Biometrics 62, 1081–1088
December 2006
DOI: 10.1111/j.1541-0420.2006.00614.x

A Method for Estimating Penetrance from Families Sampled for Linkage Analysis

Yuanjia Wang,[1,*] Ruth Ottman,[2,3] and Daniel Rabinowitz[1]

[1]Department of Statistics, Columbia University, New York, New York 10027, U.S.A.
[2]G. H. Sergievsky Center and Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, New York 10032, U.S.A.
[3]Epidemiology of Brain Disorders Department, New York State Psychiatric Institute, New York, New York 10032, U.S.A
[*]*email:* wang@stat.columbia.edu

Summary. When a gene variant is discovered to segregate with a disease, it may be of interest to estimate the risk (or the age-specific risk) of the disease to carriers of the variant. The families that contributed to the discovery of the variant would typically contain multiple carriers, and so, especially if the variant is rare, might prove a valuable source of study subjects for estimation of the risk. These families, by virtue of having brought the gene in question to the attention of researchers, however, may not be representative of the relationship between carrier status and the risk of the disease in the population. Using these families for risk estimation could bias the observed association between the variant and the risk. The purpose here is to present an approach to adjusting for the potential bias while using the families from linkage analysis to estimate the risk.

Key words: Age at onset; Bayes formula; Epilepsy; Genetics; Risk.

1. Introduction

Studies aimed at discovering genes that influence the risk of disease usually focus on families containing multiple individuals affected with the disease. Linkage analysis is used to search for regions of the genome where affected relatives share genetic marker alleles with greater frequency than would be expected by chance. If such regions are found, the nucleotide sequences of the genes in the regions are examined in affected and unaffected members of the families in which marker alleles in the regions segregate with the disease. Examination of the sequences may reveal a genetic variant shared by all the affected members. If the variant is only infrequently found in the general population, and if the variant seems likely to alter gene function, it is natural to conclude that the variant increases the risk of disease.

When such a variant is found, there may further be interest in estimating its penetrance, that is, in estimating the probability that a randomly selected individual who carries the variant will be affected with the disease. If there is not a single variant of a gene implicated in the disease, but rather many different rare variants each found in the affected members of but one or a few families, it will not be practical to estimate the penetrance of each variant. A practical alternative is to treat all variants as if they each conferred the same degree of risk, and to estimate a common penetrance.

It might be thought that a natural study design for the estimation of penetrance would be to screen a population for carriers of a variant, and then to examine the rate of disease in carriers. For a rare variant, however, it may be prohibitively expensive to obtain a sufficiently large sample of carriers. In such situations, the families from the original linkage analysis that implicated the variant in the etiology of the disease might appear an attractive source of data, as the families would typically contain multiple carriers. However, the frequency of disease in the carriers from those families may not accurately reflect the risk to carriers in the general population.

Two aspects of ascertainment schemes that draw from families included in a linkage analysis would cause the inaccuracy. First, the inclusion criteria in the original linkage study typically would have called for families with multiple diseased individuals; families in the population where carriers are not affected would tend to be underrepresented. Second, among the families that are sampled, selection for inclusion in the penetrance analysis would typically require that all, or most, of the affected members are carriers of the variant (otherwise the investigators would not have inferred that the variant caused the disease); families in the population in which the variant does not lead to disease would tend to be underrepresented. Methods that do not take these issues into account would generally be biased. The purpose here is to present an approach to estimation from these kinds of families that avoids the bias.

The approach requires that carrier status in at least some unaffected members of the families be obtained, and is based

on an application of Bayes formula to the probability that unaffected members are carriers. The Bayes calculation rests on the assumption of a rare dominant variant and a rare disease. In what follows, the Bayes calculation is described and used to develop a conditional likelihood that may be used for estimating penetrance. A version of the approach that is appropriate for congenital diseases is first described; the discussion is then extended to diseases with variable age at onset. The approach is illustrated by applying it to estimating the risk of autosomal-dominant partial epilepsy with auditory features (ADPEAF) conferred by mutations in the leucine-rich glioma inactivated 1 (LGI1) gene.

2. Estimating Penetrance

It is convenient to begin with a simple example. Consider a sample obtained from nuclear families with two or more siblings affected with a rare disease, and suppose that it is discovered that at some locus, in a portion of the families, all of the affected siblings carry a particular rare mutation. Finally, suppose that the mutation is so rare that in any given nuclear family in which the mutation is carried, almost certainly there is but one parent carrying but a single copy of the mutation.

Let π denote the risk of disease in carriers, and let r denote the probability that an unaffected sibling is a carrier. Then, by a formal application of Bayes theorem—with the marginal probability of being a carrier taken to be one half (by virtue of the assumption that there is a single parental copy of the mutation), and the conditional probability of being unaffected when not a carrier approximated to be one (by virtue of the assumption that the disease is rare)—r is approximately equal to

$$\frac{0.5(1-\pi)}{0.5(1-\pi)+0.5\times 1}. \tag{1}$$

The utility of the Bayes calculation for risk estimation may be observed by, for example, inverting the formula (1) to obtain the estimate of the penetrance,

$$\hat{\pi}=\frac{2\hat{r}-1}{\hat{r}-1}, \tag{2}$$

where $\hat{r}$ is the empirical carrier frequency in the unaffected siblings of affected carriers.

To verify the relevance of the assumption, it is useful to define some further notation. Let I denote the event that a given family in the population is chosen for the original study, and let E denote the event that all the affected members of the family are carriers. Let U denote the subset of family members who are unaffected, let u denote the observed value of U, let a denote the complement of u in the family members, and let k denote the cardinality of a. For any set of siblings s, let C_s denote the carrier status in the set s, and let $C_s^\star$ denote the event that all members of the set s are carriers. Let n denote the number of observed unaffecteds in the family, and, for any given value c of the carrier status of the subjects in u, let $m(c)$ denote the number of carriers. Finally, let π denote the risk in the population that a carrier is diseased, and let ϵ denote the corresponding risk for noncarriers.

It may be helpful to note that the notation U refers to the subset of unaffected family members, C refers to carrier status of the subjects, and the subscripts a and u refer to affected and unaffected. The added asterisk in $C_a^\star$ refers to the event that all affected members are carriers.

Two assumptions are applied in the derivation of the likelihood. The first one is the assumption that, although disease status in family members would almost certainly play a role in the inclusion criteria, family members' genotype, which certainly could influence their disease status, would not otherwise influence whether a family is included in the original study sample. More precisely, the assumption is that the inclusion of a family in the original sample is conditionally independent of the carrier status in the unaffected members of the family, given the disease status of the family members. The second assumption is that the disease status of the unaffected family members are conditionally independent, given their carrier status and the fact that all the affected members carry a mutation. This is the important assumption that there is no familial aggregation of additional risk factors. More precisely, the assumption is

$$P\left\{U=u\,\middle|\,C_u=c,C_a^\star\right\}=\prod_i P\left\{U_i=u_i\,\middle|\,C_{ui}=c_i,C_a^\star\right\}. \tag{3}$$

Here U_i is the ith unaffected member of the family, C_{ui} is the carrier status of that member, and u_i and c_i are observed values of U_i and C_{ui}.

Note that each conditional probability in the right-hand side of (3) is equal to $1-\pi$ when c_i denotes a carrier, and is equal to $1-\epsilon$ when c_i denotes a noncarrier. Then, the probability that the carrier status in the unaffected members of the family is an observed value c, conditionally given that the family is included in the sample, given that all of the affected siblings in the family are carriers, and given the affection status in all of the family members, may be expressed as

$$\begin{aligned}
&P\{C_u=c\,|\,I,U=u,E\}\\
&\quad=P\left\{C_u=c\,\middle|\,U=u,C_a^\star\right\}\\
&\quad=\frac{P\left\{C_a^\star\right\}P\left\{C_u=c\,\middle|\,C_a^\star\right\}P\left\{U=u\,\middle|\,C_u=c,C_a^\star\right\}}{\sum_{\tilde{c}}P\left\{C_a^\star\right\}P\left\{C_u=\tilde{c}\,\middle|\,C_a^\star\right\}P\left\{U=u\,\middle|\,C_u=\tilde{c},C_a^\star\right\}}\\
&\quad=\frac{P\left\{C_u=c\,\middle|\,C_a^\star\right\}\prod_i P\left\{U_i=u_i\,\middle|\,C_{ui}=c_i,C_a^\star\right\}}{\sum_{\tilde{c}}P\left\{C_u=\tilde{c}\,\middle|\,C_a^\star\right\}\prod_i P\left\{U_i=u_i\,\middle|\,C_{ui}=\tilde{c}_i,C_a^\star\right\}}\\
&\quad=\frac{P\left\{C_u=c\,\middle|\,C_a^\star\right\}(1-\pi)^{m(c)}(1-\epsilon)^{n-m(c)}}{\sum_{\tilde{c}}P\left\{C_u=\tilde{c}\,\middle|\,C_a^\star\right\}(1-\pi)^{m(\tilde{c})}(1-\epsilon)^{n-m(\tilde{c})}}.
\end{aligned}$$

Here the assumption of conditional independence between inclusion of the family and carrier status of unaffected subjects given affection status is used in the first equality. The assumption of no familial aggregation of additional risk factors is used in the third equality. Specializing to the case of a single unaffected offspring in a nuclear family (so that the conditional probability that any unaffected subject is a carrier, given $C_a^\star$, is one half) and approximating ϵ, the risk of disease in noncarriers, by zero, provides the Bayes calculation in (1).

These considerations suggest the construction of an approximate conditional likelihood for π as the product of family-specific terms of the form

$$\frac{P\{C_u = c \mid C_a^\star\}(1-\pi)^{m(c)}}{\sum_{\tilde{c}} P\{C_u = \tilde{c} \mid C_a^\star\}(1-\pi)^{m(\tilde{c})}}.$$

This likelihood has the form of a conditional logistic regression with an offset in which each family corresponds to a stratum. The subjects in the families do not correspond to the stratum members, however; rather, the values of $\tilde{c}$, the possible value of C_u, play the role of stratum members. The analog of the regression parameter in the canonical parameterization is the natural logarithm of $(1-\pi)$, the corresponding predictor is $m(\tilde{c})$, and the analog of the offset is the natural logarithm of $P\{C_u = \tilde{c} \mid C_a^\star\}$. The analog of the response variable is an indicator, that is, 1 for the observed value of C_u, and 0 otherwise; each stratum has exactly one response variable equal to 1. Once the offset terms have been obtained, the maximum likelihood estimate of π may be computed using standard statistical software.

It may be of interest to examine the calculation of the offset terms and the likelihood for some simple cases. Figure 1 shows four example pedigrees. In each, a filled symbol indicates a diseased subject, and an open symbol indicates that the subject is not diseased. The notation $M/+$ indicates a mutation carrier, the notation $+/+$ indicates a noncarrier, and the absence of both indicates that genotype information is not available.

The first of the four pedigrees corresponds to the simple case used to introduce the approach: the presence of carriers among siblings *iii* provides that one of the parents is a heterozygous carrier; the probability that the offspring *iv* is a carrier is therefore 0.5, and the contribution to the likelihood is

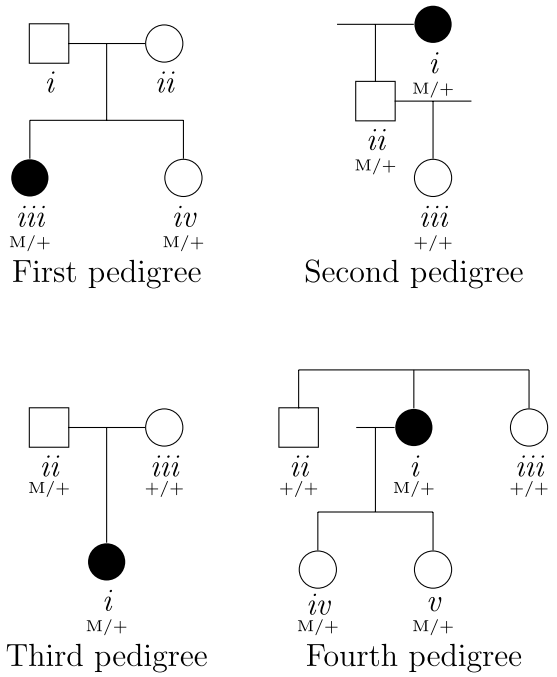

Figure 1. Example pedigrees.

$$\frac{0.5(1-\pi)}{0.5(1-\pi) + 0.5 \times 1}.$$

In the second example pedigree, disease in subject i provides that subject i is a carrier. So, conditionally, there are three possible outcomes: the variant is not transmitted to subject *ii* (and so also not to subject *iii*); the variant is transmitted to subject *ii*, but not to subject *iii*; or the variant is transmitted to subject *ii* and then to subject *iii*. The conditional probabilities for these outcomes are 0.5, 0.25, and 0.25. It is the second of the three outcomes that is depicted as having occurred, so the contribution to the likelihood is given by

$$\frac{0.25(1-\pi)}{0.5(1-\pi)^0 + 0.25(1-\pi) + 0.25(1-\pi)^2}$$
$$= \frac{1-\pi}{2 + (1-\pi) + (1-\pi)^2}.$$

In the third example pedigree, the affection status in subject i provides that either parent *ii* or *iii* is a carrier. It is convenient to introduce for the conditional probability that *ii* is a carrier of notation η. (It might be natural to take η to be 0.5, reflecting that each parent is equally likely to be the carrier. It may be, however, for example through population admixture, that one or another parent is more likely to be the carrier. In any case, whether η is taken to be 0.5 or not, the treatment of the family is the same.) The approximate conditional likelihood takes the form

$$\frac{\eta(1-\pi)}{\eta(1-\pi) + (1-\eta)(1-\pi)} = \eta.$$

The information provided by the parental pair, therefore, is ancillary, and the parental pair does not contribute to the estimate.

In the fourth example pedigree, the presence of disease in subject i ensures that the carrier status in siblings *ii* and *iii* and in offsprings *iv* and *v* are conditionally independent, given the disease status in the family members. Each of the unaffected subjects has probability 0.5 of being a carrier. The probability of carrying the variant for these subjects, given that they are not diseased, therefore, is given by (1). The contribution to the conditional likelihood is that of four independent Bernoulli variables with expectation equal to the probability.

A sampling protocol sometimes taken when collecting families begins with obtaining affected probands with multiple affected first-degree relatives, and then proceeds sequentially by examining first-degree relatives of previously included affected family members. With this strategy, the typical contributions to the conditional likelihood will have the form, as in the fourth example pedigree, corresponding to independent Bernoulli variables each with expectation as given in (1). The maximization of the likelihood in these cases has the simple form in which the maximum likelihood estimate of π is the transformation (2) of the overall carrier rate in the unaffected subjects contributing to the analysis.

The approach as proposed so far is appropriate for congenital diseases or for diseases that manifest before the ages at which subjects are recruited into the study. For diseases with variable age at onset, however, the methods are not

appropriate: an infant free of a disease that generally manifests itself in adolescence, for example, contributes entirely different information than does a disease-free adult subject, and the ages of the subjects should be reflected in the analysis. Furthermore, with a disease with variable age at onset, it may be of particular interest to estimate age-specific risk. This section concludes with a generalization to the problem of estimating age-specific risk.

The extension of the approach to estimate the age-specific risk requires an additional assumption. The assumption is that the subjects' age at ascertainment is conditionally independent of the subjects' age at onset, given their carrier status. Under this additional assumption, the essential step in moving to the estimation of age-specific risk is to modify the terms used in the construction of the conditional likelihood to reflect the age at which the disease status is ascertained. The modification is to replace the probability of having experienced onset of the disease, π, by age-specific terms, π_t, where t is the age of the subject at the time the disease status is ascertained. For any value of c of C_u, the carrier status in the unaffected family members, let $X_j(c)$ denote the indicator that the jth family member is a carrier. Let t_j denote the age of the jth subject at the time of ascertainment (if subjects are not alive or have been lost to follow-up at the time of ascertainment, t_j should be taken to be the age at death or at loss to follow-up). Then, the modified conditional likelihood with age-specific terms is given by the product of family-specific terms of the form

$$\frac{P\left\{C_u = c \,\middle|\, C_a^\star\right\} \prod_j (1-\pi_{t_j})^{X_j(c)}}{\sum_{\tilde{c}} P\left\{C_u = \tilde{c} \,\middle|\, C_a^\star\right\} \prod_j (1-\pi_{t_j})^{X_j(\tilde{c})}}.$$

The modified conditional likelihood should be maximized under the obvious monotonicity constraint that π_u should not exceed π_v for u less than or equal to v. Analogous to the likelihood for a congenital disease, the modified conditional likelihood may also be expressed in the form of a conditional logistic regression in which the vector of parameters has natural logarithms of $(1 - \pi_t)$ as its components. It is evident from the conditional logistic regression interpretation that the likelihood is convex as a function of the vector whose components are the natural logarithm of $(1 - \pi_t)$. It follows, therefore, from the results of Barlow et al. (1972), that the maximum likelihood estimates may be obtained from a pooled-adjacent-violators algorithm. A convex minorant algorithm may also be applied, as discussed in Groeneboom and Wellner (1992). In the special case where the conditional likelihood corresponds to independent Bernoulli variables, the pool-adjacent-violators and convex minorant algorithms have a relatively simple form corresponding to the estimation of a distribution function from current status data.

In some situations there might be interest in estimating the influence of factors other than age on the penetrance. To do so, one might augment the likelihood by expressing the probability of disease as a function of age and additional factors. A parameterization of the probability of remaining disease free at time t given $Z = z$ in the form $1 - \pi_{t,z} = (1 - \pi_t)e^{z\alpha}$, where Z is an encoding of the additional factors and α is a vector of regression coefficients, is particularly convenient, as the likelihood may still be represented as that of a conditional logistic regression. In the special case where the contribution to the likelihood corresponds to that of independent Bernoulli observations each with expectation $(1 - \pi_{t,z})/(2 - \pi_{t,z})$, it may be more convenient to choose a parameterization in which the logit of the expectation has the form $g(\beta, t) + z\alpha$. Here $g(\beta, t)$ might, for example, take the form of a polynomial in t whose coefficients are the components of β. To obtain $\pi_{t,z}$ from the estimated regression coefficients, the relationship, $\log(1 - \pi_{t,z}) = g(\beta, t) + z\alpha$, may be used. Standard logistic regression programs may be used to estimate the coefficients in this parameterization.

3. Illustrative Data Analysis

This section reports the results of an application of the methods proposed here to data on ADPEAF, a form of idiopathic lateral temporal lobe epilepsy with auditory symptoms as a major seizure manifestation (see, e.g., Ottman et al., 1995; Poza et al., 1999; Winawer et al., 2000, 2002). Mutations in the LGI1 gene have been found to cause this syndrome in 21 families reported so far, and this number is rapidly increasing (see, e.g., Berkovic et al., 2004; Hedera et al., 2004; Ottman et al., 2004; Pisano et al., 2005). The families with mutations have all contained multiple individuals with epilepsy, in patterns that appear consistent with autosomal-dominant inheritance with reduced and age-dependent penetrance. With one exception, each reported mutation has been found in only a single family (see, e.g., Gu, Brodtkorb, and Steinlein, 2003; Pizzuti and Giallonardo, 2003; Pizzuti et al., 2003).

The original evidence for a causal effect of LGI1 mutations on ADPEAF was obtained through a two-stage process. In the first stage, linkage analysis was carried out in a large family with apparently autosomal-dominant inheritance of epilepsy (Ottman et al., 1995). This led to the detection of a chromosome segment (on chromosome 10q24) shared more often by family members with epilepsy than would be expected by chance. In the second stage, the DNA sequences of genes known to reside within the chromosome segment were studied in affected and unaffected individuals to determine whether or not any contained a disease-associated variant from the original linkage family and from additional families with ADPEAF (Kalachikov et al., 2002). Initially one affected individual from the family was tested. After a variant in LGI1 was discovered, its association with the disease was confirmed by showing that all of the affected individuals in the family carried the variant whereas only a small proportion of unaffected individuals did. The observation that a small number of unaffected individuals carried the mutations was viewed as evidence for reduced penetrance.

Mutation testing has subsequently been carried out in additional families, selected because they contained two or more individuals with the same type of epilepsy as in the original family (Ottman et al., 2004). In each such family, the DNA from one affected subject was tested to detect a variant in LGI1. When a variant was discovered, the remaining affected subjects were tested to confirm that they also carried it. Finally, the variant's association with disease was confirmed by

testing unaffected individuals, both within the family and in a set of unrelated controls.

Six families contributed to the penetrance estimation analysis (four of the five families originally used to demonstrate a causal effect of mutations on ADPEAF, and two others; two additional families with mutations were uninformative). The numbers of unaffected subjects contributing to the analysis in the six families were 7, 4, 8, 3, 1, and 16. The numbers of unaffected carriers in these families were 2, 2, 2, 1, 0, and 4, respectively. The observed carrier rate in the unaffected members was therefore $\hat{r} = 11/39$ (28%), so applying (2) provides that the estimate of the penetrance, unadjusted for age, was $\hat{\pi} = 17/28$ (61%). Recall that in this example the likelihood in all of the families corresponded to independent Bernoulli random variables under the assumption of no familial aggregation of additional risk factors. With the same assumption, the exact confidence interval for $\hat{r}$ can be computed based on binomial distribution. To be specific, the lower bound is found by solving the equation $\sum_{k=0}^{X_{\text{obs}}-1} r_L^k(1-r_L)^{N-k} = 1-\alpha/2$, and the upper bound is found by solving the equation $\sum_{k=0}^{X_{\text{obs}}} r_U^k(1-r_U)^{N-k} = \alpha/2$. The asymptotic confidence interval can be computed by first computing the standard error of $\hat{r}$ as $(\hat{r}(1-\hat{r})/n)^{\frac{1}{2}}$ and then using the normal approximation. The resulting exact 95% confidence interval for $\hat{r}$ is (15%, 45%) and the asymptotic confidence interval is (14%, 42%). Because $\hat{\pi}$ is estimated using (2), the confidence interval for $\hat{\pi}$ can be computed by applying (2) to the endpoints of the confidence interval for $\hat{r}$. The resulting exact confidence interval for $\hat{\pi}$ is (18%, 82%) and the asymptotic confidence interval is (28%, 84%).

It should be noted that several subjects included in the analysis are younger than age 20. These younger subjects who had not yet developed the disease at the time of ascertainment may experience onset in the future. It is particularly interesting to apply an age adjustment in this case.

The contribution to the likelihood in all of the families corresponded to independent Bernoulli variables, so that the computation of the nonparametric maximum likelihood estimate of age-specific penetrance may be carried out by estimating the age-specific carrier risk, and then applying the transformation (2).

The ages of the unaffected subjects together with their mutation status are recorded in the table in the Appendix, which can be found at `www.tibs.org/biometrics`. The nonparametric maximum likelihood estimate of the age-specific probability of a positive carrier status for a decreasing function is constant on the intervals from 20 to 31, from 36 to 65, and from 69 to 70. The age-specific estimates of the carrier rate on the intervals are 1/4, 4/19, and 0, respectively. By relation (2), the corresponding values of the penetrance are 2/3, 11/15, and 1. The age-specific penetrance is 0 on the interval from 1 to 17. These rates are depicted in Figure 2.

Introducing age into the model, while providing only a fairly coarse representation of the age-specific penetrance in this example, provides perspective on the estimate derived from the raw carrier rate: the lifetime risk of disease due to the gene is estimated to eventually reach 1, and, for the most part, onset appears to occur between 20 and 31. It should be noted that in this sparse data set, this shape of the estimated curve is driven primarily by the two unaffected carriers aged

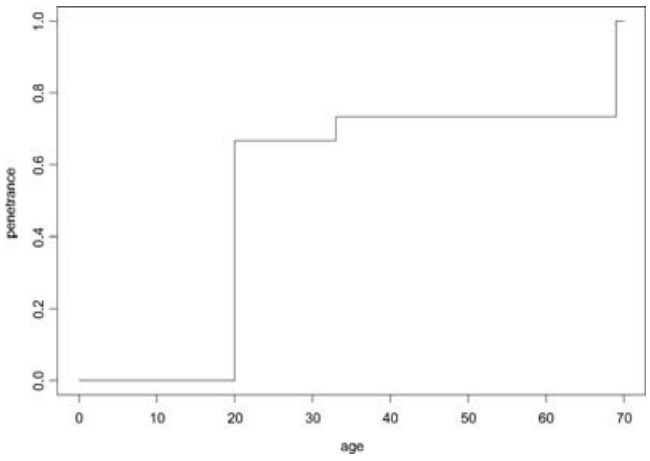

Figure 2. Nonparametric maximum likelihood estimate of the age-specific penetrance.

30 and 31 and by the two unaffected noncarriers aged 69 and 70. The two subjects with age 69 and 70 cause the estimate of penetrance to approach 1. An application of the Lynden–Bell estimator to the age-at-onset data in affected individuals, who are all carriers, suggests that onset tends to occur between ages 16 and 29. This is earlier than the results of this analysis suggest; the discrepancy can possibly be explained by random variation and the sparseness of this data set. The distribution of age at onset for the affected subjects is presented in Figure 3.

A logistic regression model for the conditional probability of positive carrier status was also fit, in which age was entered as a linear term. No intercept was fitted to ensure the penetrance is zero at age zero. The regression parameter estimate for age was −0.027 (SE: 0.010, 95% CI: (−0.047, −0.07)). The corresponding age-specific penetrance function is depicted in Figure 4, and reaches a maximum of 85% (CI: 39%, 96%) at age 70. Here the confidence interval for $\hat{\pi}_t$ is computed by applying the transformation

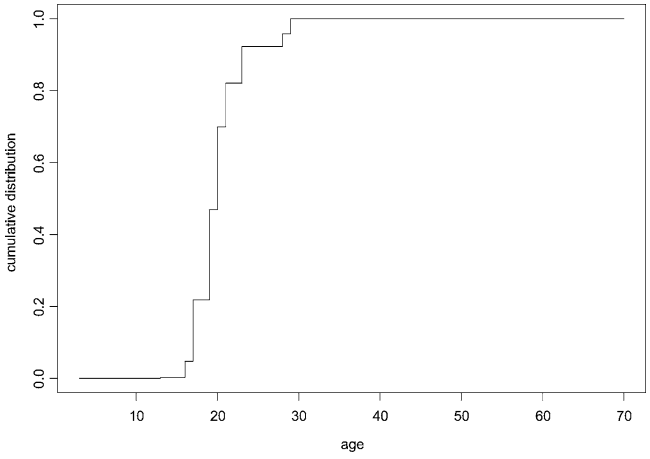

Figure 3. Nonparametric maximum likelihood estimate of the conditional distribution of age at onset for affected carrier subjects.

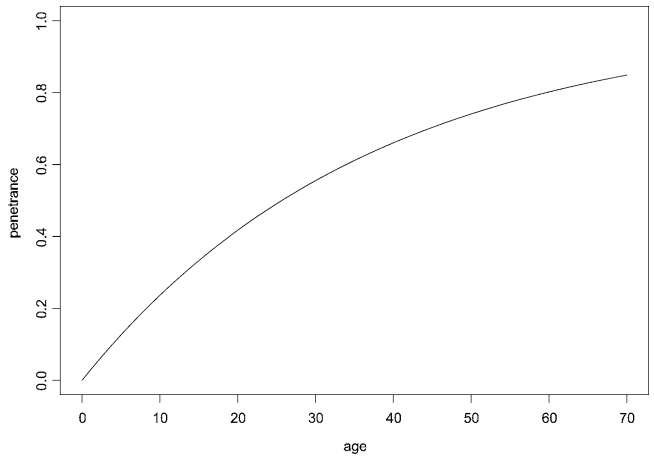

Figure 4. Parametric maximum likelihood estimate of the age-specific penetrance.

$$g(\beta) = \frac{2\exp(\beta t)/(1+\exp(\beta t)) - 1}{\exp(\beta t)/(1+\exp(\beta t)) - 1}$$

to the endpoints of the confidence interval for $\hat{\beta}$.

4. Bias Introduced by Familial Aggregation of Additional Risk Factors

An important assumption that underlies the proposed methods is that there is no additional risk factor, other than the genotype of interest, that aggregates in families. In this section, the effect of violating this assumption is investigated.

Three settings are considered. In the first scenario, the sampled families each have two siblings, one affected and one unaffected. These families correspond to the first pedigree of the example pedigrees in Figure 1. The behavior of the estimate of penetrance is investigated when there is an additional Bernoulli risk factor shared by family members. Denote the shared family-specific risk factor by X, and let the probability of a sibling being affected, given the carrier status C and risk factor X, be

$$P(D \mid C, X) = \frac{e^{\beta C + \delta X}}{1 + e^{\beta C + \delta X}}.$$

Here it is assumed that the family-specific X_i are independent and that X_i are independent of C. Denote $P(X_i = 1)$ by θ. The penetrance for the carriers is then $\pi = \theta p_1 + (1-\theta)p_0$, where $p_1 = e^{\beta+\delta}/(1 + e^{\beta+\delta})$, and $p_0 = e^{\beta}/(1 + e^{\beta})$. It is assumed that the risk of disease for the noncarriers is approximately 0.

It is shown in the Appendix, available at `www.tibs.org/biometrics`, that the bias in this setting is

$$E(\hat{\pi}) - \pi = (\alpha - \theta)(p_1 - p_0),$$

where

$$\alpha = \frac{\theta p_1(1-p_1)}{\theta p_1(1-p_1) + (1-\theta)p_0(1-p_0)}.$$

The bias is a function of the risk of the susceptible genetic variant C measured by β, the risk of the additional factor X measured by δ, and the prevalence of the risk factor measured by θ. The magnitude of the bias is evaluated relative to the correlation coefficient between the siblings. The correlation coefficient of the response variable D between the siblings in the same family is

$$r = \frac{\theta p_1^2 + (1-\theta)p_0^2 - \theta p_1 - (1-\theta)p_0}{(\theta p_1 + (1-\theta)p_0)(1 - (\theta p_1 + (1-\theta)p_0))}.$$

Table 1
Bias due to familial aggregation of an additional risk factor—first scenario

θ	True penetrance	Correlation coefficient	Bias	Relative bias
		$\beta = 0.1$, $\delta = 1.5$		
0.1	0.304	0.034	0.007	0.215
0.2	0.286	0.062	0.014	0.221
0.3	0.269	0.084	0.019	0.226
0.4	0.251	0.099	0.023	0.230
0.5	0.233	0.108	0.025	0.232
0.6	0.215	0.110	0.026	0.233
0.7	0.197	0.103	0.024	0.231
0.8	0.179	0.085	0.019	0.227
0.9	0.162	0.053	0.012	0.219
δ	True penetrance	Correlation coefficient	Bias	Relative bias
		$\beta = 0.1$, $\theta = 0.5$		
0.2	0.310	0.002	0.000	0.047
0.4	0.298	0.010	0.001	0.073
0.6	0.286	0.021	0.002	0.100
0.8	0.273	0.037	0.005	0.129
1.0	0.261	0.055	0.009	0.158
1.2	0.249	0.075	0.014	0.188
1.4	0.238	0.097	0.021	0.218
1.6	0.228	0.119	0.029	0.247
1.8	0.219	0.141	0.039	0.276
2.0	0.210	0.162	0.049	0.303

Relative bias = bias/correlation coefficient.

The bias, the relative bias, and the correlation coefficient for different combinations of θ, β, and δ are recorded in the tables. Table 1 shows that the bias increases with increasing δ. Table 1 also shows that the bias first increases and then decreases with an increasing value of θ, the prevalence of the risk factor. The magnitude of the absolute bias can be as much as 0.05, and the relative bias can be as much as 30%.

In the second scenario, it is assumed that the sampled families have more than two family members. The calculation of the bias carries through with a different value of α. To be specific, when there are m family members and k unaffected members in each family,

$$\alpha = \frac{\theta \binom{m}{k} p_1^k (1-p_1)^{m-k}}{\theta \binom{m}{k} p_1^k (1-p_1)^{m-k} + (1-\theta)\binom{m}{k} p_0^k (1-p_0)^{m-k}}.$$

The bias in this setting is recorded in Table 2. The magnitude of the bias increases with the number of subjects in each family. It can also be seen that the absolute value of

Table 2
Bias due to familial aggregation of an additional risk factor—second scenario

Family size	Number of unaffecteds	Bias	Relative bias
$\theta = 0.5$, $\beta = 0.1$, $\delta = 1.5$, $\pi = 0.23$, $r = 0.11$			
3	1	0.064	0.591
3	2	0.005	0.049
3	3	−0.067	−0.616
4	1	0.081	0.748
4	2	0.057	0.530
4	3	−0.015	−0.140
4	4	−0.077	−0.708
5	1	0.087	0.802
5	2	0.080	0.741
5	3	0.049	0.457
5	4	−0.034	−0.314
5	5	−0.082	−0.762
6	1	0.088	0.818
6	2	0.087	0.804
6	3	0.079	0.734
6	4	0.040	0.372
6	5	−0.050	−0.461
6	6	−0.085	−0.791

Relative bias = bias/correlation coefficient.

the bias first decreases and then increases for increasing numbers of unaffecteds in each family. The sign of bias changes from positive to negative. This experiment implies that the amount of bias when large families are sampled could be substantial.

In a third scenario, the behavior of the estimate is investigated where there is a continuous risk factor with density $f(x)$ shared by the family members. The probability of a family member being affected, given C and X, can be correspondingly modeled as

$$\pi(x) = P(D \mid C, X = x) = \frac{e^{\beta C + \delta x}}{1 + e^{\beta C + \delta x}}.$$

Then the penetrance is $\int \pi(x) f(x)\, dx$. As shown in the Appendix, the bias in this case is

$$\int \frac{\pi(x)(1-\pi(x))}{\int \pi(x)(1-\pi(x)) f(x)\, dx} \pi(x) f(x)\, dx - \int \pi(x) f(x)\, dx.$$

The bias evaluated with a normal density with different means and standard deviations is recorded in Table 3. It can be seen that the bias due to a large shift of mean from 0 could be substantial. However, the influence of standard deviation is moderate.

From these calculations, it can be seen that the bias introduced by the familial aggregation of additional risk factors can be large depending on certain scenarios. The bias increases with decreasing values of the true penetrance and increasing values of the correlation between family members.

5. Discussion

A standard strategy for adjusting for trait-based sampling in genetic epidemiology is to condition on observed phenotypes. See, for example, Gong and Whittemore (2003). The approach developed here differs from the standard approach in conditioning on genotypes in affected members as well as on the observed phenotypes. This is appropriate because inclusion in the analysis is not simply determined by observed phenotypes; rather, families are included in the analysis because affected family members are carriers of the variant. It is important to note that if the carrier status in the unaffected family members is used to determine that a family is worthy of inclusion in the penetrance analysis (through the observation that unaffected family members are less often carriers than would be expected under random assortment), then the approach presented here is not directly applicable.

The approach proposed here is predicated on the assumption of a rare variant. The importance of this assumption lies in being able to infer that a pair of parents carrying the variant would comprise a single heterozygous carrier and a homozygous noncarrier. It should be noted, however, that sampling strategies that call for multiple affected individuals may result in samples with nonnegligible frequency of homozygous parents or a pair of heterozygous parents. Whenever it is possible to observe that there is a homozygous parent or a pair of heterozygous parents, it should be

Table 3
Bias due to familial aggregation of an additional risk factor—third scenario

Mean	Standard deviation	True penetrance	Correlation coefficient	Bias	Relative bias
		$\beta = 0.1$, $\delta = 1.5$			
−2	1	0.104	0.192	0.112	0.586
−1	1	0.269	0.267	0.087	0.325
0	1	0.518	0.292	−0.007	−0.024
1	1	0.759	0.259	−0.095	−0.366
2	1	0.910	0.180	−0.110	−0.612
−2	10	0.424	0.894	0.070	0.078
−1	10	0.463	0.894	0.034	0.038
0	10	0.503	0.894	−0.002	−0.003
1	10	0.542	0.894	−0.039	−0.043
2	10	0.581	0.893	−0.074	−0.083

Relative bias = bias/correlation coefficient.

used in the computation of the offset terms in the conditional logistic regression.

In the illustrative example, the families that contributed to the analysis are the ones in which all the affected members in the families carried the same variant in the LGI1 gene. Generally, it is not necessary to restrict the analysis to the families where all the affected members would carry the same variant. However, as a practical matter, the genetic variant is not likely to be identified except when it is responsible for all cases within a family. Moreover, the proposed approaches are concerned with a sufficiently rare disease and sufficiently penetrant mutations, so it is unlikely that multiple etiologies will coexist in sampled families.

The methods proposed here focused on the case of a dominant variant. For a recessive variant, or more generally, for a case where penetrance in homozygous carriers is greater than penetrance in heterozygotes, the frequency in the sample families of homozygous individuals could be quite substantial. This, of course, would have an influence on what could be inferred about parental carrier status in cases where parental genotype information is not available. A reasonable strategy for handling missing genotype information might be to use estimates of the distribution of mating types to develop likelihoods that take into account ambiguities. However, there remains a question as to the extent that misspecification of the mating-type frequencies could induce bias, and whether there are methods that are robust to misspecifications.

Finally, it should be noted that the approach proposed here takes no notice of the possibility of heterogeneity in penetrance due to the presence of unmeasured genetic or environmental factors that might cluster in families. See, for example, the discussion in Begg (2002). The presence of multiple affected family members might tend to indicate the presence of additional factors that predispose toward the disease, so that samples obtained through multiple affected members might be families where additional factors are overrepresented. The methods proposed here avoid bias induced by the sampling scheme in the absence of heterogeneity, but inference about a heterogeneous population as a whole can be confounded by sampling schemes that are restricted to families containing multiple affected individuals, which cause some portion of the population to be represented disproportionately. The sensitivity of the developed method when violating the assumption of no familial aggregation of additional risk factors is investigated in several settings. The magnitude of the bias can be substantial when the true penetrance is low and the correlation is large.

Acknowledgement

This research was supported in part by NIH grants GH55978, NS36319, and NS43472.

References

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

Begg, C. B. (2002). On the use of familial aggregation in population-based case probands for calculating penetrance. *Journal of the National Cancer Institute* **94,** 1221–1226.

Berkovic, S. F., Izzillo, P., McMahon, J. M., et al. (2004). LGI1 mutations in temporal lobe epilepsies. *Neurology* **62,** 1115–1119.

Gong, G. and Whittemore, A. S. (2003). Optimal designs for estimating penetrance of rare mutations of a disease-susceptibility gene. *Genetic Epidemiology* **24,** 173–180.

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

Gu, W., Brodtkorb, E., and Steinlein, O. K. (2003). LGI1 is mutated in familial temporal lobe epilepsy characterized by aphasic seizures. *Annals of Neurology* **52,** 364–367.

Hedera, P., Abou-Khalil, B., Crunk, A. E., Taylor, K. A., Haines, J. L., and Sutcliffe, J. S. (2004). Autosomal dominant lateral temporal epilepsy: Two families with novel mutations in the LGI1 gene. *Epilepsia* **45,** 218–222.

Kalachikov, S., Evgrafoc, O., Ross, B., et al. (2002). Mutations in LGI1 cause autosomal-dominant partial epilepsy with auditory features. *Nature Genetics* **30,** 335–341.

Ottman, R., Risch, N., Hauser, W. A., et al. (1995). Localization of a gene for partial epilepsy to chromosome 10q. *Nature Genetics* **10,** 56–60.

Ottman, R., Winawer, M. R., Kalachikov, S., Barker-Cummings, C., Gilliam, T. C., Pedley, T. A., and Hauser, W. A. (2004). LGI1 mutations in autosomal dominant partial epilepsy with auditory features. *Neurology* **62,** 1120–1126.

Pisano, T., Marini, C., Brovedani, P., Brizzolara, D., Pruna, D., Mei, D., Moro, F., Cianchetti, C., and Guerrini, R. (2005). Abnormal phonologic processing in familial lateral temporal lobe epilepsy due to a new LGI1 mutation. *Epilepsia* **46,** 118–123.

Pizzuti, A., and Giallonardo, A. T. (2003). Correction. *Annals of Neurology* **54,** 137.

Pizzuti, A., Flex, E., Di, B. C., Dottorini, T., Egeo, G., and Manfredi, M. (2003). Epilepsy with auditory features: A LGI1 gene mutation suggests a loss-of-function mechanism. *Annals of Neurology* **53,** 396–399.

Poza, J. J., Saenz, A., Martinez-Gil, A., Cheron, N., Cobo, A. M., Urtasun, M., Marti-Masso, J. F., Grid, D., Beckmann, J. S., Prud'Homme, J. F., and Munain, A. L. (1999). Autosomal dominant lateral temporal epilepsy: Clinical and genetic study of a large Basque pedigree linked to chromosome 10q. *Annals of Neurology* **45,** 182–188.

Winawer, M. R., Ottman, R., Hauser, W. A., and Pedley, T. A. (2000). Autosomal dominant partial epilepsy with auditory features: Defining the phenotype. *Neurology* **54,** 2173–2176.

Winawer, M. R., Martinelli, B. F., Barker-Cummings, C., Lee, J. H., Liu, J., Mekios, C., Gilliam, T. C., Pedley, T. A., Hauser, W. A., and Ottman, R. (2002). Four new families with autosomal dominant partial epilepsy with auditory features: Clinical description and linkage to chromosome 10q24. *Epilepsia* **43,** 60–67.

Received April 2005. *Revised February* 2006.
Accepted March 2006.