

A Marginal Likelihood Approach for Estimating Penetrance from Kin-Cohort Designs

Nilanjan Chatterjee* and Sholom Wacholder**

Division of Cancer Epidemiology and Genetics, National Cancer Institute,
6120 Executive Boulevard, Rockville, Maryland 20892, U.S.A.

*email: chattern@mail.nih.gov

**email: wacholder@nih.gov

SUMMARY. The kin-cohort design is a promising alternative to traditional cohort or case-control designs for estimating penetrance of an identified rare autosomal mutation. In this design, a suitably selected sample of participants provides genotype and detailed family history information on the disease of interest. To estimate penetrance of the mutation, we consider a marginal likelihood approach that is computationally simple to implement, more flexible than the original analytic approach proposed by Wacholder et al. (1998, *American Journal of Epidemiology* 148, 623-629), and more robust than the likelihood approach considered by Gail et al. (1999, *Genetic Epidemiology* 16, 15-39) to presence of residual familial correlation. We study the trade-off between robustness and efficiency using simulation experiments. The method is illustrated by analysis of the data from the Washington Ashkenazi Study.

KEY WORDS: Correlated data; EM algorithm; Failure time data; Residual familial correlation; Sandwich variance.

1. Introduction

A disease is called fully penetrant with respect to a locus if the genotype of an individual at that locus completely determines his/her disease. Development of complex diseases, however, typically involves a combination of various genetic and environmental risk exposures. Therefore, variation in disease expression is expected even among individuals with the same genotype at the given locus due to their differences in background with respect to the other risk factors. Such uncertainty for a complex disease can be expressed in terms of its penetrance, the probability of the disease in those with the at-risk genotypes at a given disease locus.

Once a mutation in a specific locus has been identified as a risk factor for a disease, investigators interested in public health and genetic counseling need population-based estimates of the penetrance associated with carrying this mutation. Studies for identifying a disease gene, such as linkage studies, typically collect data on families with large numbers of affected individuals to enhance the power of detecting the gene. Estimates of penetrance, which do not properly account for ascertainment by highly affected families, clearly overestimate the risk in the general population due to overrepresentation of the diseased individuals in the sample. Even after accounting for ascertainment, which often is a complex task itself, risk estimates can be too high to be representative of the general population since there may exist unobserved genetic or environmental exposures in these families that had enhanced the risk of the disease from the mutation (Struewing et al., 1997; Wacholder et al., 1998). Using breast cancer-prone

families, e.g., Easton, Ford, and Bishop (1995), after accounting for ascertainment, obtained the estimate of the cumulative risk of breast cancer to age 70 among BRCA1 mutation carriers to be as high as 85%. From a more population-based design, Struewing et al. (1997) estimated this risk from BRCA1 and BRCA2 mutations to be only 56%. In this study, 5318 Ashkenazi Jewish volunteers living in the Washington, D.C. area were genotyped for three specific mutations in BRCA1 and BRCA2 genes and were interviewed for detailed personal and family history of cancer in breast and some other organs. Wacholder et al. (1998), who formally proposed this as a kin-cohort design, treated the first-degree relatives of the participants as a retrospective cohort and developed a simple analytic approach to estimate the age-specific cumulative risk of the carriers and noncarriers of the mutation using the disease history data of these relatives and the genotype data of the volunteers. More details about this design, together with its strengths and weaknesses over traditional cohort and case-control studies, can be found in Wacholder et al. (1998) and Gail, Pee, and Carroll (1999b).

Gail et al. (1999a), who called it the genotyped-proband design, considered the likelihood of the data under the conditional independent assumption that the phenotypes (disease trait) of the members of a family are independently distributed given their genotypes. An advantage of the likelihood approach over the method of moments approach of Wacholder et al. (1998) (see Section 2.3) is that it can incorporate various types of information into the analysis, such as known parametric form of the cumulative risk, covariates, disease history for

second or higher degree relatives, and the disease status of the participants themselves. Also, the problem of nondecreasing cumulative risk estimates obtained from the simple approach of Wacholder et al. (1998) can be avoided by considering a likelihood approach. The assumption underlying this likelihood, however, may be violated in practice due to presence of residual familial correlation arising from other shared genetic or environmental factors. Gail, Pee, and Carroll (1999b) found that penetrance estimates from their likelihood can have large bias in the presence of such correlation.

In this article, we propose a new method of estimation based on a marginal likelihood. The method enjoys the advantages of the likelihood approach described above and yet, compared with the likelihood approach of Gail et al. (1999a) and a pseudolikelihood approach of Moore et al. (2000), the method is computationally simpler and faster. Moreover, the marginal method is robust to the violation of conditional independence assumptions needed for the other proposed likelihood-based methods. Specifically, it can be shown that, if the participants can be regarded as a random sample from an underlying population, the marginal likelihood estimate produces consistent estimates of penetrance irrespective of the presence of residual familial correlation. When the assumption of no residual familial correlation holds, however, the marginal approach may be less efficient. We study the trade-off between bias and efficiency by using simulation experiments. We reanalyze data from the Washington Ashkenazi Study using the marginal likelihood approach to obtain a monotone estimate of the age-specific cumulative risk of breast cancer associated with carrying BRCA1/BRCA2 mutations.

2. Methods

2.1 Notation and Assumptions

Suppose K participants, sampled from an underlying target population using an appropriate sampling design, provide DNA samples and detailed personal and family history information of the disease. We assume that the locus of interest is autosomal, i.e., is not located on a sex chromosome, and each individual can inherit either of the two alleles, i.e., A , the mutated type, or a , the common (wild) type, from each of his/her parents. Thus, the genotype of an individual can be any of the three possible combinations aa , Aa , and AA . Here we assume the disease is inherited in a dominant fashion, i.e., the disease risk is the same whether one is heterozygous (Aa) or homozygous (AA) with respect to the mutant allele, but the techniques developed in this article can be easily extended for recessive inheritance or to situations where the mode of inheritance is unknown by estimating risk specific to the three distinct genotypes.

Let g_i^P denote the indicator of whether the i th of the K participants is a carrier (AA or Aa) or not (aa). The i th participant gives the family history information of the disease of interest about his/her n_i relatives. Let $y_i^R = (y_{i1}^R, \dots, y_{in_i}^R)$ denote the family history information of the i th participant and $\mathbf{g}_i^R = (g_{i1}^R, \dots, g_{in_i}^R)$ denote the genotypes of the relatives, which are not observed. Let (y_1^P, \dots, y_K^P) denote disease history of the participants themselves. Let $q_0(y; \theta_0)$ and $q_1(y; \theta_1)$ be probability mass or density functions of Y , characterized by parameters θ_0 and θ_1 , describing the distribution of the disease among the population of non-

carriers and carriers, respectively. We assume that, conditional on genotypes, the risk in the population of relatives is the same as that of the population from which the participants have been sampled. All subsequent calculations of the probability distribution of the genotypes are based on standard Mendelian genetics assumptions (Li, 1978), specifically no inbreeding, random mating, and Hardy-Weinberg equilibrium. The allele frequency of the mutation A will be denoted by f .

2.2 Likelihood with No Residual Familial Correlation

Here we briefly describe the likelihood approach of Gail et al. (1999a). Suppose we assume that, conditional on the genotypes of the members of a family, their phenotypes (Y) are independent, i.e.,

$$\begin{aligned} \text{pr} \left(y_{i1}^R, \dots, y_{in_i}^R, y_i^P \mid g_{i1}^R, \dots, g_{in_i}^R, g_i^P \right) \\ = \text{pr} \left(y_{i1}^R \mid g_{i1}^R \right) \cdots \text{pr} \left(y_{in_i}^R \mid g_{in_i}^R \right) \text{pr} \left(y_i^P \mid g_i^P \right). \end{aligned} \quad (1)$$

Under this assumption, the likelihood of the relatives' data conditional on the indexed participant's genotype can be written as

$$\begin{aligned} L^R = \prod_{i=1}^K \sum_{g_{i1}^R, \dots, g_{in_i}^R} q_{g_{i1}} \left(y_{i1}^R; \theta_{g_{i1}^R} \right) \cdots q_{g_{in_i}} \left(y_{in_i}^R; \theta_{g_{in_i}^R} \right) \\ \times \text{pr} \left(g_{i1}^R, \dots, g_{in_i}^R \mid g_i^P \right), \end{aligned} \quad (2)$$

where $\text{pr}(g_{i1}^R, \dots, g_{in_i}^R \mid g_i^P)$, the joint distribution of the genotypes of the family members given the participant's genotype, can be computed as a function of the allele frequency using the Mendelian mode of inheritance. The likelihood contribution of the participants, assuming they form a random sample, is given by

$$\begin{aligned} L^P = \prod_{i=1}^K \text{pr} \left(y_i^P \mid g_i^P \right) \text{pr} \left(g_i^P \right) \\ = \prod_{i=1}^K q_{g_i^P} \left(y_i^P; \theta_{g_i^P} \right) \text{pr} \left(g_i^P \right). \end{aligned} \quad (3)$$

To accommodate sampling of the participants conditional on their phenotypes, Gail et al. (1999a) computed the likelihood contribution of the participants as

$$L^P = \prod_{i=1}^K \text{pr} \left(g_i^P \mid y_i^P \right). \quad (4)$$

Depending on the underlying sampling mechanism of the participants, we will use (3) or (4) as the contribution of the participants in our subsequent calculations.

2.3 Marginal Likelihood with Random Sampling of Participants

The marginal likelihood is defined using a modification of L^R . Here we treat the n_i relatives of the i th participant individually, ignoring any relationship between the relatives of the participant. Thus, a family of $(n_i + 1)$ members with one participant and n_i relatives is broken into n_i pseudofamilies, each consisting of two members, i.e., the participant and a

relative. Given the i th participant's genotype, the conditional probability of his/her j th relative's phenotype is given by $\text{pr}(y_{ij}^R | g_i^P) = \sum_{g_{ij}^R} q_{g_{ij}^R}(y_{ij}^R; \theta_{g_{ij}^R}) \text{pr}(g_{ij}^R | g_i^P)$. The marginal likelihood of the relatives' data is defined by

$$L_M^R = \prod_{i=1}^K \prod_{j=1}^{n_i} \text{pr}(y_{ij}^R | g_i^P). \quad (5)$$

Note the difference in the use of the mode of inheritance in constructing the likelihoods L^R and L_M^R . Conditional on the participant's genotype, we compute the distribution of the genotypes of the relatives individually for L_M^R but jointly for L^R . This gives a computational advantage of L_M^R over L^R since computing the joint distribution for large families can be a cumbersome task.

Marginal likelihood approaches, together with robust estimators of variances, are commonly used in statistics in analysis of clustered data when parameters of the marginal distributions of the individuals are of interest but the correlation between individuals of the same cluster is considered a nuisance. As long as the marginal model is correctly specified, such an approach is known to produce consistent estimates of the parameters of the marginal model irrespective of the nature of intracluster correlation between individuals. From this well-known fact about marginal likelihood, it is easy to see that the marginal estimator we propose will produce consistent parameter estimates as long as the marginal models $q_0(y; \theta_0)$ and $q_1(y; \theta_1)$ are correct and the assumed Mendelian mode of inheritance is valid.

Now consider ways of estimating the allele frequency from the data. From Hardy-Weinberg equilibrium, we know that, in the general population, the probability of carrying the mutation is $1 - (1 - f)^2$. Thus, if p denotes the fraction of carriers among the participants, assuming the participants are sampled randomly, one can estimate f by solving the equation $1 - (1 - f)^2 = p$. When the mutation is rare, i.e., $f^2 \approx 0$, $\hat{f} \approx p/2$ (Wacholder et al., 1998). Alternatively, one can maximize the marginal likelihood jointly with respect to θ and f . The information on f from the relatives, though it may be small, can be incorporated in the second approach.

Asymptotic normality of the estimator can be established from existing theory of correlated data (Diggle, Liang, and Zeger, 1996). Let $S_{i+}(\theta, f) = \sum_{j=1}^{n_i} S_{ij}^R(\theta, f) + S_i^P(\theta, f)$ and $u_{i+}(\theta, f) = \sum_{j=1}^{n_i} u_{ij}^R(\theta, f) + u_i^P(\theta, f)$, respectively, denote the sum of the θ -scores and f -scores for the i th family. It follows that, as $K \rightarrow \infty$, $K^{1/2}\{(\hat{\theta}, \hat{f}) - (\theta, f)\}$ converges to a multivariate normal distribution with mean zero and variance-covariance matrix

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & a_{22} \end{bmatrix}^{-1} \begin{bmatrix} B_{11} & B_{12} \\ B_{12}^T & b_{22} \end{bmatrix} \begin{bmatrix} A_{11} & A_{21}^T \\ A_{12}^T & a_{22} \end{bmatrix}^{-1},$$

where

$$A_{11} = - \lim_{K \rightarrow \infty} K^{-1} \frac{\partial}{\partial \theta^T} \sum_{i=1}^K S_{i+}(\theta, f),$$

$$A_{12} = - \lim_{K \rightarrow \infty} K^{-1} \frac{\partial}{\partial f} \sum_{i=1}^K S_{i+}(\theta, f),$$

$$A_{21} = - \lim_{K \rightarrow \infty} K^{-1} \frac{\partial}{\partial \theta^T} \sum_{i=1}^K u_{i+}(\theta, f),$$

$$a_{22} = - \lim_{K \rightarrow \infty} K^{-1} \frac{\partial}{\partial f} \sum_{i=1}^K u_{i+}(\theta, f)$$

and

$$B_{11} = \lim_{K \rightarrow \infty} K^{-1} \sum_{i=1}^K \text{var} \{S_{i+}(\theta, f)\},$$

$$B_{12} = \lim_{K \rightarrow \infty} K^{-1} \sum_{i=1}^K \text{cov} \{S_{i+}(\theta, f), u_{i+}(\theta, f)\},$$

$$b_{22} = \lim_{K \rightarrow \infty} K^{-1} \sum_{i=1}^K \text{var} \{u_{i+}(\theta, f)\}.$$

Considerable simplification occurs when f is estimated by simply solving $1 - (1 - f)^2 = p$. Using the fact that $A_{21} = 0$, $B_{12} = 0$, and $a_{22} = b_{22} = 4/\{1 - (1 - f)^2\}$, it is easy to see that the asymptotic variance of $\hat{\theta}$ is given by

$$A_{11}^{-1} \left\{ B_{11} + \frac{1 - (1 - f)^2}{4} A_{12} A_{12}^T \right\} A_{11}^{-1}.$$

Thus, we see that estimation of f increases the variance of $\hat{\theta}$, the increase being small for rare mutations.

For a fixed value of f , the marginal likelihood can be maximized with respect to $\theta = (\theta_0, \theta_1)$ using an EM algorithm. For simplicity, let us assume that the density of y has the same parametric form, $q(y; \theta)$, for both noncarriers and carriers and that θ_0 and θ_1 , the parameter values corresponding to the carriers and the noncarriers, respectively, may vary independently. It follows that the i th iteration of the EM algorithm involves the following steps:

E-Step. Define two sets of weights, up to a normalizing constant, for the data as

$$W_{0i}^P = I(g_i^P = 0),$$

$$W_{0ij}^R = \text{pr}(g_{ij}^R = 0 | g_i^P) q(y_{ij}^R; \hat{\theta}_0^{(i-1)})$$

$$\div \left[\text{pr}(g_{ij}^R = 0 | g_i^P) q(y_{ij}^R; \hat{\theta}_0^{(i-1)}) \right. \\ \left. + \text{pr}(g_{ij}^R = 1 | g_i^P) q(y_{ij}^R; \hat{\theta}_1^{(i-1)}) \right],$$

and $W_{1i}^P = 1 - W_{0i}^P$, $W_{1ij}^R = 1 - W_{0ij}^R$. Define $\mathbf{W}_{0i}^R = (W_{0i1}^R, \dots, W_{0in_i}^R)$ and $\mathbf{W}_{1i}^R = (W_{1i1}^R, \dots, W_{1in_i}^R)$.

M-Step. In this step, we assume software is available for obtaining the maximum likelihood (ML) estimate of θ corresponding to the model $q(y, \theta)$ from a set of independent observations with a specified set of weights. We obtain the ML estimates from the data $\mathbf{Y} = (\mathbf{Y}_1^R, Y_1^P, \dots, \mathbf{Y}_K^R, Y_K^P)$ once corresponding to the weights $\mathbf{W}_0 = (\mathbf{W}_{01}^R, W_{01}^P, \dots, \mathbf{W}_{0K}^R, W_{0K}^P)$ and once corresponding to $\mathbf{W}_1 = (\mathbf{W}_{11}^R, W_{11}^P, \dots, \mathbf{W}_{1K}^R, W_{1K}^P)$ to obtain $\hat{\theta}_0^{(i)}$ and $\hat{\theta}_1^{(i)}$, respectively.

It is possible to cleverly modify the above steps to allow common parameters between θ_1 and θ_0 .

Let us consider estimating age-specific cumulative risk of the disease when data on age at onset of the disease are available. Let $F_g(t) = 1 - S_g(t)$, $g = 0, 1$, denote the cumulative risk to age t for the noncarriers and carriers, respectively, and $h_g(t)$, $g = 0, 1$, denote the corresponding hazard functions. Let $Y = (T, \delta)$, where T denotes the age at onset of the disease or the age at the end of follow-up, whichever is smaller, and let δ denote the indicator of incidence of the disease before the end of follow-up. Assuming that the distribution of the censoring time for an individual does not depend on his/her genotype, L_M^R and L^P can be shown to be proportional to

$$\prod_{i=1}^K \prod_{j=1}^{n_i} \sum_{g_{ij}^R=0}^1 \text{pr} \left(g_{ij}^R | g_i^P \right) S_{g_{ij}^R} \left(t_{ij}^R \right) h_{g_{ij}^R}^{\delta_{ij}^R} \left(t_{ij}^R \right)$$

and $\prod_{i=1}^K S_{g_i^P} \left(t_i^P \right) \{ h_{g_i^P} \left(t_i^P \right) \}^{\delta_i^P}$, respectively. Parametric models for the cumulative risks can be fitted by the EM algorithm described above.

Wacholder et al. (1998) proposed estimating $F_g(t)$ nonparametrically using the Kaplan-Meier estimates of the cumulative risk of the disease in the first-degree relatives of the noncarrier and carrier participants. They showed that, for a rare mutation,

$$\begin{aligned} R_0(t) &= (1-f)F_0(t) + fF_1(t), \\ R_1(t) &= (0.5-f/2)F_0(t) + (0.5+f/2)F_1(t), \end{aligned}$$

where $R_g(t)$, $g = 0, 1$, denotes the cumulative risk for the first-degree relatives of noncarriers and carriers, respectively. They proposed substituting the Kaplan-Meier estimates for R_0 and R_1 in these approximate equations and solving for $F_0(t)$ and $F_1(t)$. This method, however, does not guarantee that the resulting estimates of the cumulative risk will be monotone in finite samples. In fact, when the mutation is rare, the method often produces a nonincreasing estimate of F_1 (Struwing et al., 1997; Wacholder et al., 1998). To obtain proper nonparametric estimates of S_0 and S_1 , say \hat{S}_0 and \hat{S}_1 , we propose maximizing the marginal likelihood nonparametrically. Let $\mathcal{M} = \{t_1 < t_2 \dots < t_M\}$ be the observed event times in the data. For any T , let $l(T)$ denote the index of the largest event time less than or equal to T . Both \hat{S}_0 and \hat{S}_1 will have potential jumps within the M observed event times. Let $\{\hat{\lambda}_{g1}, \dots, \hat{\lambda}_{gM}\}$ denote the hazard components of \hat{S}_g . To obtain the nonparametric maximum marginal likelihood estimates (NPMMLE), one can use the EM algorithm described above. In the E-step, we substitute

$$\left\{ \prod_{m \leq l(T)} (1 - \hat{\lambda}_{gm}) \right\}^{1-\delta} \left\{ \hat{\lambda}_{gl(T)} \prod_{m < l(T)} (1 - \hat{\lambda}_{gm}) \right\}^{\delta}$$

for $f(y, \theta_g)$. The M-step has a closed-form solution with $\hat{\lambda}_{gm} = \sum_{i \in \mathcal{E}(t_m)} w_{gi} / \sum_{i \in \mathcal{R}(t_m)} w_{gi}$, where $\mathcal{R}(t_m)$ is the set of indices of the individuals at risk, i.e., $T \geq t_m$, and $\mathcal{E}(t_m)$ is the set of indices of the individual(s) with an event at time t_m .

Note that, if we assume the failure time data is discrete, the number of parameters to be estimated, though possibly

large, remains fixed. When the failure time data are continuous, however, the number of event times and hence the number of parameters to be estimated increase as the sample size increases. Asymptotic properties of such a truly nonparametric procedure remain to be studied.

2.4 Nonrandom Sample of Participants

So far we have assumed that the participants can be treated as a random sample from an underlying population. In practice, however, the participants may not be a random sample, both intentionally and unintentionally, and ignoring the ascertainment can cause serious bias. In the Washington Ashkenazi Study, e.g., the families were ascertained through volunteer participants. Since individuals with a family history of breast cancer are more likely to participate (Struwing et al., 1997; Wacholder et al., 1998), the estimates of risk we obtain by treating the volunteers as a random sample are biased upward. When the ascertainment is nonrandom and not under the control of the investigator, it is generally very difficult to correct for such bias. One may intentionally select the participants in a nonrandom way to increase efficiency. When Y is dichotomous, e.g., one may consider a case-control sample instead of a random sample. By increasing the proportion of case participants, one can obtain more carriers in the sample of participants (Wacholder et al., 1998; Gail et al., 1999).

The appropriate marginal likelihood when the sampling of the participants is based on their phenotypes is given by

$$L = L_M^R \times L^P = \prod_{i=1}^K \prod_{j=1}^{n_i} \text{pr} \left(y_{ij}^R | g_i^P, y_i^P \right) \times \prod_{i=1}^K \text{pr} \left(g_i^P | y_i^P \right). \quad (6)$$

Here $\text{pr}(g^P | y^P)$ depends on $q_0(y; \theta_0)$, $q_1(y; \theta_1)$, and $\pi = 1 - (1-f)^2$ through the relationship (Gail et al., 1999)

$$\text{pr}(g=1 | y) = \frac{\pi q_1(y; \theta_1)}{\pi q_1(y; \theta_1) + (1-\pi) q_0(y; \theta_0)}.$$

Thus, L^P depends on the parameters of the marginal model, θ_0, θ_1 , and f . Unfortunately, in the presence of residual familial correlation, L_M^R depends on y^P and is not determined by these parameters. In the absence of residual familial correlation, one has $\text{pr}(y_{ij}^R | g_i^P, y_i^P) = \text{pr}(y_{ij}^R | g_i^P)$ and the marginal likelihood is of the form

$$L = L_M^R \times L_M^P = \prod_{i=1}^K \prod_{j=1}^{n_i} \text{pr} \left(y_{ij}^R | g_i^P \right) \times \prod_{i=1}^K \text{pr} \left(g_i^P | y_i^P \right), \quad (7)$$

which can be computed as a function of θ_0 , θ_1 , and f . The bias of using (7) in the presence of residual correlation has been studied in Section 3.

Moore et al. (2000) observed that, as the number of parameters gets large, the score equations corresponding to the maximum likelihood method of Gail et al. (1999a) become unstable and difficult to solve. As an alternative to the maximum likelihood, they proposed maximizing L^R (ignoring L^P) with respect to θ for fixed f and maximizing L^P (ignoring L^R) with respect to f for fixed θ and iterating between them until convergence. They found that such a pseudolikelihood

approach was stable even for large numbers of parameters. The marginal version of this approach will be very similar except that one should maximize L_M^R (instead of L^R) with respect to θ for fixed f . In Section 3, the bias and efficiency of the pseudolikelihood approach of Moore et al. (2000) and the corresponding marginal version are compared using simulation experiments.

3. Simulation Studies

3.1 Binary Phenotype

In this section, we compare the marginal likelihood with alternative likelihood-based approaches in terms of bias and efficiency. The first series of simulation experiments assumes a dichotomous phenotype. The marginal probabilities of developing the disease for a carrier and a noncarrier, namely ϕ_1 and ϕ_0 , respectively, are chosen to be 0.56 and 0.05 and the allele frequency (f) is chosen to be 0.01. Residual familial correlation between the members of a family are generated using the logistic random effects model (Gail et al., 1999b)

$$\text{pr}(y = 1 | g, b) = \{1 + \exp(-\mu_g + b)\}^{-1},$$

where the random effects b are distributed as $N(0, \tau^2)$ and are independently drawn for each family. The correlation induced by such a random effects model increases with the value of τ^2 , and $\tau^2 = 0.0$ corresponds to absence of residual familial correlation. For each value of τ^2 , μ_1 and μ_0 are chosen so that ϕ_1 and ϕ_0 remain fixed at 0.56 and 0.05. The effects of family size and structure are also investigated by considering three types of families for the participant: (I) mother and one sister, (II) mother and grandmother, and (III) mother and two sisters.

First we report on the results where participants are sampled randomly. We estimated $\theta = (\theta_0, \theta_1)$, where $\theta_i = \log\{\phi_i/(1 - \phi_i)\}$ and $i = 0, 1$, using the MLE (maximum likelihood estimator) and MMLE (maximum marginal likelihood estimator), keeping the value of the allele frequency fixed at its method of moments-based estimate obtained from the proportion of carriers in the sample of participants (see Section 2.3). We note that, in the Washington Ashkenazi Study, though the participants' disease histories were available as

part of the study, Struewing et al. (1997) did not use this data in their analysis in order to avoid any potential bias caused by the effect of survival of the diseased individuals on their participation. For our simulated data, though the possibility of such bias does not exist, we perform an analysis that excluded the participants' phenotype data and compared the results with an analysis that included them. The bias and variances for θ_1 are shown in Table 1. For $\tau^2 = 0.0$, we observe that both MLE and MMLE had small biases and that the MMLE was slightly less efficient than the MLE. As the value of τ^2 increased, the MLE based on a conditional independence assumption became increasingly more biased, whereas the MMLE continued to have very small bias. Comparing the three types of families, we observe that, the higher the gain in efficiency for the MLE over the MMLE when the conditional independence assumption was true, the more severe was the bias for the MLE when the assumption did not hold. Inclusion of the participants' data in the analysis substantially improved the precision of both the MLE and MMLE and reduced the bias of the MLE. For estimation of θ_0 , the difference between the MLE and MMLE were very small (results not reported). For $\tau^2 = 0.0$, the MMLE was as efficient as the MLE, and for $\tau^2 > 0.0$, the MLE had only a small bias.

We repeat the above experiments with a case-control sample of the participants. From a random sample of 5000 participants, we select all the cases and only 6.1% of the controls so that the numbers of cases and controls are approximately the same. The results for the pseudolikelihood procedure of Moore et al. (2000) (PLE), the corresponding marginal version (MPLE), the MLE, and the MMLE are shown in Table 2. Here we observe that, for $\tau^2 > 0.0$, all the methods were biased; the biases of the MPLE and the MMLE, however, were less severe than those of the PLE and the HLE. Somewhat surprisingly, we found that, for $\tau^2 > 0.0$, the PLE and MLE not only had larger bias than the corresponding marginal estimators, they often had larger variances, too. For $\tau^2 = 0.0$, all the methods have small bias and similar efficiency.

Table 1

Kin-cohort design with randomly selected probands: bias and variability in estimation of $\log\{\phi_1/(1 - \phi_1)\}$

		I		II		III		
Family type:		Bias	Var	Bias	Var	Bias	Var	
Using relatives' phenotype data	$\tau^2 = 0.0$	MLE	-0.020	0.065	0.000	0.099	0.005	0.041
		MMLE	-0.015	0.077	-0.087	0.131	0.012	0.050
	$\tau^2 = 1.0$	MLE	0.342	0.106	0.620	0.166	0.521	0.056
		MMLE	-0.002	0.086	-0.083	0.130	0.014	0.056
	$\tau^2 = 2.0$	MLE	0.916	0.186	1.506	0.331	1.134	0.100
		MMLE	0.028	0.084	-0.076	0.133	0.023	0.051
Using participants' and relatives' phenotype data	$\tau^2 = 0.0$	MLE	0.007	0.025	0.001	0.029	-0.002	0.018
		MMLE	0.006	0.027	-0.025	0.031	-0.004	0.020
	$\tau^2 = 1.0$	MLE	0.122	0.034	0.161	0.038	0.239	0.027
		MMLE	0.000	0.034	-0.018	0.037	0.004	0.030
	$\tau^2 = 2.0$	MLE	0.238	0.036	0.303	0.041	0.514	0.032
		MMLE	-0.017	0.031	-0.040	0.037	0.008	0.031

Table 2
Kin-cohort design with case-control sample of probands: bias and variability in estimation of $\log\{\phi_1/(1 - \phi_1)\}$ for PLE, MPLE, MLE, and MMLE

Family type:	I		II		III		
	Bias	Var	Bias	Var	Bias	Var	
$\tau^2 = 0.0$	PLE	0.021	0.132	0.022	0.233	0.002	0.083
	MPLE	0.027	0.130	0.049	0.288	0.005	0.082
	MLE	0.014	0.100	0.030	0.142	0.019	0.073
	MMLE	0.015	0.101	0.050	0.160	0.017	0.073
$\tau^2 = 1.0$	PLE	0.444	0.191	0.579	0.554	0.473	0.134
	MPLE	0.321	0.168	0.331	0.851	0.274	0.110
	MLE	0.338	0.125	0.472	0.200	0.357	0.093
	MMLE	0.289	0.120	0.369	0.150	0.264	0.087
$\tau^2 = 2.0$	PLE	0.679	0.272	1.446	3.852	0.975	0.408
	MPLE	0.354	0.181	0.327	0.690	0.344	0.126
	MLE	0.605	0.230	0.859	0.644	0.702	0.118
	MMLE	0.465	0.165	0.528	0.411	0.420	0.096

3.2 Phenotype: Age at Onset of Disease

In this section, we study the property of the nonparametric marginal likelihood estimator through simulation studies. The time to disease onset is generated from a Weibull distribution for both carriers and noncarriers. The shape and scale parameters for carriers and noncarriers are chosen to be (4.0474, 0.0164) and (5.1479, 0.0133), respectively. These parameter values correspond to a mean age at onset of 69 years for noncarriers and of 55 years for carriers. We assumed that each participant provided data on time to disease onset for his/her mother and one sister. To allow residual correlation between the mother and the sister, we used the Clayton-Oak's bivariate copula model (Clayton and Cuzick, 1978) for their joint distribution so that the marginal distributions for carriers and noncarriers remain fixed at the corresponding Weibull distributions. We chose the association parameter in the copula model so that it corresponds to a Pearson's correlation of 0.5 between two relatives' ages at onset. The censoring mechanism was generated by simulating ages for the mothers and sisters from two normal distributions with mean and variance corresponding to the mothers and sisters of the Washington Ashkenazi data. A woman is considered censored if her simulated age at onset of disease exceeds her simulated age. After generating continuous age data from this model, we rounded them to the nearest integer. We chose the allele frequency to be 0.01. In each replication, we estimated the cumulative risk functions nonparametrically based on relatives' phenotype data using the NPMML. In Figure 1, the average of the estimated risks over 100 simulations has been compared with the corresponding true risk from the Weibull distribution for the carriers. We see that, until age 50, the estimated risk followed the true risk very closely. After age 50, the estimated risk had a small downward bias. The bias, however, went away as we increased the number of families (results not shown).

4. Washington Ashkenazi Study

Struewing et al. (1997) and Wacholder et al. (1998) published a series of plots showing the cumulative risk for various

cancers estimated from the cancer history information of the first-degree relatives of the volunteer participants. They used a total of 4873 family sets, each corresponding to an index participant for whom the genotype information was available. All of the estimated cumulative risk curves were nonmonotone, with the problem being more serious for some of the graphs than the others. Using the same data, we reestimate those cumulative risk curves using the nonparametric marginal likelihood approach described in Section 2.3 (Figure 2). We observe that our estimates, after imposing monotonicity constraints, remain very close to the original estimates of Struewing et al. (1997) (not shown here). As a result, the most substantial differences between the estimates are observed in those age intervals where the original estimates had a drop in cumulative risk. A major goal of the study was to estimate the lifetime penetrance of the mutations for breast cancer, defined as the cumulative risk to age 70. The marginal likelihood approach estimated this quantity as 60%, which is slightly larger than the original estimate of 56%.

5. Discussion

In this paper, we have proposed the use of a marginal likelihood approach to the kin-cohort design for estimation of penetrance. Computationally, the method is simple to implement; the EM algorithm we have defined for nonparametric estimation of the age-specific penetrance has been found to be remarkably faster than similar algorithms for alternative likelihood-based methods discussed here. The robustness property of our method stems from the fact that, if the participants are randomly sampled, the marginal likelihood scores, which ignore the known Mendelian correlation structure between the genotypes of the relatives of the same participant, are unbiased irrespective of the nature of the correlation between the phenotypes of the relatives given their genotypes. On the other hand, the scores corresponding to likelihood-based methods, which assume the phenotypes of the relatives given their genotypes are independent but use the known genotypic correlation of the relatives to construct the true likelihood, are biased if the conditional independence assumption is violated.

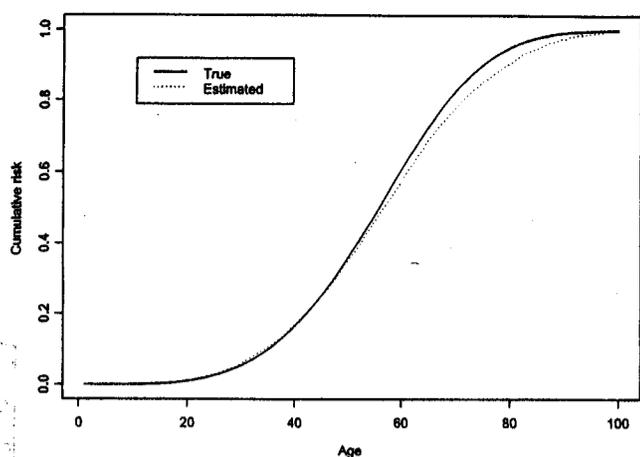


Figure 1. Average of nonparametric estimates of cumulative risk over 100 replications (dotted line) and the true cumulative risk (solid line) for carriers.

When the conditional independence assumption does hold, ignoring the genotypic correlation reduces efficiency. Though we have seen that, for small families, the loss of efficiency is small, for large families, more substantial loss of efficiency is possible. Generalized estimating equation (GEE) (Zhao and Prentice, 1990; Liang, Zeger, and Quaqish, 1992) techniques, with suitable modification for the fact that genotypes of the relatives are unknown, can be used to enhance efficiency of the marginal approach.

Although the marginal likelihood method may be robust to the violation of the conditional independence assumption discussed above, the method can be sensitive to violation of assumptions involving the ascertainment and the Mendelian mode of inheritance, just like other methods. In general, accounting for ascertainment is a complex process and has to be addressed on a case-by-case basis. In Section 2.4, we considered an important special case of nonrandom ascertainment where participants are sampled randomly, conditional on their disease status, for which the marginal likelihood is given in (6). We note that, besides considering a retrospective likelihood for the participants' data, one also needs to condition on the disease status of the participants while computing the contribution of the relatives. Failing to condition can lead to biased estimates of penetrance in the presence of residual correlation between participants and their relatives. We found that this bias, though less severe than that of the alternative likelihood methods discussed, can be substantial (Table 2). Larger biases for the alternative methods in this situation can be attributed to the fact that these methods are affected not only by residual correlation between the participants and the relatives but also by the residual correlation between the relatives of the same participants. To account for conditioning of the participants' disease status in the relatives' contribution to the marginal likelihood, a model for the joint distribution of diseases in relative-participant pairs, defined in terms of the marginal penetrances and some additional parameters describing residual correlation, can be considered. Sensitivity of the marginal likelihood method to the violation of the Mendelian

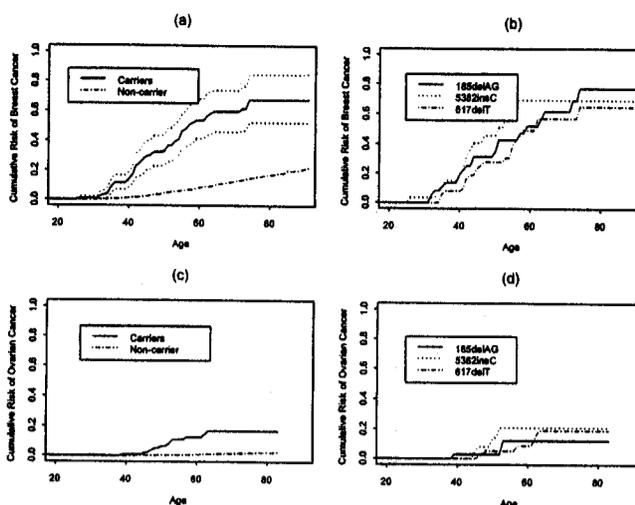


Figure 2. a. Estimated risk of breast cancer and 95% pointwise confidence interval among carriers of a BRCA1 or BRCA2 mutation. The estimated risk for noncarriers is also shown. b. Estimated risk of breast cancer among carriers of each of the three mutations. c. Same as (a) for ovarian cancer. d. Same as (b) for ovarian cancer.

mode of inheritance assumption needs future investigation. Gail et al. (1999b) studied the bias of their likelihood approach to the violation of such an assumption due to population stratification. They found the bias to be small for rare mutations and noticeable for more common alleles. Since, in the marginal approach, the probability calculation for the genotypes uses the mode of inheritance assumption only at a marginal level, we believe that the bias of the marginal approach to violation of this assumption should be less than or equal to the bias in Gail's method.

ACKNOWLEDGEMENTS

We are indebted to Dr Mitch Gail and Dr Dirk More for their constructive comments, which helped to improve the presentation of this paper.

RÉSUMÉ

Le protocole d'étude de cohorte familiale est une alternative prometteuse aux protocoles classiques d'études de cohorte ou cas-témoins pour estimer la pénétrance d'une mutation autosomale rare connue. Dans ce protocole d'étude, un échantillon de participants sélectionné de manière adéquate fournit les informations génotypiques et d'histoire familiale détaillée sur la maladie d'intérêt. Pour estimer la pénétrance de la mutation, nous considérons une approche de vraisemblance marginale qui est simple à réaliser sur le plan calculatoire, plus flexible que l'approche analytique originale proposée par Wacholder et al (1998), et plus robuste à la présence de corrélation familiale résiduelle que l'approche de vraisemblance considérée par Gail et al (1999). Nous étudions le compromis entre robustesse et efficacité par des expériences de simulation. La méthode est illustrée par l'analyse des données de l'étude ashkénaze de Washington.

REFERENCES

- Clayton, D. and Cuzick, J. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**, 141-151.
- Diggle, P., Liang, K., and Zeger, S. (1996). *Analysis of Longitudinal Data*, Oxford Statistical Science Series. Oxford: Clarendon Press.
- Easton, D., Ford, D., and Bishop, D. (1995). Breast and ovarian cancer incidence in BRCA1-mutation carriers. *American Journal of Human Genetics* **56**, 265-271.
- Gail, M., Pee, D., Benichou, J., and Carroll, R. (1999a). Designing studies to estimate the penetrance of an identified autosomal mutation: Cohort, case-control, and genotyped-proband designs. *Genetic Epidemiology* **16**, 15-39.
- Gail, M., Pee, D., and Carroll, R. (1999b). Effects of violations of assumptions on likelihood methods for estimating the penetrance of the autosomal mutations from kin-cohort studies. *Journal of Statistical Planning and Inference*, in press.
- Gail, M., Pee, D., and Carroll, R. (1999c). Kin-cohort designs for gene characterization. *National Cancer Institute Monograph* **26**, 55-60.
- Li, C. C. (1978). *First Course in Population Genetics*. Pacific Grove, California: Boxwood.
- Liang, K., Zeger, S., and Quaqish, B. (1992). Multivariate regression analysis for categorical data. *Journal of the Royal Statistical Society, Series B* **54**, 3-40.
- Moore, D., Chatterjee, N., Pee, D., and Gail, M. (2000). Pseudo-likelihood estimates of the cumulative risk of an autosomal dominant disease from a kin-cohort study. *Genetic Epidemiology*, in press.
- Struwing, J. P., Hartge, P., Wacholder, S., Baker, S. M., Berlin, M., McAdams, M., Timmerman, M. M., Lawrence, B. C., and Tucker, M. A. (1997). The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *The New England Journal of Medicine* **336**, 1401-1408.
- Wacholder, S., Hartge, P., Struwing, J. P., Pee, D., McAdams, M., Lawrence, B., and Tucker, M. A. (1998). The kin-cohort study for estimating penetrance. *American Journal of Epidemiology* **148**, 623-629.
- Zhao, L. and Prentice, R. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* **77**, 642-648.

Received January 2000. Revised July 2000.

Accepted July 2000.