - y_0}$  and  $f(y_0 | g_0 = 0; \varphi_0, \varphi_1) = \varphi_0^{y_0} (1 - \varphi_0)^{1 - y_0}$ . Here  $\varphi_0$  and  $\varphi_1$  are penetrance parameters for non-carriers and carriers, respectively, and  $\varphi = (\varphi_0, \varphi_1)$ . In the case of time to response data,  $f$  is the density of a survival curve characterized by parameters  $\varphi_0$  for non-carriers and  $\varphi_1$  for carriers.

In order to write the likelihood for the kin-cohort design, we use the commonly made assumption that the phenotypes of family members are conditionally independent given their genotypes. Also, by assuming Hardy-Weinberg equilibrium, we are ignoring the possibility that an individual's phenotype influences the chance that that person will transmit his or her genes. From the kin-cohort sampling scheme, we can write the likelihood for a given family as

$$f_0(g_0 | y_0; \varphi, \pi) f_1(y_1 | g_0; \varphi, \pi). \quad (1)$$

Here  $f_0$  is a conditional probability mass function for  $g_0$  and  $f_1$  is a conditional density or mass function for  $y_1$ , the vector of the phenotypes of the relatives.

The first factor in (1) reflects the assumption that probands are selected at random, conditional on their phenotypes. From Bayes' theorem,

$$f_0(g_0 | y_0; \pi, \varphi) = \frac{\pi^{g_0} (1-\pi)^{1-g_0} f(y_0 | g_0; \varphi)}{\sum_u \pi^u (1-\pi)^{1-u} f(y_0 | u; \varphi)}. \tag{2}$$

The second factor in the likelihood (1) follows from the conditional independence assumption because the conditional density of  $y_1$  given  $g_0$  and  $y_0$  is

$$f_1(y_1 | g_0; \varphi, \pi) = \sum_{g_1} \prod_{j=1}^m f(y_{1j} | g_{1j}; \varphi) p(g_1 | g_0; \pi). \tag{3}$$

The full likelihood is  $e^l = e^{l_1} \cdot e^{l_0}$ , where  $e^{l_1}$  and  $e^{l_0}$  are the product over families of  $f_1(y_1 | g_0; \varphi, \pi)$  and  $f_0(g_0 | y_0; \varphi, \pi)$ , respectively. In principle, the log-likelihood  $l$  can be maximized over  $\pi$  and  $\varphi$  and variances of  $\hat{\pi}$  and  $\hat{\varphi}$  determined from the observed information matrix. In practice, it is often convenient to evaluate the observed information matrix by numerical differentiation of the log-likelihood evaluated at the parameter estimates.

Because full maximum likelihood score equations can lead to unstable estimates and failure of convergence for piecewise exponential survival models with many parameters, we consider two pseudo-likelihood approaches instead. In the first approach, we solve the following estimating equations [Godambe, 1991], which we refer to as pseudo-likelihood equations:

$$U_{1\varphi}(\varphi, \pi) = \frac{\partial l_1}{\partial \varphi} = 0. \tag{4}$$

$$U_{0\pi}(\pi, \varphi) = \frac{\partial l_0}{\partial \pi} = 0. \tag{5}$$

Solving these equations is equivalent to alternately maximizing  $l_1$  with respect to  $\varphi$  for fixed  $\pi$  and maximizing  $l_0$  with respect to  $\pi$  for fixed  $\varphi$ , and continuing until the parameter estimates converge. Viewing  $l_1$  as a log-likelihood, we regard substitution of  $\hat{\pi}$  as a pseudo-likelihood procedure [Gong and Samaniego, 1981]. Likewise, substituting  $\hat{\varphi}_0$  and  $\hat{\varphi}_1$  into  $l_0$  can be regarded as a pseudo-likelihood procedure, which justifies our terminology. In the second pseudo-likelihood method, we consider a modification of (3) using the marginal approach introduced by Chatterjee and Wacholder [2001], which we shall refer to subsequently as CW. In the marginal approach, one ignores the relationships between the relatives of a proband, but uses the relationships between each relative and his/her proband. Thus, if a proband has a mother and a sister, the contribution of this family in (3) is obtained by treating this family as two separate families, one consisting of the mother and the proband and the other consisting of the sister and the proband. Details of the marginal approach

are presented in CW. For the rest of this study, the two pseudo-likelihood estimators will be referred to as PLE and MPLE (marginal PLE), respectively.

Using a standard Taylor series argument (see Appendix), it can be shown that as the number of families goes to  $\infty$ , PLE estimates will be asymptotically normally distributed with a variance-covariance matrix which can be consistently estimated by  $\hat{B}^{-1}\hat{\Omega}(\hat{B}^{-1})'$ , where

$$\hat{B} = \begin{bmatrix} \frac{\partial^2 l_0}{\partial \pi^2} & \frac{\partial^2 l_0}{\partial \pi \partial \varphi} \\ \frac{\partial^2 l_1}{\partial \varphi \partial \pi} & \frac{\partial^2 l_1}{\partial \varphi \partial \varphi'} \end{bmatrix}_{\hat{\pi}_{PLE}, \hat{\varphi}_{PLE}}$$

and

$$\hat{\Omega} = \begin{bmatrix} \hat{\Omega}_{11} & 0 \\ 0 & \hat{\Omega}_{22} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 l_0}{\partial \pi^2} & 0 \\ 0 & \frac{\partial^2 l_1}{\partial \varphi \partial \varphi'} \end{bmatrix}_{\hat{\pi}_{PLE}, \hat{\varphi}_{PLE}}$$

Variance calculation for MPLE is similar to that of PLE, except that one needs to replace  $\hat{\Omega}_{22}$  by an empirical estimate of the variance-covariance matrix of  $U_{1\varphi}$  (see CW).

For fixed  $\pi$ , one can solve  $U_{1\varphi} = 0$  for  $\varphi$  using an EM algorithm. If  $g_1$  were known, then standard algorithms could be used to maximize the “complete data” likelihood  $L_1 = \prod f(y_{1j}|g_1; \varphi)$  over  $\varphi$  (the “M-step”). The previous product is over families, and  $f(y_{1j}|g_1; \varphi)$  is the product over relatives of  $f(y_{1j}|g_{1j}; \varphi)$ . For the “E-step,” we need the expected value of  $g_{1i}$  given  $y_{1i}$  and  $g_{0i}$  [see McLachlan and Krishnan, 1997 for EM calculations for a mixture distribution]. We first need to calculate the joint conditional density for a particular family

$$h(g_1 | y_1, g_0) = \frac{f(y_1 | g_1) p(g_1 | g_0)}{\sum_g f(y_1 | g) p(g | g_0)} \tag{6}$$

Then the conditional expectation of the  $j$ 'th element of  $g_1$  is given by

$$E(g_{1j} | y_1, g_0) = h(g_{1j} | y_1, g_0) = \sum_{u_i \neq j} h(u_1, \dots, g_{1j}, \dots, u_m | y_1, g_0) \tag{7}$$

We iterate between the M- and E-steps to solve (4) for fixed  $\pi$ . To obtain the pseudo-likelihood estimates  $(\hat{\pi}, \hat{\varphi})$ , we iterate between the EM algorithm for solving (4) and a one-dimensional search of (5) for  $\pi$  for fixed  $\varphi$ , as discussed earlier.

**Dichotomous Outcomes**

If the phenotypes  $Y_0$  and  $Y_1$  are dichotomous with  $y = 1$  or  $0$  corresponding to the presence or absence of disease, then  $f(y | g = 0) = \varphi_0^y (1 - \varphi_0)^{1-y}$  and  $f(y | g = 1) = \varphi_1^y (1 - \varphi_1)^{1-y}$ .

The M-step of the EM algorithm yields the estimates

$$\hat{\phi}_0 = \frac{\sum_i \sum_j y_{1ij} (1 - \hat{g}_{1ij})}{\sum_i \sum_j (1 - \hat{g}_{1ij})} \tag{8}$$

and

$$\hat{\phi}_1 = \frac{\sum_i \sum_j y_{1ij} \hat{g}_{1ij}}{\sum_i \sum_j \hat{g}_{1ij}}. \tag{9}$$

where  $\hat{g}_{1ij} = E(g_{1ij} | y_{1ij}, g_{0i}; \hat{\pi}, \hat{\phi})$  as calculated in the E-step (Equation 7).

**Survival Outcomes**

Let  $t_{1ij}$ ,  $i = 1, \dots, I$ ,  $j = 1, \dots, m$ , denote the relatives' ages at the earlier of a disease event or censoring, and let  $\delta_{1ij}$  be the corresponding indicator of the disease event. We similarly define  $t_{0i}$ , and  $\delta_{0i}$ ,  $i = 1, \dots, I$ , for the probands. Thus, the phenotype  $y$  in Equation (1) corresponds to the pair  $(t, \delta)$ . We model the disease hazard for non-carriers and carriers as piecewise constant with a common set of cut-points  $v_0 = 0, v_1, v_2, \dots, v_k$ . These cut-points define the intervals  $[v_0, v_1), [v_1, v_2), \dots, [v_k, \infty)$ . Let  $\lambda_1^g, \lambda_2^g, \dots, \lambda_k^g$  be the corresponding hazards for carriers ( $g = 1$ ) and non-carriers ( $g = 0$ ). In the parametric case, the number and location of the cut-points are pre-specified (see Performance on Simulated Weibull Survival Data). Let  $Y_i(t)$  be the indicator function for whether or not individual  $i$  is at risk at age  $t$ . Then the probability that the person will not have been observed to have the event at or before age  $t$  is

$$\begin{aligned} S_i^g(t_i) &\equiv S(t_i; \lambda^0, \lambda^1, g_i) = \exp\left(-\int_0^{t_i} Y_i(t)[g_i \lambda^1(t) + (1 - g_i) \lambda^0(t)] dt\right) \\ &= \exp\left(-\sum_{l=1}^k \int_{v_{l-1}}^{v_l} Y_i(t)[g_i \lambda_l^1 + (1 - g_i) \lambda_l^0] dt\right) \\ &= \exp\left(-\sum_{l=1}^k [g_i \lambda_l^1 + (1 - g_i) \lambda_l^0] V_{il}\right), \end{aligned} \tag{10}$$

where  $V_{il} = \int_{v_{l-1}}^{v_l} Y_i(t) dt$  denotes the number of person years that individual  $i$  spends in the  $l$ 'th interval. We assume any censoring is independent of genotype. Under this assumption, one can replace  $f(y_{1j} | g_{1j}; \varphi)$  in equation (3) by  $\lambda^{g_{1j}}(t_{1j})^{\delta_{1j}} S^{g_{1j}}(t_{1j})$ . Likewise we can replace  $f(y_0 | g_0; \varphi)$  in equation (2) by  $\lambda^{g_0}(t_0)^{\delta_0} S^{g_0}(t_0)$ , provided, in addition, survival following disease onset is independent of carrier status [see Gail et al., 1999b].

Define the following expressions for the  $l$ 'th interval in terms of person-years ( $PY_l^{1NC}$  and  $PY_l^{1C}$ ) and deaths per interval ( $D_l^{1NC}$  and  $D_l^{1C}$ ) for non-carriers ( $g_{ij} = 0$ ) and carriers ( $g_{ij} = 1$ ), respectively:

$$\begin{aligned}
PY_l^{1NC} &= \sum_i \sum_j (1 - g_{1ij}) V_{1ijl} \\
D_l^{1NC} &= \sum_i \sum_j \delta_{1ij} (1 - g_{1ij}) I[v_{l-1} < t_{1ij} \leq v_l] \\
PY_l^{1C} &= \sum_i \sum_j g_{1ij} V_{1ijl} \\
D_l^{1C} &= \sum_i \sum_j \delta_{1ij} g_{1ij} I[v_{l-1} < t_{1ij} \leq v_l]
\end{aligned} \tag{11}$$

where  $V_{1ijl} = \int_{v_{l-1}}^{v_l} Y_{1ij}(t) dt$ . Then the M-step is defined by the complete data maximum likelihood estimates of the hazards (based only on the relatives), given by

$$\hat{\lambda}_{0l} = \frac{D_l^{1NC}}{PY_l^{1NC}} \tag{12}$$

$$\hat{\lambda}_{1l} = \frac{D_l^{1C}}{PY_l^{1C}} \tag{13}$$

using  $\hat{g}_{1ij}$ . For the E-step, we use the conditional expectations of the carrier status as defined by Equation (7).

We may obtain weakly parametric hazard estimates by increasing the number of age intervals. The properties of such estimates remain to be investigated, however, as the number of intervals increases, with increasing sample size.

## NUMERICAL RESULTS

### Performance on Dichotomous Outcomes

We studied analytically the performance of the full pseudo-likelihood (PLE) and maximum likelihood (MLE) methods with respect to bias, precision, efficiency, computation time, and stability of the estimates. We considered the effect of varying the proportion  $\rho$  of the probands with disease. We studied all control probands ( $\rho = 0$ ), random sampling of probands from the population ( $\rho = 0.105$  or  $\rho = 0.252$ , depending on the allele frequency in the population), case-control sampling with equal numbers of case and control probands ( $\rho = 0.5$ ), and all case probands ( $\rho = 1$ ). Each proband had only two informative relatives for breast cancer, a sister and a mother. The efficiencies and variances-per-family are presented in Table I for a rare mutant allele ( $\pi = 0.0066$ ) and for a more common mutant allele ( $\pi = 0.19$ ). We used the penetrance parameters  $\varphi_1 = 0.92$  and  $\varphi_0 = 0.1$  to correspond to the values of the penetrance of mutations of an autosomal dominant gene for breast cancer [Claus et al., 1991]. We evaluated the efficiency of PLE compared to MLE. The variances of MLE and PLE were computed exactly. The results are plotted in Figure 1.

For  $\pi = 0.0066$ , the efficiency of PLE is especially poor (8%) for  $\rho = 0.03$  and increases as  $\rho$  tends to unity. This makes sense because for small  $\rho$ , few probands will be carriers. Therefore, there will be very few families contributing to  $P(Y_1, Y_2 | g_0 = 1)$  (rela-

**TABLE I. Variances Per Family of MLE and PLE Parameter Estimates for a Dictotomous Outcome, and Efficiencies of PLE**

Sampling method	Carrier frequency ( $\pi$ )	Proportion of probands who are cases ( $\rho$ )	MLE variance			PLE variance (efficiency as %)		
			$\hat{\phi}_1$	$\hat{\phi}_0$	$\hat{\pi}$	$\hat{\phi}_1$	$\hat{\phi}_0$	$\hat{\pi}$
100% controls	0.0066	0.000	39.052	0.087	0.2186	59.858 (65%)	0.116 (75%)	0.3773 (58%)
Population	0.0066	0.105	9.676	0.046	0.0066	71.415 (14%)	0.047 (98%)	0.0067 (98%)
Case-control	0.0066	0.500	8.912	0.046	0.0020	17.210 (52%)	0.046 (99%)	0.0024 (84%)
100% cases	0.0066	1.000	8.731	0.047	0.0014	8.744 (100%)	0.047 (100%)	0.0014 (100%)
100% controls	0.19	0.000	0.526	0.221	0.8752	0.548 (96%)	0.240 (92%)	0.9974 (88%)
Population	0.19	0.252	0.388	0.066	0.1972	1.558 (25%)	0.101 (66%)	0.2110 (94%)
Case-control	0.19	0.500	0.433	0.060	0.1728	1.314 (33%)	0.081 (74%)	0.2274 (76%)
100% cases	0.19	1.000	1.044	0.114	0.4337	1.091 (96%)	0.120 (95%)	0.4691 (92%)

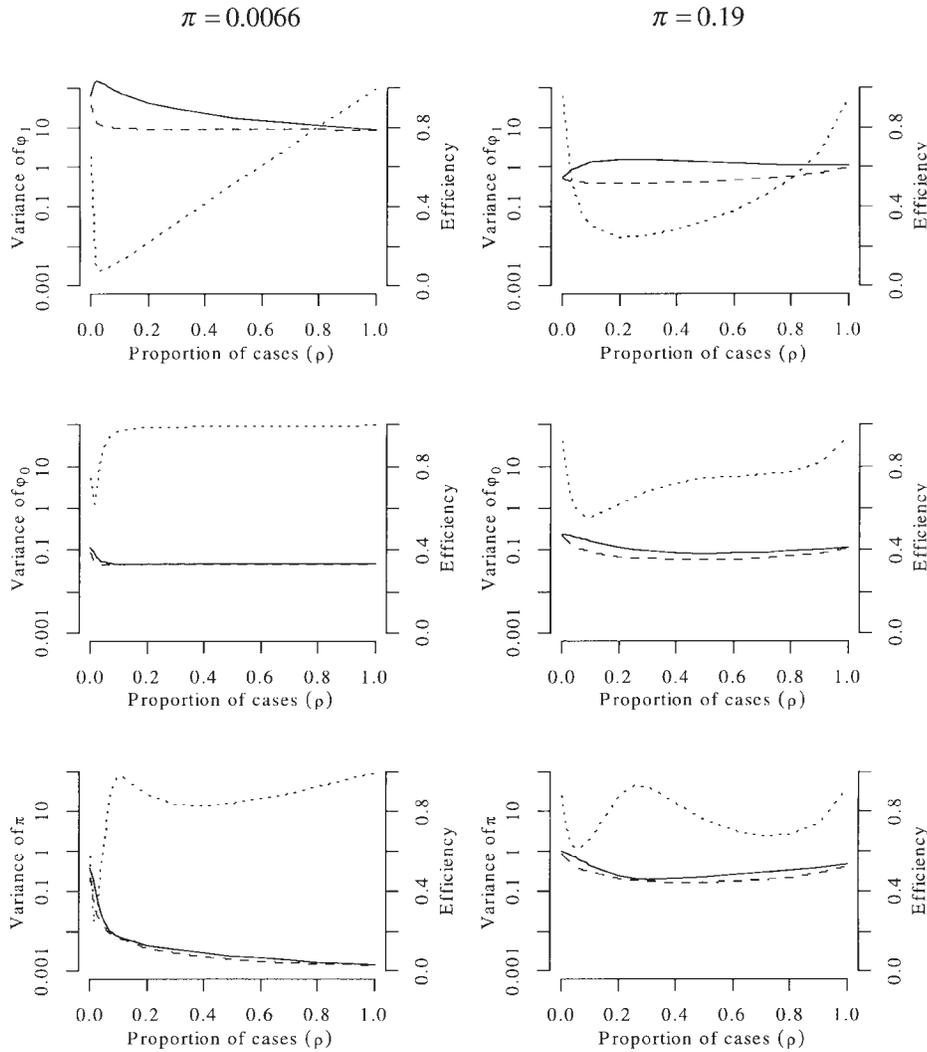


Fig. 1. Plots of the variances-per-family of the PL and ML estimates of  $\varphi_1$ ,  $\varphi_0$ , and  $\pi$ , and of their efficiencies. The variances (per family) of the pseudo- and full-likelihood parameter estimates for the dichotomous outcome model are the solid and dashed lines, respectively, and the values are plotted on a log scale on the left vertical axes. The efficiencies are the dotted lines, and the values are plotted on the right vertical axes. The proportion of mutation carriers in the population is denoted by  $\pi$  and the proportion of probands who are cases by  $\rho$ . The three plots on the left are for a carrier frequency  $\pi = 0.0066$ , and the three on the right are for  $\pi = 0.19$ .

tives of carriers), and most families will contribute only to  $P(Y_1, Y_2 | g_0 = 0)$  (relatives of non-carriers). These latter families tell us very little about  $\varphi_1$ . While MLE can recapture some of this information from  $P(g_0 | Y_0 = 0)$ , PLE cannot since PLE estimation of  $\varphi_1$  is based only on the information from the relatives. For a more common allele ( $\pi = 0.19$ ), the variances are one or two orders of magnitude smaller and more stable across the

range of  $\rho$ . The efficiency of PLE is usually higher for  $\pi = 0.19$  as compared to a rare allele ( $\pi = 0.0066$ ), but otherwise follows a similar pattern.

For  $\pi = 0.0066$ ,  $\text{var}(\hat{\varphi}_0)$  is much smaller and much more stable than  $\text{var}(\hat{\varphi}_1)$  over the full range of  $\rho$ . This may be because the dominant contribution to estimating  $\varphi_0$  is from  $P(Y_1, Y_2 | g_0 = 0)$ , and there are always many probands with  $g_0 = 0$ , even if all probands are cases. The efficiency of PLE is never worse than 75% with  $\pi = 0.0066$ . With  $\pi = 0.19$ ,  $\text{var}(\hat{\varphi}_0)$  increases slightly, compared to  $\pi = 0.0066$ , and the efficiency of PLE is never worse than 66%.

For  $\pi = 0.0066$ ,  $\text{var}(\hat{\pi})$  decreases as  $\rho$  approaches 1 for both MLE and PLE, and the efficiency of PLE is at least 85% for values of  $\rho$  above 0.105. For a more common allele ( $\pi = 0.19$ ),  $\text{var}(\hat{\pi})$  is higher than for  $\pi = 0.0066$  for both PLE and MLE, and the efficiency of PLE is greater than 76% for all values of  $\rho$ .

### Performance on Simulated Weibull Survival Data

We assumed that the time to disease onset (“survival time”) for mutation carriers followed a Weibull distribution with shape parameter 2.1334 and scale parameter 0.0130, whereas non-carriers followed a Weibull distribution with shape parameter 3.2893 and scale parameter 0.0078. The parameters of the Weibull distributions were chosen to match the cumulative risk of breast cancer reported by Struewing et al. [1997] at ages 50 and 70, namely 0.33 and 0.56, respectively, for carriers and 0.045 and 0.13, respectively, for non-carriers. We chose a carrier frequency of  $\pi = 0.0243$  to match that estimated by Carroll et al. [2000] from the Washington Ashkenazi data set. As for the dichotomous case, we assumed each proband provided data on a mother and a sister. We fit separate models with piecewise-constant hazards on seven age intervals, [0,30), [30,40), [40,50), [50,60), [60,70), [70,80), [80,∞), for mutation carriers and non-carriers. We refer to this as a six-interval model because we are only interested in estimates to age 80. We used simulations to estimate the precision of our estimated survival parameters, and we estimated the efficiencies of PLE and MPLE as the ratio of sample variances of MLE to those of PLE and MPLE, respectively, obtained from 100 independent simulations. In each simulation, we generated data from 5,000 families. We assumed that probands were sampled at random from the population with the proportion of families having carrier probands set to  $\pi = 0.0243$  as discussed above. Based on this value of  $\pi$ , the trinomial distribution with support of eight points (the proband, mother, and sister can each be a carrier or a non-carrier) was produced. For each pedigree, we used this distribution to randomly determine with which trinomial point the pedigree was associated. To determine the proband’s age we generated a normal ( $\mu = 51.96$ ,  $\sigma = 14.12$ ) random variable. Given the proband’s age, the sister’s age was obtained by adding the proband’s age to a random variable generated from a normal ( $\mu = -0.47$ ,  $\sigma = 6.433$ ) distribution. Similarly, the mother’s age was obtained by adding the proband’s age to a random variable generated from a normal ( $\mu = 28.39$ ,  $\sigma = 5.34$ ) distribution. The parameters of these three distributions matched the means and standard deviations observed in the Washington Ashkenazi data. A woman was a case if her simulated age at breast cancer onset was less than or equal to her simulated age at the time of study. All of the age information was finally rounded to the nearest year. All computations were carried out using the Gauss System [Aptech Systems, 1999].

We found no evidence of bias with any of the methods, since the average hazard

estimates from MLE, PLE, and MPLE agreed well with the Weibull average hazard in each interval (Table II). The estimated efficiency of PLE for hazard estimation ranges from 54 to 91%, and of MPLE from 51 to 89% (Table II). The survival and hazard estimates for carriers are plotted in Figure 2 for MLE, PLE, and MPLE, and 95% confidence intervals are also shown. As expected, the MLE confidence intervals are somewhat narrower than both of the pseudo-likelihood confidence intervals. Somewhat surprisingly, the widths of the confidence intervals based on PLE and MPLE are similar. Weibull data for non-carriers were also simulated. MLE, PLE, and MPLE produced unbiased estimates of the average hazards for non-carriers.

Using 13 age intervals,  $[0,30)$ ,  $[30,35)$ , . . . ,  $[30,35)$ , . . . ,  $[80,85)$ , . . . ,  $[85,\infty)$ , PLE and MPLE yielded excellent agreement with the carrier survival distribution (data not shown). We refer to this as a 12-interval model because we are studying results up to age 85. We were unable to obtain MLE hazard estimates with this model because the constrained maximization algorithm, called CML in Gauss [Aptech Systems, 1999], failed to converge, even when we started the iterations at the PLE estimates. This sophisticated maximization algorithm uses a Newton-Raphson approach with a sequence of step sizes. If no steps yield an increase in the objective function, numerous random directions are explored to start a new Newton-Raphson search. When we allowed each event time to define the upper limit of a hazard interval (the “non-parametric” approach), both PLE and MPLE yielded survival estimates that were close to the correct model (solid line in Fig. 3, top).

Both PLE and MLE require considerable computation, but MLE requires even more calculation than PLE. MLE required 6–8 times longer to fit a six-interval model than PLE. MLE failed to converge with the twelve-interval model, while PLE took 2,170 cpu seconds in a Pentium Pro 200 MHz computer. The non-parametric PLE model required approximately 50,000 cpu seconds to fit one data set, and again MLE failed to converge. Thus, PLE is stable computationally even for fully nonparametric survival estimation, but the time required for convergence can be quite long. By contrast, our algorithm for MLE converged for a six-interval model, but failed to converge for large numbers of intervals. MPLE is much faster than PLE.

### Re-Analysis of Data From the Washington Ashkenazi Study

We used the PLE and MPLE procedures on a subset of 1960 family sets from the data used by Struewing et al. [1997]. This subset consisted of all family sets where one relative was a mother and where there was at least one sister of the proband. For sets where the proband had several sisters, one sister was selected at random, so that every family set contained exactly two relatives. (The original data set consisted of 4,873 family sets with up to seven first-degree relatives in a set.) For consistency with the paper by Struewing et al., we presented cumulative distribution functions (CDF's) instead of survival functions. We obtained three CDF estimates: one each for PLE and MPLE, and for comparison, a CDF calculated using the methods described in Struewing et al. [1997] and Wacholder et al. [1998] (Fig. 4). Note that the Wacholder-Struewing estimate is not monotone, but nevertheless provides an estimated CDF that is in close agreement with both of the pseudo-likelihood estimates. We used bootstrap re-sampling of families to estimate standard deviations of the Wacholder-Struewing CDF estimates (1,000 bootstrap samples) and of the pseudo-likelihood CDF estimates (155 bootstrap samples). Pointwise 95% confidence intervals were taken as the point estimate  $\pm 1.96$  times the

**TABLE II. Comparison of MLE, PLE, and MPLE Piecewise-Constant Hazard Estimates With the Corresponding Average Weibull Hazard for Mutation Carriers\***

Age interval (years)	Average Weibull hazard in interval <sup>a</sup>	Weibull survival probability at end of interval	MLE		PLE		MPLE		Efficiency (%)	
			Estimated hazard (s.e.) <sup>b</sup>	Estimated survival (s.e.) <sup>b</sup>	Estimated hazard (s.e.) <sup>b</sup>	Estimated survival (s.e.) <sup>b</sup>	Estimated hazard (s.e.) <sup>b</sup>	Estimated survival (s.e.) <sup>b</sup>	PLE hazard <sup>c</sup>	MPLE hazard <sup>c</sup>
0–30	0.00449	0.8740	0.00433 (0.00079)	0.8783 (0.0208)	0.00410 (0.00098)	0.8846 (0.0259)	0.00407 (0.00099)	0.8855 (0.0261)	65	64
31–40	0.01141	0.7797	0.01165 (0.00256)	0.7821 (0.0296)	0.01189 (0.00345)	0.7860 (0.0352)	0.01184 (0.00358)	0.7871 (0.0360)	55	51
41–50	0.01517	0.6700	0.01539 (0.00375)	0.6710 (0.0367)	0.01534 (0.00509)	0.6752 (0.0468)	0.01516 (0.00505)	0.6772 (0.0464)	54	55
51–60	0.01904	0.5538	0.01884 (0.00556)	0.5567 (0.0434)	0.01847 (0.00736)	0.5630 (0.0585)	0.01818 (0.00726)	0.5663 (0.0584)	57	59
61–70	0.02301	0.4400	0.02293 (0.00918)	0.4442 (0.0508)	0.02182 (0.00964)	0.4546 (0.0636)	0.02183 (0.00989)	0.4576 (0.0664)	91	86
71–80	0.02706	0.3357	0.02734 (0.01701)	0.3435 (0.0701)	0.02560 (0.01778)	0.3587 (0.0838)	0.02577 (0.01802)	0.3605 (0.0852)	91	89

\*The Weibull models for carriers and non-carriers are described in Performance on Simulated Weibull Survival Data. The population mutation carrier frequency was  $\pi = 0.0243$ . The average ML estimate of  $\pi$  was 0.0244 with standard error 0.00212, the average PL estimate of  $\pi$  was 0.0245 with standard error 0.00212, and the average marginal estimate of  $\pi$  was 0.0245 with standard error 0.00212.

<sup>a</sup>The average hazard is the integrated hazard divided by the interval width.

<sup>b</sup>Standard errors of the mean hazard rates were estimated empirically from the 100 simulations.

<sup>c</sup>Efficiency is estimated as the squared ratio of standard errors of MLE to PLE and MPLE.

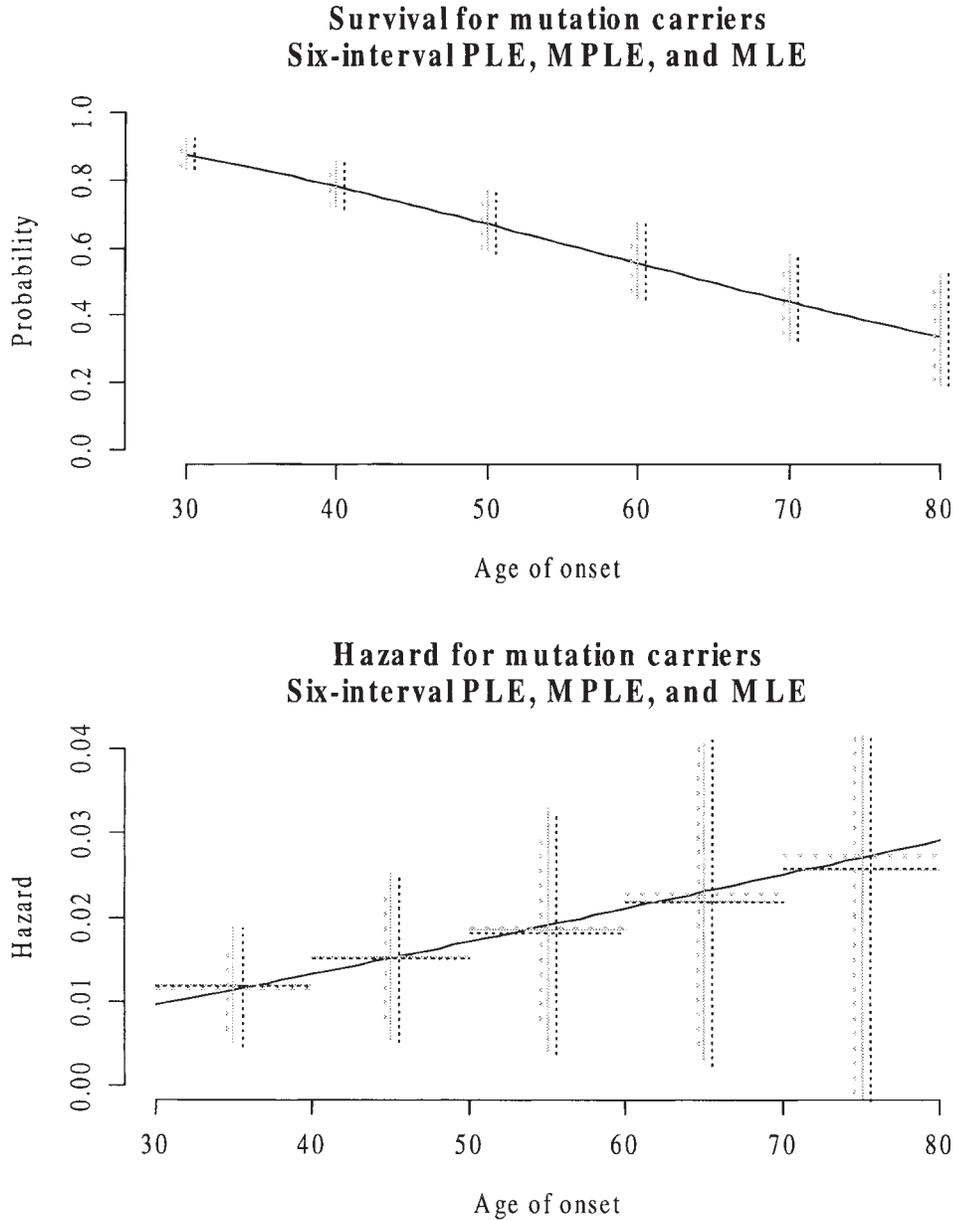


Fig. 2. Average survival and hazard estimates for mutation carriers for the six-interval constant hazard model. Data were generated from Weibull models for mutation carriers (see Performance on Simulated Weibull Survival Data) with carrier frequency  $\pi = 0.0243$ . There were 100 independent simulations, each based on 5,000 families. In these figures, the Weibull survival and hazard functions are plotted in solid black, and the means of the 100 PL, MPL, and ML survival and hazard estimates and 95% empirical confidence intervals are plotted in solid gray, dotted black, and dashed gray, respectively. **Top:** Only the Weibull survival distribution is shown; the PL, MPL, and ML survival estimates are not shown since they are indistinguishable from the Weibull at this plot resolution. **Bottom:** All three hazard estimates are shown. The mean MLE, PLE, and MPLE estimates of  $\pi$  are 0.0244, 0.0245, and 0.0245, respectively.

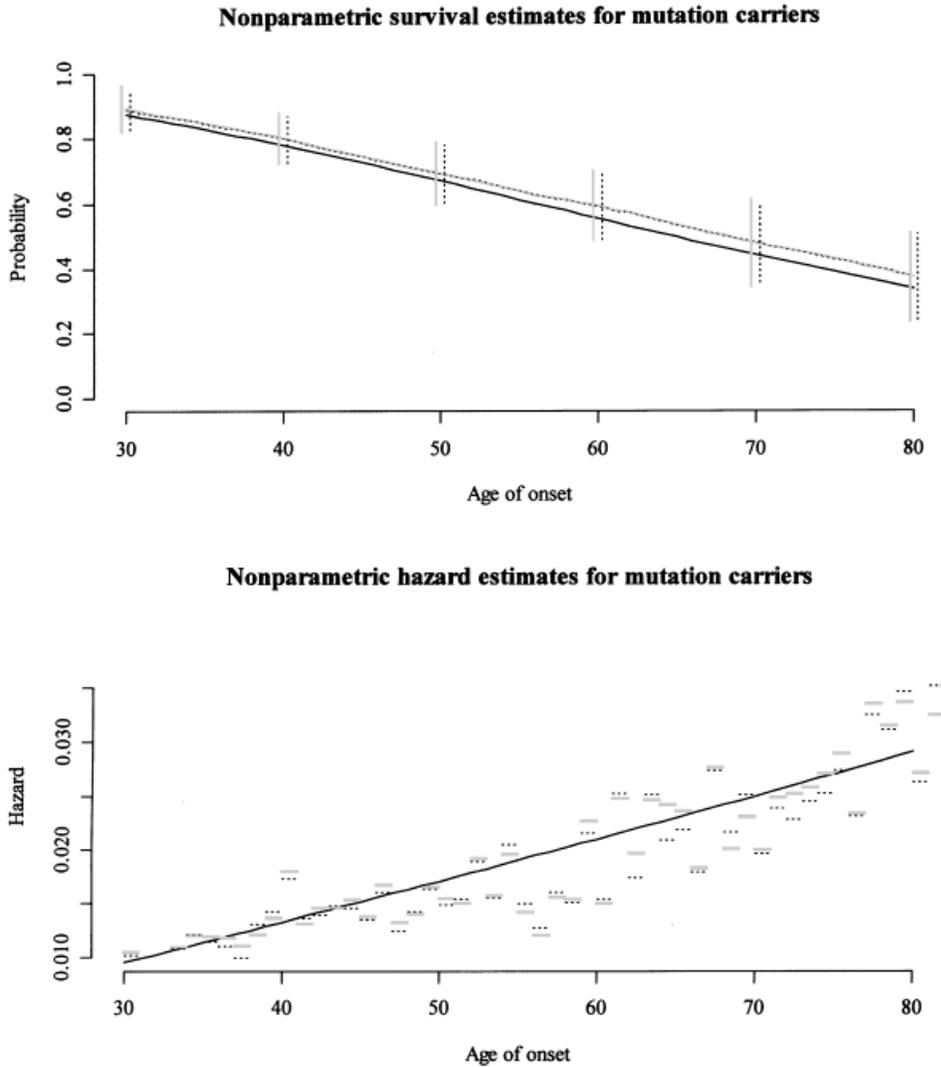


Fig. 3. Average survival and hazard estimates vs. age from the nonparametric PLE and MPLE models for carriers of mutations, from a simulation of Weibull data with a carrier frequency of 0.023. The data for each simulation were generated from a Weibull distribution as described in Performance on Simulated Weibull Survival Data. There were 5,000 family sets and 50 independent simulations. **Top:** Weibull survival distribution is solid black, the PL estimate is solid gray, and the MPL estimate is dotted black. **Bottom:** Solid black line is the Weibull hazard, and the PL and MPL hazard estimates are indicated by the same line types as at the top. The estimates and standard errors of the carrier frequency are  $\hat{\pi}_{PLE} = 0.0230 \pm 0.0022$  and  $\hat{\pi}_{MPLE} = 0.0229 \pm 0.0021$ .

estimated standard deviation (Fig. 4). The PLE and MPLE confidence intervals are both narrower than those of the Wacholder-Stuewing procedure. The improved precision of PLE and MPLE may reflect the fact that PLE and MPLE implicitly impose monotonicity constraints on the cumulative risk.

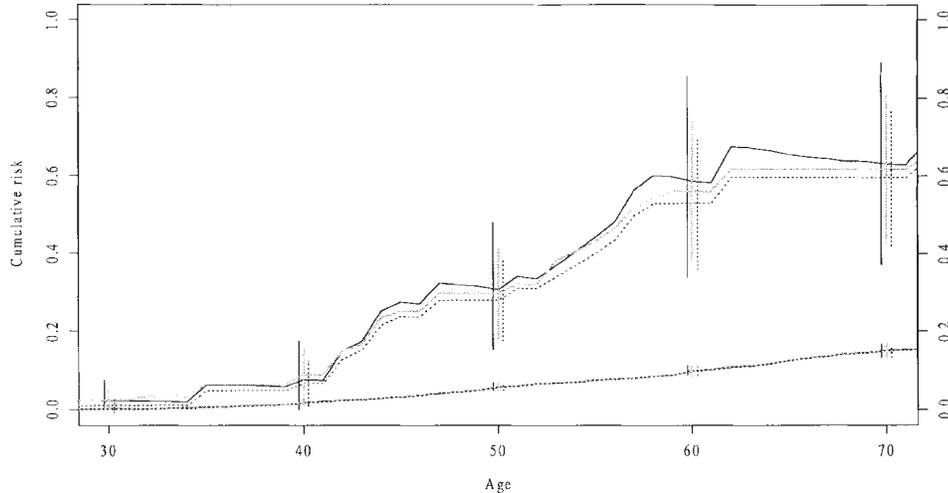


Fig. 4. This graph represents the Wacholder-Stuewing estimate (solid black), the non-parametric PLE estimate (solid gray), and the non-parametric MPLE estimate (dotted) of the cumulative distribution functions and pointwise 95% confidence intervals at 10-year intervals. The top lines are for mutation carriers. Note the non-monotonicity of the Wacholder-Stuewing estimates. The bottom lines are for non-carriers, which are virtually superimposable. The Wacholder-Stuewing estimate of the carrier frequency is  $\hat{\pi}_{ws} = 0.0228$ , the PLE estimate is  $\hat{\pi}_{PLE} = 0.0213$ , and the MPLE estimate is  $\hat{\pi}_{MPLE} = 0.0227$ . The confidence intervals for the Wacholder-Stuewing estimates are based on bootstrap samples of size 1,000, while those for the pseudo-likelihood and marginal estimates are based on bootstrap samples of size 155.

## DISCUSSION

We have used three approaches to obtain monotone estimates of cumulative risk (or survival) functions for carriers and non-carriers of an autosomal dominant disease gene from a kin-cohort design. The likelihood factors as (1) the probability of the genotype of the proband given the proband's phenotype and (2) the probability of the relatives' phenotypes, given the genotype of the proband. Maximum likelihood estimation of the carrier frequency and the survival curves based on Newton-Raphson procedures works for models with a modest number of parameters, as discussed by Gail et al. [1999b], but becomes numerically unstable and very time-consuming as the number of parameters increases. A second procedure, PLE, maximizes the two components of the likelihood separately, using the first part to estimate the carrier frequency and the latter part to estimate cumulative risk functions. PLE is computationally stable but time-consuming, and it can be considerably less efficient than MLE for some values of the parameters (Tables I and II, Fig. 1). Because of its greater efficiency, we recommend MLE whenever it can be computed. However, further work is needed to define a reliable algorithm to compute MLE for nonparametric or weakly-parametric survival estimates. A third procedure, MPLE, originally proposed by CW, treats the second factor in the likelihood as if each relative-proband pair came from an independent family. This procedure is very fast, because there is no need to sum over combinations of genotypes involving more than two family members. MPLE is only slightly less efficient than PLE, and MPLE has the

additional advantage that it is more robust to failure of the assumptions of conditional independence of family phenotypes given genotypes than are MLE or PLE (see CW). In particular, if probands are selected at random from the general population, MPLE yields valid estimates of the marginal cumulative risk functions even if there are sources of familial aggregation unrelated to the gene under study.

MLE, PLE, and MPLE all yield more precise estimates of cumulative risk than the procedure used in Struewing et al. [1997], probably because the implicit monotonicity constraints reduce random variability of the estimates.

Several additional issues may be important in particular applications. First, we have assumed that our sample of probands is representative, conditional on phenotype. Biases can result if the willingness of a potential proband to participate depends on the phenotypes of his or her relatives [Struewing et al., 1997; Wacholder et al., 1998; Gail et al., 1999b]. Second, there may be a need to introduce covariates that modify cumulative risks. Third, a proband is available for study only if he or she survives competing causes of mortality to the date of study, and the likelihood could be affected if survival probabilities differ between gene carriers and non-carriers [Gail et al., 1999a]. These biases can be minimized only through careful attention to the methods used for sampling probands in kin-cohort studies.

A computer program in GAUSS is available from David Pee or Dirk Moore. This program carries out the required calculations for a pedigree consisting of a proband, her sister, and her mother. This program would require non-trivial modifications by the user for more general applications.

## ACKNOWLEDGMENTS

D. Moore's research was partially supported by a Temple University study leave program. The authors are grateful to J. Struewing for permission to use a subset of the Washington Ashkenazi Study data.

## REFERENCES

- Aptech Systems, Inc. 1999. The Gauss System, Version 3.2. Maple Valley, Washington.
- Carroll RJ, Gail MH, Benichou J, Pee D. 1999. Score tests for familial correlation in genotyped-proband designs. *Genet Epidemiol* 18:293–306.
- Chatterjee N, Wacholder S. 2001. A marginal likelihood approach for estimating penetrance from a kin-cohort design. *Biometrics* (in press).
- Claus EB, Risch NJ, Thompson WD. 1991. Genetic analysis of breast cancer in the Cancer and Steroid Hormone Study. *Am J Hum Genet* 48:232–42.
- Gail MH, Pee D, Benichou J, Carroll R. 1999a. Designing studies to estimate the penetrance of an identified autosomal dominant mutation: cohort, case-control and genotyped-proband designs. *Genet Epidemiol* 16:15–39.
- Gail MH, Pee D, Carroll R. 1999b. Kin-cohort designs for gene characterization. *J Nat Cancer Inst Monogr* 26:55–60.
- Godambe VP. 1991. Estimating equations. Oxford: Clarendon Press.
- Gong G, Samaniego FJ. 1981. Pseudo maximum likelihood estimation: theory and applications. *Ann Stat* 9:861–9.
- Li CC. 1976. First course in population genetics. Pacific Grove, CA: Boxwood Press.
- McLachlan GJ, Krishnan T. 1997. The EM algorithm and extensions. New York: Wiley.
- Struewing JP, Hartge P, Wacholder S, Baker SM, Berlin M, McAdams M, Timmerman MM, Bordy LC, Tucker MA. 1997. The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *New Eng J Med* 336:1401–8.

Wacholder S, Hartge P, Struwing JP, Pee D, McAdams M, Brody L, Tucker M. 1998. The kin-cohort study for estimating penetrance. *Am J Epidemiol* 148:623–9.

## APPENDIX

To compute the variance of  $U_0$ , consider a family with given  $y_0$ . Because  $f_0(g_0|y_0; \pi, \varphi)$  is a conditional density and the support of  $g_0$  is independent of  $\pi$ , standard arguments show

$$E\left(\frac{\partial}{\partial \pi} \ln f_0\right) = 0$$

and

$$E\left(\frac{\partial}{\partial \pi} \ln f_0\right)^2 = -E\left(\frac{\partial^2}{\partial \pi^2} \ln f_0\right),$$

where the expectation is over  $g_0$  given  $y_0$ . Because

$$U_{0\pi} = \sum_{i=1}^l \frac{\partial}{\partial \pi} \ln \{f_0(g_{0i} | y_{0i}; \varphi, \pi)\},$$

$E(U_{0\pi}) = 0$ . The variance of  $U_{0\pi}$ , conditional on the observed set  $\{y_{01}, y_{02}, \dots, y_{0l}\}$ , is

$$\text{var}(U_{0\pi} | y_{01}, y_{02}, \dots, y_{0l}) = \sum_{i=1}^l -E\left(\frac{\partial^2}{\partial \pi^2} \ln \{f_0(g_{0i} | y_{0i}; \varphi, \pi)\}\right), \quad (14)$$

where the expectations in the summand are over  $g_{0i}$  given  $y_{0i}$ . Thus, the “observed” variance of  $U_{0\pi}$ , analogous to the observed Fisher information, is simply

$$-\frac{\partial^2 l_0}{\partial \pi^2},$$

which can be obtained by numerical differentiation of  $f_0$ . This quantity approximates the unconditional variance of  $U_{0\pi}$  in repeated samples of families.

To compute the variance of  $U_{1\varphi}$ , consider a family with given  $g_0$ . Because the conditional density  $f_1(y_1 | g_0; \pi, \varphi)$  has support independent of  $\varphi$ , the previous argument shows that

$$E\left(\frac{\partial}{\partial \varphi} \ln f_{1\varphi}\right) = 0,$$

where the expectation is over  $y_1$  given  $g_0$ . Because

$$U_{1\varphi} = \sum_{i=1}^l \frac{\partial}{\partial \varphi} \ln f_1(y_{1i} | g_{0i}; \varphi, \pi),$$

$E(U_{1\varphi}) = 0$ . The conditional variance of  $U_{1\varphi}$  given the set of values of  $g_0$  is

$$\text{var}(U_{1\varphi} | \underline{g}_0) = \sum_{i=1}^I -E_{Y_i} \left( \frac{\partial^2 \ln f_i(y_{1i} | g_{0i}; \underline{\varphi}, \underline{\pi})}{\partial \underline{\varphi} \partial \underline{\varphi}'} \right), \quad (15)$$

where the expectations are over  $y_{1i}$  given  $g_{0i}$ . As before, the ‘‘observed’’ variance of  $U_{1\varphi}$ , which approximates its unconditional variance, is

$$-\frac{\partial^2 l_1}{\partial \varphi^2}.$$

To show that  $\text{cov}(U_{0\pi}, U_{1\varphi}) = 0$ , consider the covariance between

$$U_{0\pi i} = \frac{\partial}{\partial \pi} \ln f_0(g_{0i} | y_{0i}; \underline{\varphi}, \underline{\pi})$$

and

$$U_{1\varphi i} = \frac{\partial}{\partial \varphi} \ln f_1(y_{1i} | g_{0i}; \underline{\varphi}, \underline{\pi})$$

for family  $i$ . Conditional on  $y_{0i}$ ,  $E(U_{0\pi i}) = 0$ , and conditional  $g_{0i}|y_{0i}$  on  $E(U_{0\varphi i}) = 0$ . Therefore

$$\begin{aligned} \text{cov}(U_{0\pi i}, U_{1\varphi i}) &= E_{y_{0i}} \{ \text{cov}(U_{0\pi i}, U_{1\varphi i} | y_{0i}) \} \\ &= E_{y_{0i}} E_{g_{0i}|y_{0i}} \{ \text{cov}(U_{0\pi i}, U_{1\varphi i} | y_{0i}, g_{0i}) \} = 0 \end{aligned}$$

because  $U_{0\pi i}$  is constant, given  $g_{0i}$  and  $y_{0i}$ .

$$\text{Letting } \Omega = \begin{pmatrix} \text{var } U_{0\pi} & 0 \\ 0 & \text{var } U_{1\varphi} \end{pmatrix}, \text{ and}$$

$$B = \begin{pmatrix} \frac{\partial U_{0\pi}}{\partial \pi} & \frac{\partial U_{0\pi}}{\partial \varphi} \\ \frac{\partial U_{1\varphi}}{\partial \pi} & \frac{\partial U_{1\varphi}}{\partial \varphi} \end{pmatrix},$$

we use a Taylor series expansion to estimate

$$\text{cov}(\hat{\pi}, \hat{\varphi}) \approx V_{pl} = B^{-1} \Omega (B^{-1})'. \quad (16)$$

Estimates of  $B$  may be obtained by numerical differentiation at the pseudo-likelihood estimate  $(\hat{\pi}, \hat{\varphi})$ ; the ‘‘observed’’ estimate of  $\Omega$  is also obtained by numerical differentiation at  $(\hat{\pi}, \hat{\varphi})$ .