

Robustness of Inference on Measured Covariates to Misspecification of Genetic Random Effects in Family Studies

Ruth M. Pfeiffer,^{1*} Allan Hildesheim,¹ Mitchell H. Gail,¹ David Pee,² Chien-Jen Chen,³ Alisa M. Goldstein,¹ and Scott R. Diehl⁴

¹Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland

²Information Management Services, Inc., Rockville, Maryland

³Graduate Institute of Epidemiology, College of Public Health, National Taiwan University, Taipei, Taiwan

⁴Center for Pharmacogenomics and Complex Disease Research, New Jersey Dental School, Newark, New Jersey

Family studies to identify disease-related genes frequently collect only families with multiple cases. It is often desirable to determine if risk factors that are known to influence disease risk in the general population also play a role in the study families. If so, these factors should be incorporated into the genetic analysis to control for confounding. Pfeiffer et al. [2001 *Biometrika* 88: 933–948] proposed a variance components or random effects model to account for common familial effects and for different genetic correlations among family members. After adjusting for ascertainment, they found maximum likelihood estimates of the measured exposure effects. Although it is appealing that this model accounts for genetic correlations as well as for the ascertainment of families, in order to perform an analysis one needs to specify the distribution of random genetic effects. The current work investigates the robustness of the proposed model with respect to various misspecifications of genetic random effects in simulations. When the true underlying genetic mechanism is polygenic with a small dominant component, or Mendelian with low allele frequency and penetrance, the effects of misspecification on the estimation of fixed effects in the model are negligible. The model is applied to data from a family study on nasopharyngeal carcinoma in Taiwan. *Genet Epidemiol* 24:14–23, 2003. © 2003 Wiley-Liss, Inc.

Key words: ascertainment; conditional logistic regression; correlated binary data; misspecified model; nested random effects model

*Correspondence to: Ruth Pfeiffer, National Cancer Institute, 6120 Executive Blvd., EPS 8030, Bethesda, MD 20892-7244. E-mail: pfeiffer@mail.nih.gov

Received for publication 21 May 2002; Revision accepted 19 June 2002

Published online in Wiley InterScience (www.interscience.wiley.com).

DOI: 10.1002/gepi.10191

INTRODUCTION

This paper was motivated by a linkage study to find genes that predispose to nasopharyngeal carcinoma (NPC). The NPC study was a collaborative effort between the National Institutes of Health (Bethesda, MD) and the Institute of Epidemiology at the National Taiwan University in Taipei, and is based on a sample of approximately 150 Taiwanese families. Families were included in the study only if they had two or more affected family members. Although the primary purpose of this study is to detect genes linked to NPC, it is important to determine if measured environmental factors that are known to influence disease risk in the general population also play a role in the study of families. If so, these environmental factors should be incorporated into

the genetic analysis to control for confounding and potentially improve the power to find NPC-associated genes. It is thus desirable to be able to use data from family studies to estimate the effects of environmental exposures before attempting to identify the susceptibility genes.

A second, related problem to which the methods we discuss apply is the estimation of the effect of a measured major gene in the presence of residual genetic correlations induced by other unmeasured genes or environmental factors.

A natural approach to account for ascertaining families with a fixed number of cases would be to conduct a matched case-control analysis with matching on family, and to use a conditional logistic regression that conditions on the number of cases in the family. This approach can lead to underestimates of exposure effects if genetic

correlations are ignored [Pfeiffer et al., 2001]. Pfeiffer et al. [2001] gave conditions under which conditional logistic regression yields unbiased estimates of measured exposure effects, and proposed a variance components or random effects model that accounts for common familial effects and for varying genetic correlations among family members. After adjusting for ascertainment, they found maximum likelihood estimates of the measured exposure effects based on this model.

Although it is appealing that these procedures account for genetic correlations as well as for the ascertainment of the families, in order to perform an analysis one needs to specify the distribution of random genetic effects. A criticism of the model could be that as the gene (or genes) inducing correlations among the response variables is unmeasured, no straightforward diagnostics are available to evaluate the model assumptions. It is thus important to assess the robustness of the inference to misspecifications of the random effects distribution. This issue has been considered for linear mixed effects models [e.g., Butler and Louis, 1992; Muthen and Shedden, 1999], and simulations have shown that estimation of the fixed parameters is often not severely compromised [e.g., Verbeke and Lesaffre, 1997]. However, it is not clear how conclusions for the linear case carry over to the nonlinear setting. Hartford and Davidian [2000] investigated violations of the assumptions of normality of the random effects distribution in nonlinear mixed effects models via simulations, using first-order expansions and Laplace approximations to evaluate integrals. Due to the ascertainment correction, our likelihood does not fall into any of the standard mixed-effects model frameworks that were investigated by Hartford and Davidian [2000].

In this paper, we first modify the ascertainment correction used in Pfeiffer et al. [2001] to more accurately reflect the ascertainment in this study and also greatly simplify the computations. We then examine the robustness of the estimates of fixed effects parameters against violations of the assumptions made for the random genetic effects in a simulation study. The misspecifications of the random effects distribution that we study are motivated by plausible underlying genetic mechanisms of NPC. We apply the model to a subset of the NPC data, and also fit the simpler standard conditional logistic regression model for comparison.

METHODS

Let Y_{ij} denote the binary disease status for the j th member of the i th family ($Y_{ij} = 1$ if diseased and 0 otherwise), and X_{ij} , the corresponding vector of measured covariates for $j = 1, \dots, n_i$, and $i = 1, \dots, m$. Let n_i denote the size of the i th family and m the total number of families in the study.

In the two-level random effects model, the probability $p_{ij} = P(Y_{ij} = 1)$ is a function of the covariate X_{ij} , the random familial effect a_i , which affects all family members equally, and an individual level random genetic effect g_{ij} for the j th individual in the i th family:

$$\begin{aligned} \text{logit}(p_{ij}) &= \text{logit} P(Y_{ij} = 1 | a_i, g_{ij}, X_{ij}) \\ &= \mu + \sigma_a a_i + \sigma_g g_{ij} + \beta X_{ij}. \end{aligned} \quad (1)$$

The a_i are assumed to be independent and identically distributed with $E(a_i) = 0$ and $\text{var}(a_i) = 1$, and are assumed to be independent of the g_{ij} s in the general population. The g_{ij} s have mean zero, variance one, and are correlated within the i th family. The specification of the correlation structure is given below.

Model (1) is appealing because it allows one to combine information on an individual's measured characteristics and covariates, X_{ij} , with a measure of the genetic liability, g_{ij} . In this model, β describes the increase in log relative odds from a unit increase in exposure, X , for an individual conditional on the random effects. In the NPC study, scientists plan to measure candidate genes. Once the genes are measured, they can be included as known covariates in model (1), making β the most relevant exposure parameter.

Under the logistic model (1), the marginal probability of the response in the i th family requires multidimensional integration over the random effects distribution, an operation that cannot be carried out in closed form, and is written as

$$\begin{aligned} P(Y_{i1}, \dots, Y_{in_i} | X_{i1}, \dots, X_{in_i}) \\ = \int \dots \int \prod_{j=1}^{n_i} p_{ij}^{y_{ij}} q_{ij}^{1-y_{ij}} dF(a, g), \end{aligned} \quad (2)$$

where $q_{ij} = 1 - p_{ij}$.

ASCERTAINMENT CORRECTION

To account for the fact that the selected families are not a random sample of families in the population, but have at least two cases, the likelihood function of the data should be

conditioned on the ascertainment event. Let $Y_i = \sum_{j=1}^{n_i} Y_{ij}$ denote the number of cases in the i th family. Conditioning on the ascertainment event, $Y_i \geq 2$, leads to the following likelihood for m families in the sample:

$$\begin{aligned}
L(\underline{Y}_1, \dots, \underline{Y}_m, \beta) &= \prod_{i=1}^m P(Y_{i1}, Y_{i2}, \dots, Y_{in_i} | X_{i1}, \dots, X_{in_i}, Y_i \geq 2) \\
&= \prod_{i=1}^m \frac{P(Y_{i1}, Y_{i2}, \dots, Y_{in_i}, Y_i \geq 2 | X_{i1}, \dots, X_{in_i})}{P(Y_i \geq 2)} \\
&= \prod_{i=1}^m \frac{P(Y_{i1}, Y_{i2}, \dots, Y_{in_i} | X_{i1}, \dots, X_{in_i})}{1 - P(Y_i \leq 1)} \\
&= \prod_{i=1}^m \frac{\prod_{j=1}^{n_i} \exp(\beta Y_{ij} X_{ij}) \int \prod \lambda_{ij}^Y(\beta) dF(a, g)}{1 - \int \prod_{l=1}^{n_i} \lambda_{il}^0(\beta) dF(a, g) - D_i} \quad (3)
\end{aligned}$$

where

$$D_i = \sum_l \exp(\beta X_{il}) \int \lambda_{il}^1 \prod_{j \neq l} \lambda_{ij}^0(\beta) dF(a, g)$$

and

$$\lambda_{ij}^Y(\beta) = \frac{\exp\{Y_{ij}(\mu + \sigma_a a_i + \sigma_g g_{ij})\}}{1 + \exp(\mu + \sigma_a a_i + \sigma_g g_{ij} + \beta X_{ij})}, \quad (4)$$

for $Y = 0, 1$.

Note that the dimension of the integral in (3) is $n_i + 1$, where n_i is the length of the vector g and an additional integral is added for the familial random effect a .

An alternative approach to correct for ascertainment, as used in Pfeiffer et al. [2001], is to condition on a slightly stronger event, the exact number Y_i of cases in a family. After integrating over the unobserved random effects, the second conditional likelihood function for m families is the product

$$\begin{aligned}
L(\underline{Y}_1, \dots, \underline{Y}_m, \beta) &= \prod_{i=1}^m P(Y_{i1}, Y_{i2}, \dots, Y_{in_i} | X_{i1}, \dots, X_{in_i}, Y_i) \\
&= \prod_{i=1}^m \frac{\prod_{j=1}^{n_i} \exp(\beta Y_{ij} X_{ij}) \prod \lambda_{ij}^Y(\beta) dF(a, g)}{\sum_{l=1}^{n_i} \exp(\beta Y_{ij} X_{ij}) \prod \lambda_{il}^Y(\beta) dF(a, g)} \quad (5)
\end{aligned}$$

where the summation is over all possible choices of Y_i cases out of n_i family members; see, for example, Breslow and Day [1980]. For example, if

$Y_i = 2$, there are $n_i(n_i - 1)/2$ summands, corresponding to all choices of 2 cases out of the n_i family members.

If there is no residual correlation, i.e., $g_{ij} = 0$ for all i, j , model (1) reduces to the simpler logistic model that assumes a family-specific intercept $\mu + a_i$, and independence of the Y_{ij} s given a_i :

$$\begin{aligned}
\text{logit}(p_{ij}) &= \text{logit}P(Y_{ij} = 1 | \mu, a_i, X_{ij}) \\
&= \mu + \sigma_a a_i + \beta X_{ij}. \quad (6)
\end{aligned}$$

In this model, conditioning on Y_i yields the standard conditional likelihood

$$L(\underline{Y}_1, \dots, \underline{Y}_m, \beta) = \prod_{i=1}^m \frac{\prod_{j=1}^{n_i} \exp(\beta Y_{ij} X_{ij})}{\prod_{k=1}^{n_i} \exp(\beta Y_{ik} X_{ik})} \quad (7)$$

that was studied, for example, by Kraft and Thomas [2000] for sibship-based case control studies.

Note that while (3) as well as (5) yield asymptotically unbiased estimates of β , the likelihood (3) that is obtained by conditioning on $Y_i \geq 2$ contains more information on the parameters than likelihood (5). In particular, if $g_{ij} = 0$, the parameters μ and σ_a^2 cancel out the likelihood (5) but not (3). Our simulations indicate that using (3) will result in more efficient estimates of β , as well as the intercept parameters, than using (5). O'Neill and Barry [1995] found similar results on the efficiency of conditioning for a logistic regression model without random effects.

Conditioning on $Y_i \geq 2$ instead of on Y_i also requires less computation, because the number of terms in the summand in (5) grows exponentially with family size, whereas the number of terms in the denominator of (3) grows linearly.

SPECIFICATION OF RANDOM EFFECTS DISTRIBUTIONS

Following Fisher [1918], we assume a polygenic model in which the g_{ij} s are normally distributed, and the covariance matrix can be represented as the sum of an additive genetic variance and a dominance variance. We further assume that the g_{ij} s in (3) have only an additive component of variance. The additive covariance matrix Σ_i for the i th family is a function of the degree of kinship $k(j, l)$ between members j and l in the family:

$$\text{cov}(g_{ij}, g_{il}) = (\Sigma_i)_{j,l} = 2^{-k(j,l)}. \quad (8)$$

For example, $k(j, j) = 0$, and $k(j, l) = 1$ if j and l are first-degree relatives, e.g., siblings, and $k(j, l) = 2$

if j and l represent second-degree relatives, such as a grandparent and a grandchild, or an aunt and a nephew. Thus for each extra generation that separates two members of a family, the correlation is multiplied by a factor of $1/2$. For unrelated members of the family, such as spouses, $k(j, l) = \infty$ and $(\Sigma_i)_{j,l} = 0$.

The random familial intercept a_i is assumed to arise from a standard normal distribution, and is assumed to be independent of the g_{ij} s.

RESULTS

A SIMULATION STUDY

We use simulations to study the robustness of the estimates of β based on the likelihood (3) under the polygenic model with covariance (8) when, in fact, the data derive from 1) a polygene that incorporates both an additive and a dominant covariance component, or 2) a single Mendelian gene (dominant, recessive, or additive). We want to stress that the main interest lies in estimating β , and not the parameters of the random effects distributions.

Each data set in the simulations consists of 100 families. For simplicity, we assume that all families have the same size, $n_i = 6$, and the same correlation structure: mother, father, and four offspring, i.e., $(\Sigma_i)_{j,l} = 0$ for $j = 1$ and $l = 2$, and $(\Sigma_i)_{j,l} = 1/2$ for all other $j \neq l$.

The true value of β is 1.0, and the univariate covariate X is simulated from a Bernoulli distribution with probability parameter $p = 1/2$. Independently of X , the genetic liabilities g are simulated, as described in more detail below. Phenotypes were generated according to equation (1), and families were included in the sample only if $Y_i \geq 2$.

The parameters of the likelihood were estimated by direct maximization. To evaluate the integrals in the conditional likelihood function, we used Monte Carlo integration with a Monte Carlo sample size of $N = 5,000$. To avoid convergence problems with respect to μ and σ_a^2 , we performed a grid search on those two parameters. For further details on the computations, see Pfeiffer et al. [2001]. In addition to assessing the robustness of model (3), we also estimate β based on standard conditional logistic regression (7), that completely ignores varying correlations between family members. Estimates based on conditional logistic regression are denoted by $\hat{\beta}_{CLR}$ while $\hat{\beta}_{RE}$ stands for the random effects model estimates. We let

$\bar{\beta}_{CLR}$ represent the mean of the simulated estimates $\hat{\beta}_{CLR}$, and define $\bar{\beta}_{RE}$ similarly. These estimates and their standard errors (Tables I and II) were based on at least 100 simulated data sets for each condition examined. The minimum number of simulations was 114, and the maximum was 148. The estimates $\hat{\beta}_{RE}$ and $\hat{\beta}_{CLR}$ converged in every situation.

Genetic effects: polygenic with covariance including a dominant component. The covariance matrix of a polygenic random effects model that allows for both an additive and a dominant component depends on a second parameter, σ_d , and is expressed as $(\Sigma)_{j,j} = 1 + \sigma_d^2/\sigma_g^2$, $(\Sigma)_{j,l} = 1/2 + \sigma_d^2/(4\sigma_g^2)$ if j and l are siblings. Except for sibling (including twin) pairs, the dominant component of covariance occurs only when there is inbreeding. For all other types of relatives, $(\Sigma)_{j,l} = 2^{-k(j,l)}$, where, again, $k(j, l)$ denotes the degree of kinship between members j and l in the family. In this paper we do not consider twinning or inbreeding.

Ignoring this dominant component of variance can lead to biased estimates of β , and the bias is more severe for $\hat{\beta}_{CLR}$ than for $\hat{\beta}_{RE}$ in all cases studied (Table I). Nonetheless, even $\hat{\beta}_{RE}$ can be substantially biased when $\sigma_d^2/\sigma_g^2 \approx 1$. For $\mu = -5$, $\sigma_g^2 = 1$ and $\sigma_d^2 = 1$, $\hat{\beta}_{RE} = 0.87$, while $\hat{\beta}_{CLR} = 0.83$, and for $\mu = -3$ with the same values for σ_g^2 and σ_d^2 , $\hat{\beta}_{RE} = 0.88$ and $\hat{\beta}_{CLR} = 0.82$. When $\sigma_d^2 < \sigma_g^2$, as for $\sigma_d^2 = 0.5$ and $\sigma_g^2 = 1$, we obtain for $\mu = -5$ an estimate of $\hat{\beta}_{RE} = 0.96$, while $\hat{\beta}_{CLR} = 0.92$ has an 8% bias. In the same situation for $\mu = -3$, the estimate $\hat{\beta}_{RE} = 0.98$, while $\hat{\beta}_{CLR} = 0.91$ shows a 9% bias toward the null. When the magnitude of the omitted variance component, σ_d^2 , is smaller than the magnitude of the variance component in the model, σ_g^2 , the effects on $\hat{\beta}_{RE}$ are less pronounced, as the random effects model captures most of the correlation present in the data. The larger the dominant component is compared to σ_g^2 , the stronger the impact of ignoring the dominant component is on the bias of the regression estimates in the random effects model. While the magnitude of μ does not influence $\hat{\beta}_{RE}$, it does impact $\hat{\beta}_{CLR}$. As noted by Pfeiffer et al. [2001], the bias in $\hat{\beta}_{CLR}$ increases as μ gets closer to zero, namely in more common diseases.

Not surprisingly, the coverage of confidence intervals is below the nominal 95% level for all situations that exhibit noticeable bias. The coverage of the likelihood ratio-based intervals for the random effects model is always closer to the nominal 95% level than the coverage of the

TABLE I. Results for estimation of β when data are simulated from polygenes having additive and dominant covariance components, while the random effects analysis model assumes only an additive component^a

Simulation parameters $\mu, \sigma_a^2, \sigma_g^2, \sigma_d^2$	Random effects model		CLR	
	Mean estimates $\bar{\mu}, \bar{\sigma}_a^2, \bar{\beta}_{RE}, \bar{\sigma}_g^2$	CI coverage ^b for β_{RE}	Mean β_{CLR}	CI coverage ^b for β_{CLR}
-5, 1.0, 1.0, 1.0	-4.37, 0.85, 0.87, 0.86 (1.25, 0.96, 0.20, 0.95)	0.86	0.83 (0.19)	0.79
-5, 1.0, 1.0, 0.5	-4.43, 0.78, 0.96, 0.86 (1.22, 0.90, 0.19, 0.89)	0.95	0.92 (0.17)	0.95
-5, 1.0, 1.5, 0.5	-4.44, 0.81, 0.97, 1.31 (1.20, 0.88, 0.18, 1.20)	0.96	0.90 (0.16)	0.93
-3, 1.0, 1.0, 1.0	-2.83, 0.91, 0.88, 1.02 (1.06, 0.99, 0.20, 1.19)	0.83	0.82 (0.16)	0.80
-3, 1.0, 1.0, 0.5	-2.89, 0.93, 0.98, 1.01 (1.10, 1.05, 0.21, 1.24)	0.92	0.91 (0.18)	0.90
-3, 1.0, 1.5, 0.5	-3.12, 1.20, 0.93, 1.41 (1.20, 1.22, 0.22, 1.38)	0.85	0.83 (0.17)	0.81

^aThe empirical standard errors are given the line below the estimates.

^bCoverage for nominal 95% confidence interval for β . A test-based confidence interval derived from the likelihood ratio statistic was used.

confidence intervals from standard conditional logistic regression estimates.

Because the dominance component only affects sibling pairs (apart from the exceptions noted above), the family structure we studied (two parents and four offspring) poses a severe test of robustness from omitting the dominant term. Families with relatives with varying degrees of kinship would likely yield less biased estimates for β_{RE} than indicated in Table I.

Genetic effects: Mendelian. Here, we study bias in β_{RE} that results when an additive polygenic model is fit to data in which genetic correlations are induced by a single biallelic gene.

Let d denote the wild-type allele, and D the disease-associated allele. Let $D_{ij} = 0, 1, 2$ denote the number of alleles D that individual ij is carrying. To simulate the data, the score functions $g_{ij} = g(D_{ij})$ in model (1) are defined as follows for various genetic models. For the dominant model, $g_{ij} = g(D_{ij}) = 1$ for $D_{ij} = 1, 2$, and 0 otherwise. For a recessive model, $g(D_{ij}) = 1$ for $D_{ij} = 2$, and 0 otherwise. For the additive model on the logit scale, $g(D_{ij}) = 1$ for $D_{ij} = 2$, $1/2$ for $D_{ij} = 1$, and 0 for $D_{ij} = 0$. The allele frequencies used for the simulations presented in Table 2 are $p = 0.01$ for the dominant model and $p = 0.1$ for the recessive model to reflect moderate gene frequencies, and $p = 0.05$ for the dominant model and $p = 0.25$ for the recessive model to assess the case of a more prevalent genetic disease component. To generate genotypes for a random family, we first select the parental genotypes at random from the general population assuming Hardy-Weinberg equilibrium,

and then generate the genotypes for the offspring assuming Mendelian transmission. Other features of the simulations are the same as described previously for polygenes.

The values of μ, σ_a, σ_g , and p in Table II were chosen to reflect plausible levels of genetic risk for NPC. Setting $\beta = 0$ in (1), we can calculate the attributable risk and penetrances for Mendelian models (Appendix). Note that σ_g in this model denotes the change in log relative odds for a unit change in g , and σ_g^2 no longer represents the variance of genetic effects, which is instead $\sigma_g^2 \text{Var}(g)$. For $\mu = -5, \sigma_a = 1, \sigma_g = 1$, and a dominant gene with allele frequency $p = 0.01$, the attributable risk is 3.1%, the penetrance for a noncarrier is 1%, and the penetrance for a carrier is 2.9%. If the allele frequency is changed to $p = 0.05$, then the attributable risk increases to 13.5%. For $\mu = -5, \sigma_a = 1, \sigma_g = 2$, and a dominant gene with allele frequency $p = 0.01$, the penetrance for a noncarrier is 1%, and the penetrance for a carrier increases to 7.0%. The attributable risk in this situation is 5.4%. When μ changes to -3 , the attributable risk is 2.4% for $p = 0.01$, while it is 10.82% when $p = 0.05$. The penetrance for a noncarrier in this setting is 7%, and the penetrance for a carrier is 15.7%.

Table II summarizes the simulations. We first discuss dominant models. In the rare disease case where $\sigma_g = 1.0$ and $\mu = -5$, standard conditional logistic regression yields nearly unbiased estimates of $\beta_{CLR} = 0.99$ [see Pfeiffer et al., 2001], while the estimate based on the random effects model is $\beta_{RE} = 1.01$. Even for the intercept of

TABLE II. Results for estimation of β when data are simulated from a biallelic single gene model, with empirical standard errors given on line below estimates

Simulation model and parameters $\mu, \sigma_a, \sigma_g, p$	Random effects model		CLR	
	Mean estimates $\bar{\mu}, \bar{\sigma}_a^2, \bar{\beta}_{RE}, \bar{\sigma}_g^2$	CI coverage ^a for β_{RE}	Mean $\bar{\beta}_{CLR}$	CI coverage ^a for β_{CLR}
Dominant: -5, 1.0, 1.0, .01	-4.33, 0.49, 1.01, 0.38 (1.18, 0.64, 0.18, 0.57)	0.95	0.99 (0.18)	0.96
Dominant: -5, 1.0, 1.0, .05	-4.59, 0.64, 1.03, 0.50 (1.22, 0.75, 0.16, 0.66)	0.97	1.01 (0.16)	0.98
Dominant: -5, 1.0, 5.5, .05	-4.88, 0.57, 0.94, 14.44 (1.50, 1.12, 0.38, 9.52)	0.93	0.48 (0.17)	0.26
Dominant: -3, 1.0, 1.0, .01	-3.29, 1.06, 1.04, 0.49 (1.04, 0.91, 0.19, 0.49)	0.93	1.00 (0.11)	0.94
Dominant: -3, 1.0, 1.0, .05	-3.16, 1.04, 1.00, 0.51 (1.17, 0.97, 0.19, 0.75)	0.95	0.96 (0.18)	0.95
Dominant: -5, 1.0, 2.0, .01	-4.52, 0.61, 0.99, 0.22 (1.24, 0.69, 0.18, 0.72)	0.94	0.96 (0.17)	0.97
Dominant: -3, 1.0, 2.0, .01	-3.23, 0.89, 1.02, 0.84 (1.09, 0.88, 0.19, 1.13)	0.97	0.95 (0.16)	0.98
Recessive: -5, 1.0, 1.0, 0.1	-4.63, 0.63, 1.04, 0.44 (1.32, 0.69, 0.22, 0.63)	0.90	1.01 (0.21)	0.91
Recessive: -5, 1.0, 1.0, 0.25	-4.60, 0.58, 1.00, 0.43 (1.19, 0.70, 0.18, 0.59)	0.95	0.98 (0.18)	0.97
Recessive: -5, 1.0, 2.0, 0.1	-4.72, 0.71, 1.08, 0.41 (1.34, 0.81, 0.19, 0.54)	0.94	1.00 (0.18)	0.98
Recessive: -3, 1.0, 1.0, 0.1	-3.37, 1.09, 1.04, 0.43 (1.16, 0.94, 0.20, 0.69)	0.95	1.01 (0.18)	0.94
Recessive: -3, 1.0, 1.0, 0.25	-3.25, 0.99, 1.04, 0.54 (1.18, 0.98, 0.19, 0.80)	0.94	1.00 (0.18)	0.97
Recessive: -3, 1.0, 2.0, 0.1	-3.27, 1.00, 1.01, 0.54, (1.16, 0.91, 0.19, 0.92)	0.92	0.97 (0.18)	0.97
Additive: -5, 1.0, 1.0, 0.1	-4.74, 0.65, 1.00, 0.11 (1.35, 0.77, 0.19, 0.59)	0.91	0.98 (0.19)	0.94
Additive: -5, 1.0, 2.0, 0.1	-4.59, 0.74, 1.02, 0.57 (1.39, 0.83, 0.18, 0.66)	0.96	0.99 (0.17)	0.94
Additive: -3, 1.0, 1.0, 0.1	-3.38, 1.18, 1.03, 0.57 (1.16, 1.01, 0.19, 0.83)	0.93	0.98 (0.16)	0.97
Additive: -3, 1.0, 2.0, 0.1	-3.11, 1.05, 1.02, 0.67 (1.10, 0.98, 0.22, 1.00)	0.91	0.97 (0.19)	0.99

^aCoverage for nominal 95% confidence interval for B. A test-based confidence interval derived from the likelihood ratio statistic was used.

$\mu = -3$, both fixed-effects parameters are nearly unbiased, with $\bar{\beta}_{RE} = 1.04$ and $\bar{\beta}_{CLR} = 1.00$. For $\mu = -5$ and $\sigma_g = 2.0$, the conditional logistic regression estimate $\bar{\beta}_{CLR} = 0.96$ is slightly smaller than $\bar{\beta}_{RE} = 0.99$. For $\mu = -3$ and $\sigma_g = 2.0$, the rare disease assumption is violated, and the conditional logistic regression estimate, $\bar{\beta}_{CLR} = 0.95$ shows a 5% bias, whereas $\bar{\beta}_{RE} = 1.02$. When $p = 0.05$, there is no bias for $\mu = -5$ and $\sigma_g = 1$, but when $\mu = -3$, $\bar{\beta}_{CLR} = 0.96$ shows a small bias that is not seen for $p = 0.01$. When $\mu = -5$, $p = 0.05$, and $\sigma_g = 5.5$, a parameter setting that corresponds to 60%c regression estimate $\bar{\beta}_{CLR} = 0.96$ is slightly smaller than $\bar{\beta}_{RE} = 0.99$. For $\mu = -3$ and $\sigma_g = 2.0$, the rare disease assumption is violated, and the conditional logistic regression estimate, $\bar{\beta}_{CLR} = 0.95$ shows a 5% bias, whereas $\bar{\beta}_{RE} = 1.02$. When $p = 0.05$, there is no bias for $\mu =$

-5 and $\sigma_g = 1$, but when $\mu = -3$, $\bar{\beta}_{CLR} = 0.96$ shows a small bias that is not seen for $p = 0.01$. When $\mu = -5$, $p = 0.05$, and $\sigma_g = 5.5$, a parameter setting that corresponds to 60% penetrance for gene carriers, $\bar{\beta}_{RE} = 0.94$ exhibits a small bias, while $\bar{\beta}_{CLR} = 0.48$ shows a 52% bias, due to the strong influence of the omitted genetic component.

Slightly different patterns for bias are seen in the recessive model. With an allele frequency of $p = 0.1$, all parameter settings result in virtually unbiased estimates of β for both models. Even when the allele frequency is increased to $p = 0.25$, we see very little bias for $\mu = -5$, $\sigma_g = 1$, as $\bar{\beta}_{CLR} = 0.97$.

For the additive model we see a small bias even for $\sigma_g = 1$ and $\mu = -5$ for conditional logistic regression, $\bar{\beta}_{CLR} = 0.98$, while $\bar{\beta}_{RE}$ is unbiased. In all situations the bias is less than 3%, however.

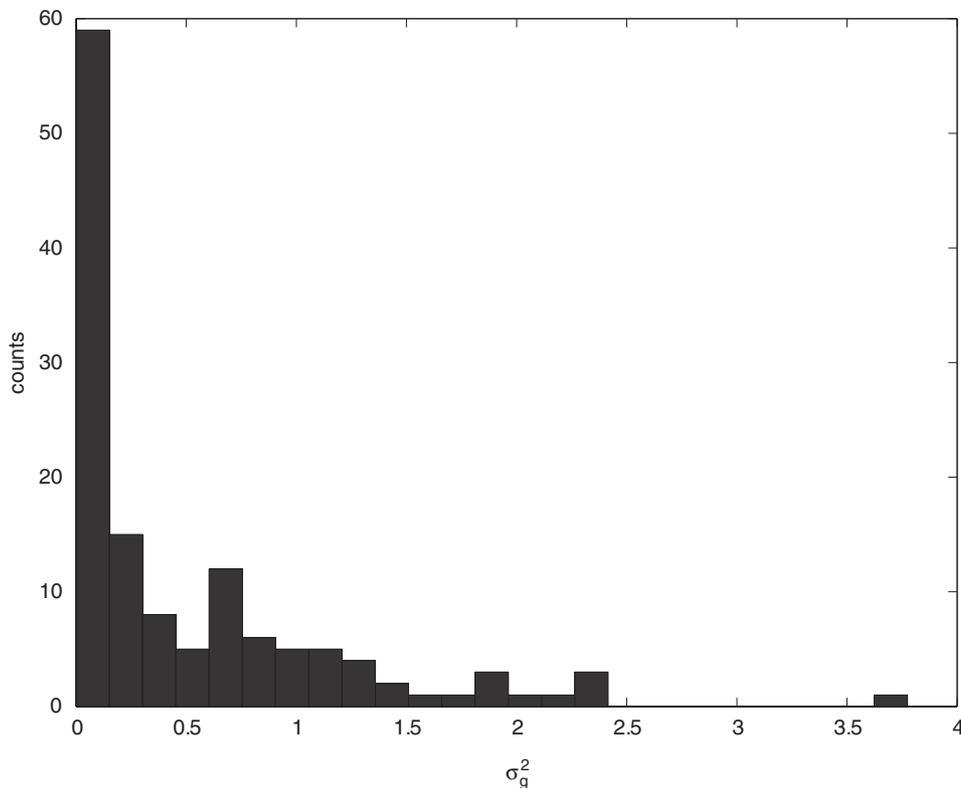


Fig. 1. Histogram of estimates $\hat{\sigma}_g$ for $\mu = -5$, $\sigma_a^2 = 1.0$, $\sigma_g^2 = 1$, and $p = 0.05$.

The robustness of the estimates in the dominant setting is not surprising, and can be explained as follows. For a dominant trait, $P(g_{ij} = 1) = p(2 - p)$ and $\text{Var}(g_{ij}) = p(2 - p)(1 - p)^2$. The correlation between a parent and a child is given by $(1 - p)/(2 - p)$, and the correlation between siblings is $(4 - 3p)/(8 - 4p)$ [Elandt-Johnson, 1971]. For p close to zero, both correlations are nearly $1/2$. The correlation structure for a pedigree for a dominant gene with small p is thus close to the correlation structure of an additive polygenic trait. For a recessive trait, $P(g_{ij} = 1) = p^2$, and $\text{Var}(g_{ij}) = p^2(1 + p)q$ [Elandt-Johnson, 1971]. Here the correlation between a parent and a child is given by $p/(1 + p)$, and the correlation between siblings is $(1 + p)/(4 - 4p)$. For p close to zero, the first correlation is close to zero, while the second one is close to $1/4$. The fact that in the recessive case our additive correlation structure does not fit the data as well as in the dominant case may be the reason that in the recessive case, the random effects model tends to slightly overestimate the true β .

We also estimate μ and σ_a^2 with reasonable accuracy, especially when the disease is less rare, even though those two parameter estimates were obtained via a crude grid search. For example,

when $\mu = -3$, $\sigma_a^2 = 1$, and $\sigma_g = 1$, we obtain $\bar{\mu} = -3.29$ and $\bar{\sigma}_a^2 = 1.06$ when $p = 0.01$, and $\bar{\mu} = -3.16$ and $\bar{\sigma}_a^2 = 1.04$ when $p = 0.05$ in the dominant model. When the disease is rare, $\mu = -5$, the intercept parameters μ and σ_a^2 are underestimated for all settings. For example, in the dominant case with $p = 0.05$ and $\sigma_g = 1$, $\bar{\mu} = -4.59$ and $\bar{\sigma}_a^2 = 0.64$. For results for the other parameter settings, see Table 2. Thus, even though we only observe a highly selected sample of families from the population, the random effects model provides insight into the population prevalence of the disease, whereas conditional logistic regression does not.

The distribution of estimates of σ_g^2 for a dominant trait with $\mu = -5$, $\sigma_g = 1$, and $p = 0.05$ is skewed to the right, and 60% of the estimates of σ_g^2 are zero (Fig. 1). The median estimate of σ_g^2 , 0.22, was less than the mean estimate, 0.5. We observed similar skewness for each of the simulations in Table II. For example, for a recessive trait with $\mu = -5$, $\sigma_g^2 = 1$, and $p = 0.25$, the median was 0.17, while the mean was 0.43. It is evident from histograms such as in Figure 1 that estimates of σ_g^2 will often be zero for almost all situations tested in Table II and provide no evidence for a genetic

random effect. This is not surprising because, as mentioned above, $\sigma_g^2 \text{Var}(g)$ is typically small for Mendelian traits. When the data arose from a dominant gene with allele frequency $p=0.05$ and $\sigma_g = 5.5$, the mean of the estimates for σ_g was 14.44 (excluding one simulation in which $\hat{\sigma}_g$ tended to infinity), and the median estimate was 11.54.

The overall conclusion from the simulation study in Table II is that using the additive covariance structure and normally distributed random effects yields reliable estimates of the fixed-effects parameters $\hat{\beta}_{RE}$, even when the true underlying genetic distribution is discrete and arises from a single gene. The estimator $\hat{\beta}_{CLR}$ also performs well for small or moderate allele frequencies and penetrances in Table II.

DATA EXAMPLE: NPC STUDY

We fitted the two-level random effects model and conditional logistic regression to a subset of NPC data. We studied the effects of sex, age, and the strongest risk factor for NPC, Epstein-Barr seropositivity, on disease risk. Different families required the use of different covariance matrices. These covariance matrices were found using Proc Inbreed, SAS 8.0 [SAS Institute, Inc., 1999].

The covariates we considered were a gender indicator, $X_1 = 1$ for male and 0 for female, and two age-group indicators, $X_2 = 1$ for age 46–57 years, $X_2 = 0$ otherwise, and $X_3 = 1$ for age >57 years, $X_3 = 0$ otherwise. The ≤ 46 -year age-group was the reference group. Age refers to age at diagnosis for cases, and age at interview for controls. X_4 denotes the indicator for Epstein-Barr virus (EBV) seropositivity, defined as an antibody-positive test against one or more of the following four EBV antigens: VCA IgA, EBNA1 IgA, Anti-DNAse, and TK IgA [Connolly, 2001; Hildesheim et al., 2001]. We had to limit our analysis to families in which at least two cases had measurements on EBV exposure. Subjects without EBV status were excluded from the analysis. The problem of missing data affects both the random effects model analysis and the conditional logistic regression analysis. Imputation methods to address missing data would need to be used with caution in our example, because missing values are more likely to occur in older and diseased cases, from whom serum samples could not be obtained. We therefore based our analysis on data from the 38 families with at least two cases with EBV measurements and with a total of 385 subjects. The family sizes ranged from 2–22. Three

TABLE III. Results for estimation of β and σ_g^2 for 38 NPC families, with standard errors in parentheses

Parameter	Two-level random effects model	Conditional logistic regression (CLR)
Male indicator	0.99 (0.31)	0.94 (0.31)
Age 46–57	1.28 (0.26)	1.22 (0.37)
Age >57	0.75 (0.38)	0.74 (0.38)
Epstein Barr	1.74 (0.34)	1.74 (0.38)
σ_g^2	0.00004 (0.001)	NA
Log likelihood	–115.62	–132.26

families had three, and 35 families had two affected members. The estimates and their standard errors are given in Table III.

The two-level random effects model yielded the log odds estimates 0.99 for men, 1.28 for the 46–57-year age group, and 0.75 for the oldest age group. The estimate of random effects variance was $\hat{\sigma}_g^2 = 0.00$, with a standard error of 0.001. The estimates based on conditional logistic regression were very similar, with log odds of 0.94 for men, 1.22 for the 46–57-year age group, and 0.74 for the oldest age group. These findings are consistent with earlier work demonstrating lower risk in women and elevated risk in the 46–57-year age group in Taiwan [Hildesheim and Levine, 1993]. The estimate $\hat{\beta}_{CLR}$ for EBV exposure was 1.74, in perfect agreement with the estimate $\hat{\beta}_{RE}$.

Note that even though the random effects model fits more parameters than conditional logistic regression, it yields smaller standard errors for the β estimates. This may result from the fact that the ascertainment correction used in the random effects model is less stringent (and therefore more realistic), and thus the likelihood contains more information than the conditional logistic regression likelihood.

DISCUSSION

We assessed the robustness of a two-level random effects model for binary disease outcomes in family data that corrects for ascertainment and includes measured covariates as well as random genetic effects that are modeled as polygenes with an additive covariance structure. We were interested in the sensitivity of estimates of the fixed-effects parameters to misspecifications of distribution of underlying genetic liability. In related work, Neuhaus et al. [1992] considered misspecified mixing distributions in logistic-normal models and found that, although regression estimates are asymptotically biased, the magnitude of the

bias is typically small. In support of the conclusions of Neuhaus et al. [1992], Heagerty and Kurland [2001] found that a marginally specified regression structure that is estimated by maximum likelihood is generally not very susceptible to bias resulting from misspecifications of the mixing distribution. Incorrect modeling of random effects in nonlinear mixed models that are not in the class of generalized linear models can have a big impact on the estimates of parameters [Hartford and Davidian, 2000]. Our ascertainment-corrected likelihood does not fall into the framework studied by Neuhaus et al. [1992], Heagerty and Kurland [2001], or Hartford and Davidian [2000].

We simulated the liability from a polygenic model that included a dominant as well as an additive genetic variance component. When the magnitude of the dominant component was less than the magnitude of the additive effect, the bias in the fixed-effects parameters was small. We also studied genetic random effects arising from Mendelian models. When the allele frequencies were small, chosen to reflect the allele frequencies that were expected for nasopharyngeal carcinoma, the model estimates were completely robust to the misspecification. As the allele frequencies increased, a small bias could be detected, but in all situations, even including a dominant model with 60% penetrance, the bias was less than 6%. Thus, modeling g_{ij} as a polygene with additive covariance yields robust estimates of β .

For comparison, we estimated the fixed-effects parameters using standard conditional logistic regression, that completely ignores genetic correlations, and treats each family as a matched set by conditioning on the number of cases in the family. Estimates of the fixed effects based on the simpler model exhibited a stronger bias when the genetic liability was a polygene with a dominant and additive component than the estimates based on the random effects model. As reported by Pfeiffer et al. [2001], conditional logistic regression also leads to estimates of β biased toward zero in the presence of a polygene with additive covariance only. In the Mendelian setting with low allele frequencies and moderate penetrance, the performance of both models was similar with regard to bias of the estimates of fixed effects. In one example of a dominant model with 60% penetrance, however, $\hat{\beta}_{CLR}$ was downwardly biased by 52%, whereas $\hat{\beta}_{RE}$ had only a 6% bias.

An advantage of conditional logistic regression is that the bias is known to be towards the null

[Pfeiffer et al., 2001], while the estimates based on the random effects model may overestimate the magnitude of the fixed effects slightly, as in the case of a Mendelian gene, or underestimate the magnitude of fixed effects, as in the case of a polygene with an additive and a dominant component.

We applied the random effects model and conditional logistic regression to a subset of 38 families from the NPC study, to estimate the effects of age, gender, and exposure to Epstein-Barr virus. The magnitude of random effects variance was estimated to be zero, and the estimates of covariate effects were very similar for the two models. This is not surprising, as the simulations showed that our model might fail to detect an effect of a single gene with low allele frequency and penetrance. An indication in the NPC data that this might be a plausible genetic mechanism is the fact that most families in the study have exactly two affected members, and only very few have more than three cases, as might be expected for a highly penetrant gene [e.g., Bishop, 1999]. Pfeiffer et al. [2001] showed that if genetic random effects are omitted, the estimates of fixed effects from conditional logistic regression are biased toward the null. As the estimates $\hat{\beta}_{RE}$ and $\hat{\beta}_{CLR}$ are virtually identical in the NPC data, and the magnitude of the random effects is estimated to be zero, we conclude that conditional logistic regression using standard software, such as Proc Phreg, SAS 8.0 [SAS Institutes, inc., 1999], would yield valid estimates of exposure effects for these data and is computationally simpler than the random effects model.

In some applications it may be useful to decompose the effect of an individual level covariate X_{ij} into two components, X_i and $X_{ij} - X_i$, in order to estimate "between"- and "within"- cluster exposure effects [Neuhaus and Kalbfleisch, 1998]. In the absence of genetic correlations, CLR estimates the within-cluster effect. In principle it should be possible to estimate both components from the random effects model with data ascertained as in our study, because the likelihood contains some, if little, information on the intercept parameters. When we used the between- and within-family parameterization for Epstein-Barr virus exposure in our data example of 38 NPC families, however, the maximization algorithm did not converge. A larger data set might provide adequate information to estimate both components.

It might be worth exploring the use of regressive logistic models [Bonney, 1986] to account for

familial correlations, provided a suitable ascertainment correction is available. The interpretation of β in these models depends on family structure and size, however, unlike model (1).

In conclusion, if the disease of interest is rare and includes an unmeasured genetic component that is small or, in the case of a single gene, has low allele frequency and low penetrance, then treating the family members as independent within a given family and relying on conditional logistic regression will result in nearly unbiased estimates of the fixed-effects parameters. If the disease of interest is less rare, and is based on polygenic effects of moderate size, or if a highly penetrant autosomal-dominant gene is present, then using the random effects model is preferable to the conditional logistic analysis, and the estimates of fixed effects will be robust to misspecifications of the underlying genetic mechanism.

APPENDIX

In the absence of covariates, i.e., $\beta = 0$, model (1) reduces to $\text{logit}P(Y_{ij} = 1|a_i, g_{ij}) = \mu + \sigma_a a_i + \sigma_g g_{ij}$. The prevalence of disease in the population is given by

$$P(Y_{ij} = 1) = \int_a \int_g P(Y_{ij} = 1|a_i, g_{ij})dF(a)dF(g).$$

When the genetic component is Mendelian, the integration with respect to the distribution of g is replaced by a summation over the appropriate genotypes. Recall that $D_{ij} = 0, 1, 2$ denotes the number of alleles D with allele frequency p that an individual is carrying. For a dominant trait, the score function $g(D_{ij}) = 1$ for $D_{ij} = 2$ or $D_{ij} = 1$, and 0 otherwise. The overall disease probability for the dominant model is

$$P(Y_{ij} = 1) = \int_a P(Y_{ij} = 1|a_i, g_{ij} = 1)(p^2 + 2p(1 - p)) + P(Y_{ij} = 1|a_i, g_{ij} = 0)(1 - p)^2 dF(a),$$

the baseline penetrance for a given set of parameters is $\int_a P(Y_{ij} = 1|a_i, g_{ij} = 0)dF(a)$, and the penetrance for a carrier is $\int_a P(Y_{ij} = 1|a_i, g_{ij} = 1)dF(a)$. The population attributable risk is

$$AR = 1 - \frac{\int P(Y_{ij} = 1|a_i, g_{ij} = 0)dF(a)}{(p^2 + 2pq) \int P(Y_{ij} = 1|a_i, g_{ij} = 1)dF(a) + (1 - p)^2 \int P(Y_{ij} = 1|a_i, g_{ij} = 0)dF(a)},$$

where $q = 1 - p$. The calculations are similar for additive and recessive traits.

REFERENCES

Bonney GE. 1986. Regressive logistic model for familial disease and other binary traits. *Biometrics* 42:611–25.

Bishop DT. 1999. BRCA1 and BRCA2 and breast cancer incidence: a review. *Ann Oncol* 10:113–9.

Butler SM, Louis TA. 1992. Random effects models with nonparametric prior. *Stat Med* 11:1981–2000.

Connolly Y. 2001. Antibodies to EBV thymidine kinase: a characteristic marker for the serological detection of NPC. *Int J Cancer* 91:692–7.

Elandt-Johnson RC. 1971. Probability models and statistical methods in genetics. New York: Wiley.

Fisher RA. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinb* 52:399–433.

Hartford A, Davidian M. 2000. Consequences of misspecifying assumptions in nonlinear mixed effects models. *Comput Stat Data Ana* 34:139–64.

Heagerty PJ, Kurland BF. 2001. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* 88:973–85.

Hildesheim A, Levine PH. 1993. Etiology of nasopharyngeal carcinoma—a review. *Epidemiol Rev* 15:466–85.

Hildesheim A, Dosemeci M, Chan CC, Chen CJ, Cheng YJ, Hsu MM, Chen IH, Mittl BF, Sun B, Levine PH, Chen JY, Brinton LA, Yang CS. 2001. Occupational exposure to wood, formaldehyde, and solvents and risk of nasopharyngeal carcinoma. *Cancer Epidemiol Biomarkers Prev* 10:1145–53.

Kraft P, Thomas DC. 2000. Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. *Am J Hum Genet* 66:1119–31.

Muthen B, Shedden K. 1999. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* 55:463–9.

Neuhaus JM, Kalbfleisch JD. 1998. Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics* 54:638–45.

Neuhaus JM, Hauck WW, Kalbfleisch JD. 1992. The effects of mixture distribution misspecification when fitting mixed effects logistic models. *Biometrika* 79:755–62.

O'Neill TJ, Barry SC. 1995. Truncated logistic regression. *Biometrics* 51:533–41.

Pfeiffer RM, Gail MH, Pee D. 2001. Inference for covariates that accounts for ascertainment and random genetic effects in family studies. *Biometrika* 88:933–48.

SAS Institute, Inc. 1999. SAS release 8.0. Cary, NC: SAS Institute, Inc.

Verbeke G, Lesaffre E. 1997. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Comput Stat Data Anal* 23:541–56.