**Genetic Epidemiology 16:261–273 (1999)**

# Case-Parents Design for Gene-Environment Interaction

**Daniel J. Schaid***

*Departments of Health Sciences Research and Medical Genetics,
Mayo Clinic/Mayo Foundation, Rochester, Minnesota*

The scientific and public health implications of gene-environment interaction warrant that the most powerful study designs and methods of analysis be used. Because traditional case-control designs, which use nonrelated subjects, have demonstrated the need for large samples to detect interactions, alternative study designs may be worthwhile, such as sampling diseased cases and their parents. If the transmission of particular alleles from parents to their diseased child appears to be distorted from Mendelian expectation, then this suggests an etiologic association of the alleles with disease; if the frequency of transmission differs between exposed and nonexposed cases, then gene-environment interaction is suggested. We present likelihood-based methods to assess interaction, as well as an extension of the transmission/disequilibrium test (*TDT*). For these statistical tests, we also derive methods to compute sample size and power. Comparisons of sample size requirements between the case-parents design and the case-control design indicate that the case-parents design can be more powerful to detect gene-environment interactions, particularly when the disease susceptible allele is rare. Also, one of the derived likelihood methods, based on additive effects of alleles, tended to be the most robust in terms of power for a broad range of genetic mechanisms, and so may be useful for broad applications to assess gene-environment interactions. Genet. Epidemiol. 16:261–273, 1999. © 1999 Wiley-Liss, Inc.

**Key words: association; conditional logistic regression; genotype relative risk; likelihood; transmission disequilibrium**

Contract grant sponsor: NIH; Contract grant number: GM51256.

*Correspondence to: Dr. Daniel J. Schaid, Harwick 7, Mayo Clinic, 200 First Street S.W., Rochester, MN 55905. E-mail: schaid@mayo.edu

Received 21 January 1998; Revised 26 May 1998; Accepted 26 June 1998

## INTRODUCTION

Recent advances in molecular genetics give hope that the influence of genetic susceptibility on phenotypic response to environmental risk factors can now be evaluated [Perera, 1997]. Gene-environment interactions, which we denote GxE, have important scientific and public health implications: (1) the estimated genetic effects depend on the environmental conditions that exist for any particular study, limiting inference of genetic effects to only the range of environments studied, and vice versa, and (2) interactions can imply dramatically increased (or decreased) risks to individuals, depending on their genotype and environment.

The definition of GxE is critical, yet it is often left ambiguous; sometimes as biologic interaction, and sometimes as statistical interaction. Biological interaction can be defined as a specific form of interdependent operation of a gene and an environmental risk factor to produce disease [Rothman et al., 1980]. From an epidemiologic viewpoint, mechanistic models of biological interaction have been proposed [Ottman, 1990]. It is important to recognize that interaction depends on a model, either a biologic mechanistic model or a statistical model. Although interactions are probably easiest to detect when a mechanistic model is assumed, and animal models may offer insights [Knudson and Hino, 1997], this is a rare situation for human studies because of our current limited knowledge. In the absence of such a mechanistic model, or when generalizing genetic effects to populations, or when developing models to predict disease risk, statistical models of interaction can be useful. Note that statistical interactions depend on the scale of measurement. Statistical interaction means that the joint effects of genetic and environmental factors cannot simply be added, if an additive model is assumed, or cannot simply be multiplied, if a multiplicative model is assumed.

Statistical issues relating to GxE can be complex [MacMahon, 1968; Greenland, 1993], and the sample size required to have adequate power can be so large that traditional case-control studies are sometimes not feasible [Greenland, 1983; Goldstein et al., 1997], unless the interaction effect is large and the genetic and environmental risk factors are both common [Hwang et al., 1994; Khoury et al., 1995]. Alternative epidemiologic designs have been proposed to assess GxE [Khoury and Flanders, 1996], and in particular the use of diseased cases and their parents [Khoury, 1994; Khoury and Flanders, 1996; Schaid and Sommer, 1994]. In the absence of GxE, using cases and their parents has proven to be powerful to assess disease associations with candidate genes, as well as assessment of both linkage and allelic association with genetic markers. The transmission/disequilibrium test (*TDT*) proposed by Spielman et al. [1993] is a popular method to assess these types of associations, and Risch and Merikangas [1996] have recently evaluated the power of this type of design/analysis to perform genome-wide association tests. The power to detect associations can be improved by using likelihood methods of analysis, and the sample size requirements can be estimated for more general models [Schaid, 1999] than those proposed by Risch and Merikangas [1996]. However, the feasibility of the case-parents design to assess GxE has not been evaluated. The purposes of this paper are to derive methods to determine sample size and power for the *TDT* and likelihood methods to detect GxE, and then compare the sample size requirements for different methods of analysis for specific alternative hypotheses of GxE. To examine the feasibility of different designs to assess GxE, we compare the required sample size of the case-parents design with the traditional case-control design that uses nonrelated subjects.

## STATISTICAL METHODS

### Likelihood Methods

To illustrate the likelihood methods, consider the distribution of genotypes among the diseased children (denoted cases), conditional on their parents' genotypes. Because it is necessary to condition on sampling through a case, the distribution of genotypes among the cases depends on genotype relative risks, where an arbitrary genotype is chosen as the baseline. Schaid and Sommer [1993] presented likelihood methods to estimate the genotype relative risks and to test for association of the genetic marker with disease. Extension of these likelihood methods to assess GxE can be performed by noting that GxE implies that the genotype relative risk parameters vary over the levels of the environmental risk factor. For illustrative purposes, we shall consider a binary environmental risk factor ($X_E = 1$ if exposed; $X_E = 0$ if not exposed), and a genetic marker with two alleles (*A*, the putative high-risk allele, and *a* the low-risk allele), although more general models can be used [Schaid, 1996], as discussed later. There are four genotype relative risks when considering exposure status: $r_1^+$ and $r_2^+$ for the exposed cases having genotypes *Aa* and *AA*, respectively (with $r_0^+ = 1$ for the exposed baseline genotype *aa*), and $r_1^-$ and $r_2^-$ for the nonexposed cases having genotypes *Aa* and *AA*, respectively (again with $r_0^- = 1$ for the nonexposed baseline genotype *aa*). Note that exposed and nonexposed genotypes each have their own baseline. This is because the main effects of environmental risk factors cannot be estimated by using the case-parents design, due to the implicit stratification on each nuclear family, but only the interaction of environmental risk factors with susceptible genotypes can be estimated [Self et al., 1991; Schaid, 1995].

When cases are independent, as when there is only a single case per family, the likelihood ratio statistic to test for GxE is $LR = 2[\ln L^+(\hat{r}_1^+, \hat{r}_2^+) + \ln L^-(\hat{r}_1^-, \hat{r}_2^-) – \ln L^T(\hat{r}_1, \hat{r}_2)]$, where $\ln L^+(\hat{r}_1^+, \hat{r}_2^+)$ is the maximized loglikelihood for the subset of exposed cases, $\ln L^-(\hat{r}_1^-, \hat{r}_2^-)$ is that for the nonexposed cases, and $\ln L^T(\hat{r}_1, \hat{r}_2)$ is that for the pool of all cases. Because the test for GxE is testing simultaneously whether $r_1^+ = r_1^-$ and $r_2^+ = r_2^-$, the asymptotic chi-square distribution of *LR* has two degrees-of-freedom. Alternatively, a specific genetic model could be assumed to reduce the degrees-of-freedom to one, with the intent of increasing power to detect GxE.

Under a specific genetic model, $\ln L(r_1,r_2)$ can be maximized subject to the constraints of the model. For example, multiplicative allele effects ($r_1 = r$, $r_2 = \mathrm{r}^2$), additive effects ($r_1 = r$, $r_2 = 2r – 1$), dominant effects ($r_1 = r_2 = r$), or recessive effects ($r_1 = 1$, $r_2 = r$) can be assumed, and maximum likelihood estimates can be easily computed [Schaid, 1999]. One can then test for GxE, subject to the constraints of an assumed genetic relative risk model.

As an example, consider the multiplicative model, which implies that $r_1^+ = r^+$, $r_2^+ = (r^+)^2$ for the exposed cases, and $r_1^- = r^-$, $r_2^- = (r^-)^2$, for the nonexposed cases. One can then test for GxE by testing whether $r^+ = r^-$ with a single degree of freedom. Note that the logistic regression model $\log[\pi/(1 – \pi)] = \beta_A + \beta_I X_E$ has been proposed to assess GxE [Maestri et al., 1997], where $\pi$ is the probability that a heterozygous

*Aa* parent transmits an *A* allele to their diseased child. Here, the dependent variable is $Y^{TDT} = 1$ if an *A* allele is transmitted from a heterozygous parent to diseased case; $Y^{TDT} = 0$ if *A* is not transmitted. The regression coefficients are log relative risks: $r_A = e^{\beta_A}$ is the relative risk associated with each *A* allele, and $r_1 = e^{\beta_I}$ is a multiplicative interaction parameter that summarizes the effect of environmental exposure on each *A* allele. The relationship between this model, which considers alleles, and our likelihood method, which considers genotypes, is as follows:

$$r_1^- = r_A,$$
$$r_2^- = r_A^2,$$
$$r_1^+ = r_A r_E,$$
$$r_2^+ = r_A^2 r_E^2.$$

This illustrates that modeling transmission of alleles using the logistic model implicitly assumes much larger relative risks for homozygous *AA* cases, with even more dramatic relative risks for those that are exposed. If the risk of exposure for *AA* homozygotes is not as large as predicted by this model, then the risk associated with allele *A* ($\beta_A$), and the interaction parameter ($\beta_I$), can be underestimated, with the amount of bias depending on the frequency of *AA* homozygotes and the frequency of exposure. Alternatively, any of the other genetic relative risk models can be assumed, with allowance for assessing GxE.

### *TDT* Method

Because the *TDT* method of analysis evaluates the probability that an *Aa* parent transmits the *A* allele to an affected child, GxE is suggested when the observed transmission probability differs between exposed and nonexposed cases. The statistical test for interaction can be based on the statistic

$$z = \frac{\hat{\pi}^+ - \hat{\pi}^-}{\sqrt{\hat{\pi}(1-\hat{\pi})\left[\frac{1}{n^+} + \frac{1}{n^-}\right]}},$$

where $\hat{\pi}^+$ and $\hat{\pi}^-$ are the estimated transmission probabilities for the exposed and nonexposed cases, respectively, $\hat{\pi}$ is that for the pool of all cases, and $n^+$ and $n^-$ are the number of heterozygous *Aa* parents of exposed and nonexposed cases, respectively. The statistic *z* has an approximate standard normal distribution.

## SAMPLE SIZE AND POWER

### Likelihood Methods

To derive sample size and power to detect GxE for an assumed genetic-environmental model, using any of the likelihood ratio statistics or the *TDT* method, we need to consider the distribution of the genotypes of the cases, conditional on their exposure status and their parents' genotypes. The general formulas for these probabilities, not considering exposure status, are given in Table I. Assuming that allele frequencies do not depend on exposure status, similar results pertain to exposed and nonexposed cases, except that the relative risk parameters in Table I are replaced with $r^+$ for exposed cases and $r^-$ for nonexposed cases.

To determine sample size and power in a general manner, we consider the relative risks for an alternative hypothesis (which we shall refer to as the "true" model), and a model for analysis (referred to as the "assumed" model). For an assumed model (with relative risks $r_1^{\,+}$, $r_2^{\,+}$, $r_1^-$, and $r_2^-$), which may or may not agree with the "true" model (with relative risks $t_1^{\,+}$, $t_2^{\,+}$, $t_1^-$, and $t_2^-$), we need to maximize the *expected* log-likelihoods among the expected number of exposed cases ($\ln L^{+*}(\hat{r}_1^+, \hat{r}_2^-)$), nonexposed cases ($\ln L^{-}*(\hat{r}_1^-, \hat{r}_2^-)$), and the total of all cases ($\ln L^{T*}(\hat{r}_1, \hat{r}_2)$). Using these, note that *LR* has an approximate non-central chi-square distribution with non-centrality parameter $\lambda = N\delta^2$, where $N\delta^2 = 2[\ln L^{+*}(\hat{r}_1^+, \hat{r}_2^+) + \ln L^{-}*(\hat{r}_1^-, \hat{r}_2^-) - \ln L^{T*}(\hat{r}_1, \hat{r}_2)]$.

The expected log-likelihood among the expected exposed cases can be written as

$$lnL^{+*}(r_1^+, r_2^+) = \sum_{i=2,4,5} \sum_{j=0,1,2} n_{ij}^{+*} log\, P_{j|i}(r_1^+, r_2^+),$$

where $P_{j|i}\ (r_1^+, r_2^+)$ is the conditional probability of the $j^{th}$ genotype category for the case, given the $i^{th}$ parental mating type for the assumed model of genotype relative risks (one of columns 5–9 of Table I); $n_{ij}^{+*} = NP_EP_{ij}\ (\mathrm{t}_1^+, t_2^+)$, the expected number of exposed cases in the $ij^{th}$ genotype category under the true model; $P_E$ is the probability of exposure; and $P_{ij}\ (t_1^+, t_2^+)$ is the unconditional true probability of the $ij^{th}$ genotype category for exposed cases. These unconditional probabilities can be computed by using $t_1^+$ and $t_2^+$ to compute the probability of the parental mating type (column 2 of Table I), and then multiplying by the probability of the child's genotype, given parental mating type $[P_{j|i}\ (t_1^+, t_2^+)$ - one of the columns 5–9 of Table I]. The expected log-likelihood among the nonexposed, $\ln L^{-*}(r_1^-, r_2^-)$, can be computed in an analogous manner, using relative risks for nonexposed and $n_{ij}^{-*} = N(1 - P_E)P_{ij}(t_1^-, t_2^-)$. The expected log-likelihood among all cases, $\ln L^{T*}(r_1, r_2)$, can also be computed in this manner by using

$$n_{ij}^* = N[P_E P_{ij}(t_1^+, t_2^+) + (1 - P_E)P_{ij}(t_1^-, t_1^-)]$$

To maximize the expected log-likelihood for an assumed model, replace the $n_{ij}$ values with their expected values in the expressions for the maximum likelihood estimates [i.e., expressions 2, 4–6 in Schaid, 1999]. For the multiplicative, additive, dominant, and recessive models, each having 1 degree-of-freedom, power or sample size can be determined by solving the equation $z_\beta = \sqrt{N\delta} - z_\alpha$ for either $z_\beta$ or $N$, respectively, where $z_\alpha$ and $z_\beta$ are the $(1\text{-}\alpha)^{th}$ and $(1\text{-}\beta)^{th}$ percentiles of a standard normal distribution, giving Type-I error of $\alpha$ and power of $(1\text{-}\beta)$. The sample size, $N$, is the total number of cases required. This method allows one to specify any type of true genetic relative risk, and then evaluate the power of the likelihood ratio statistic for an assumed genetic model. In other words, the power of the *LR* can be evaluated for both correct and misspecified genetic models.

**TABLE I. Genotype Relative Risk Models and Probablities of Genotypes for a Diseased Child (Case)**

| | | | | P (case genotype, given mating type) = $P_{j\|i}(r_1,r_2)$ | | | | |
|---|---|---|---|---|---|---|---|---|
| Parental mating type | P (mating type)[a] | Case genotype | $Y^{TDT}$ | General arbitrary $r_1,r_2$ | Multiplicative[b] $r_1 = r$, $r_2 = r^2$ | Additive $r_1 = r$, $r_2 = 2r–1$ | Dominant $r_1 = r_2 = r$ | Recessive $r_1 = 1$, $r_2 = r$ |
| 1. AA × AA | $p^4\frac{r_2}{R}$ | AA | — | 1 | 1 | 1 | 1 | 1 |
| 2. AA × Aa | $2p^3q\,\frac{(r_1 + r_2)}{R}$ | AA | 1 | $r_2/(r_1 + r_2)$ | $\pi$ | $(2r – 1)/(3r – 1)$ | ½ | $r/(r + 1)$ |
| | | Aa | 0 | $r_1/(r_1 + r_2)$ | $(1 – \pi)$ | $r/(3r –1)$ | ½ | $1/(r + 1)$ |
| 3. AA × aa | $2p^2q^2\frac{r_1}{R}$ | Aa | — | 1 | 1 | 1 | 1 | 1 |
| 4. Aa × Aa | $p^2q^2\frac{(r_2 + 2r_1 +1)}{R}$ | AA | 1,1 | $r_2/(r_2 + 2r_1 + 1)$ | $\pi^2$ | $(2r - 1)/(4r)$ | $r/(3r + 1)$ | $r/(r +3)$ |
| | | Aa | 1,0 | $2r_1/(r_2 + 2r_1 +1)$ | $2\pi(1 – \pi)$ | 1/2 | $2r/(3r + 1)$ | $2/(r + 3)$ |
| | | aa | 0,0 | $1/(r_2 + 2r_1 + 1)$ | $(1 – \pi)^2$ | $1/(4r)$ | $1/(3r + 1)$ | $1/(r + 3)$ |
| 5. Aa × aa | $2pq^3\,\frac{(r_1 + 1)}{R}$ | Aa | 1 | $r_1/(r_1 + 1)$ | $\pi$ | $r/(r + 1)$ | $r/(r + 1)$ | 1/2 |
| | | aa | 0 | $1/(r_1 + 1)$ | $(1 –\pi)$ | $1/(r + 1)$ | $1/(r + 1)$ | 1/2 |
| 6. aa × aa | $q^4\,\frac{1}{R}$ | aa | — | 1 | 1 | 1 | 1 | 1 |

[a] $R = r_2p^2 + r_12pq + q^2$.
[b] $\pi = r/(r + 1)$.

### *TDT* Method

For the *TDT* method, power and sample size can be calculated by solving the following equation for $z_\beta$ or $N$:

$$\sqrt{N}|\pi^+ - \pi^-| = z_\alpha \sqrt{V_o} + z_\beta \sqrt{V_I}$$

where $\pi^+$ and $\pi^-$ are the expected transmission probabilities under the true model, and the variance of ($\hat{\pi}^+ - \hat{\pi}^-$) is $V_o/N$ when there is no interaction, and $V_I/N$ when there is interaction. Again, the sample size $N$ is the total number of cases required, so that the total number of cases and parents to genotype is $3N$.

To compute $\pi^+$ and $\pi^-$ under a genetic model, the methods in Schaid [1999] can be used. Briefly, using the transmission indicator variable $Y^{TDT}$, as illustrated in column 4 of Table I, the expected value of $\hat{\pi}^+$ among the exposed cases can be determined by writing

$$\hat{\pi}^+ = \frac{\sum Y_i^{TDT}}{n^+}, \tag{1}$$

where the summation in the numerator is over all the informative transmissions, as indicated in Table I. For a true genetic relative risk model, the expected value of $\hat{\pi}^+$ can be approximated by replacing the values of the numerator and denominator of expression (1) with their expected values. Consider a sample size of $N = 1$ diseased case. The expected value of the numerator can be determined by multiplying $\{P_E P_{ij}\ (t_1^+, t_2^+)\}$ times the $Y^{TDT}$ scores in Table I, and then summing over all informative categories. The sum of the values of $\{P_E P_{ij}\ (t_1^+, t_2^+)\}$ over all informative categories, with a weight of 2 when both parents are heterozygous, gives the expected value of the denominator, the number of heterozygous parents of an exposed case ($h^+$). A similar method can be used to compute $\pi^-$ and $h^-$. The variances factors can then be computed by

$$V_I = \frac{\pi^+(1-\pi^+)}{h^+} + \frac{\pi^-(1-\pi^-)}{h^-}, \text{ and } V_o = \pi(1-\pi)\left[\frac{1}{h^+} + \frac{1}{h^-}\right],$$

where $\pi$ is computed under the null hypothesis; simulations (now shown) have indicated that computing $h^+$ and $h^-$ under the alternative for $V_o$, rather than under the null hypothesis, give more accurate estimates of sample size and power.

### Case-Control Design for Nonrelated Subjects

For a traditional case-control design, statistical interaction can be assessed using logistic regression. Similar to the likelihood models used for the case-parents design, we can use an assumed genetic model (e.g., multiplicative, additive, dominant, recessive) with a single parameter representing genetic risk, so that the test for interaction requires a single degree-of-freedom. For example, if allelic effects are multiplicative, the logistic regression model for the case-control design is

$$logit = \beta_0 + \beta_A X_G + \beta_E X_E + \beta_I X_I,$$

where $X_G$ has values 0, 1, or 2 for genotypes *aa*, *Aa*, or *AA*, respectively; $X_I = X_G \times X_E$; $\beta_A$ is the log odds-ratio for the *A* allele; $\beta_E$ is the log odds-ratio for exposure; and $\beta_I$ is the log odds-ratio representing GxE interaction. The statistical test for interaction can be based on the likelihood ratio statistic $LR = 2[\ln L(\hat{\boldsymbol{\beta}}) – \ln L(\hat{\boldsymbol{\beta}}_{\mathbf{o}})]$, where $\hat{\boldsymbol{\beta}}$ is the vector of maximum likelihood estimates allowing for interaction, and $\hat{\boldsymbol{\beta}}_{\mathbf{o}}$ is that for when $\beta_I = 0$.

For large samples, *LR* has an approximate non-central chi-square distribution with one degree-of-freedom and with noncentrality parameter

$$N\delta^2 = 2[\ln L^*(\hat{\beta}) - \ln L^*(\hat{\beta}_o)], \tag{2}$$

where $\ln L^*(\hat{\boldsymbol{\beta}})$ is the maximized *expected* log-likelihood allowing for interaction, and $\ln L^*(\hat{\boldsymbol{\beta}}_o)$ is that without interaction. To compute $\ln L^*(\boldsymbol{\beta})$ note that

$$ln\, L(\beta) = \sum Y_i X_i' \beta - ln(1 + e^{X_i'\beta}) \tag{3}$$

where $Y_i = 1$ if case ($Y_i = 0$ if control), $\mathbf{X}_i$ is the vector of covariates, and the summation is over all subjects. To evaluate power of the *LR* statistic under correct and misspecified models, we shall let $\mathbf{T}$ be a vector of covariates for the true model, with corresponding coefficients $\boldsymbol{\gamma}$, and let $\mathbf{X}$ and $\boldsymbol{\beta}$ be the corresponding vectors for an assumed model. Taking the expectation of expression (3) over all values of *Y* given $\mathbf{T}$, and then over the distribution of $\mathbf{T}$ vectors, results in

$$ln\, L^*(\beta) = N \sum_{\mathbf{T}} \left[ P_{Y|\mathbf{T}}(\mathbf{T}'\gamma)\mathbf{X}'\beta - ln(1 + e^{\mathbf{X}'\beta}) \right] P(\mathbf{T}),$$

where $P_{Y|\mathbf{T}}(\mathbf{T}'\boldsymbol{\gamma}) = e^{\mathbf{T}'\gamma}/(1 + e^{\mathbf{T}'\gamma})$, $P(\mathbf{T})$ is the population probability of vector $\mathbf{T}$, and the summation is over all possible $\mathbf{T}$ vectors. Assuming that the distributions of genotypes and exposure are independent in the population, $P(\mathbf{T})$ can be determined by the product of probabilities of exposure status and genotype, as illustrated in Table II for a multiplicative genetic model. To compute genotype probabilities, we assume Hardy-Weinberg proportions of $p^2$, $2pq$, and $q^2$ for genotypes *AA*, *Aa*, and *aa*, where *p* and *q* are the population frequencies of alleles *A* and *a*, respectively. For the case-control design, the true intercept parameter, $\gamma_o$, is determined by the fraction of cases

**TABLE II. Coding Genotypes for a Multiplicative Model and the Distribution of T Vectors**

| Exposure status | Genotype | $\mathbf{T}' = T_G$ | $T_E$ | $T_I$ | $P(\mathbf{T})$ |
|---|---|---|---|---|---|
| – | aa | 0 | 0 | 0 | $(1 - P_E)q^2$ |
| | Aa | 1 | 0 | 0 | $(1 - P_E)2pq$ |
| | AA | 2 | 0 | 0 | $(1 - P_E)p^2$ |
| + | aa | 0 | 1 | 0 | $P_E q^2$ |
| | Aa | 1 | 1 | 1 | $P_E 2pq$ |
| | AA | 2 | 1 | 2 | $P_E p^2$ |

in the sample ($f$), and the remaining regression coefficients, as determined by the following expression:

$$f = \sum_{\mathbf{T}} P_{Y|\mathbf{T}}(\mathbf{T}'\gamma)P(\mathbf{T}). \qquad (4)$$

For specified values of $\gamma_A$, $\gamma_E$, and $\gamma_I$, we solved for $\gamma_o$ in expression (4) by the bisect algorithm. Then, with $\boldsymbol{\gamma}$ specified, we maximized ln $L^*$ ($\boldsymbol{\beta}$) over the values of $\boldsymbol{\beta}$ by the simplex method [Press et al., 1992]. This method allows computation of $\delta^2$ in expression (2), so that power or sample size can be determined by solving $z_\beta = \sqrt{N}\,\delta - z_\alpha$ for either $z_\beta$ or $N$, respectively. Note that $N$ is the total sample size, so that the total number of required cases is $fN$.

## RESULTS

The sample sizes needed to achieve 80% power to detect various levels of GxE, for one-sided tests having Type-I error of 5%, were computed for a variety of genetic and environmental models. The $A$ allele frequency varied over $p$ = .01, .10, .50; the frequency of exposure was fixed at 30%; the odds-ratio associated with exposure in the absence of genetic risk, $OR_{Env}$, varied over values of 1, 2, and 4 (this odds-ratio only affects sample sizes for the case-control design using logistic regression, because the main effects of exposure do not affect the case-parents design); among the nonexposed, the true relative risk models were multiplicative ($r_1^- = r$, $r_2^- = r^2$), additive ($r_1^- = r$, $r_2^- = 2r - 1$), dominant ($r_1^- = r_2^- = r$), and recessive ($r_1^- = 1$, $r_2^- = r$), with $r$ having values of 2 and 4; among the exposed, the relative risks for the high-risk genotypes were determined by a multiplicative relative risk interaction ($r_{\text{int}}$) such that $r_1^+ = r_{\text{int}}r_1^-$ and $r_2^+ = r_{\text{int}}r_2^-$, with the exception for the recessive model for which $r_1^+ = r_1^- = 1$, implying that only $AA$ homozygotes have an increased risk of disease, albeit modified by environmental exposure. The magnitude of the interaction was $r_{\text{int}}$ = 2 or 4. For the case-control design, the odds-ratios for exposed are also inflated by a factor of $OR_{Env}$. For these 144 combinations of parameters, sample sizes were computed for the case-parents design (using the statistics $TDT$, $LR_{Mul}$, $LR_{Add}$, $LR_{Dom}$, $LR_{Rec}$, where $LR$ denotes the likelihood ratio statistic to test GxE, and the subscript indicates the assumed model for analysis) and for the case-control design (using logistic regression likelihood ratio statistics $Logit_{Mul}$, $Logit_{Dom}$, $Logit_{Rec}$, where the subscript indicates the assumed model for analysis). For the case-parents design, $N_{CP}$ is the estimated number of cases (the total sample size, including parents, is $3N_{CP}$). For the case-control design we assumed equal number of cases and controls (total sample size = $2N_{CC}$, where $N_{CC}$ is the number of cases needed for the case-control design).

The total sample sizes required by the case-parents design and those required by the case-control design were compared by computing relative efficiencies: the ratio of the required sample size divided by the smallest required sample size for all eight statistical tests. A portion of these results are presented in Table III for when $OR_{Env}$ = 2 and the interaction was $r_{\text{int}}$ = 4. In Table III, the relative efficiencies are presented, with values of 1 indicating the most efficient method, and values greater than 1 indicating the amount of inflation of the sample size needed to achieve the same power as the most efficient method. A number of key points are demonstrated

**TABLE III. Relative Efficiencies of Statistics to Detect Gene-Environment Interaction: Case-Parents Design and Nonrelated Case-Control Design**

| | | | | | | | Relative efficiencies, compared to smallest sample size[b] | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Genotype relative risks[a] | | | | Case-parents | | | | | Case-control | | | $N_{CP}$ cases for |
| $p$ | $r$ | Model | $r_1^-$ | $r_2^-$ | $r_1^+$ | $r_2^+$ | $TDT$ | $LR_{Mul}$ | $LR_{Add}$ | $LR_{Dom}$ | $LR_{Rec}$ | $Logit_{Mul}$ | $Logit_{Dom}$ | $Logit_{Rec}$ | $LR_{Add}$ |
| 0.01 | 2 | mul | 2 | 4 | 8 | 16 | 1.26 | 1.04 | 1 | 1 | 2809.01 | 2.36 | 2.36 | 809.26 | 963 |
| | | add | 2 | 3 | 8 | 12 | 1.26 | 1.04 | 1 | 1 | 3312.99 | 2.36 | 2.36 | 645.71 | 962 |
| | | dom | 2 | 2 | 8 | 8 | 1.26 | 1.04 | 1 | 1 | 4322.94 | 2.35 | 2.36 | 477.75 | 962 |
| | | rec | 1 | 2 | 1 | 8 | 10.51 | 10.51 | 12.06 | 163.99 | 1 | [c] | [c] | 6.03 | 999,963 |
| | 4 | mul | 4 | 16 | 16 | 64 | 1.42 | 1.08 | 1 | 1 | 4157.45 | 4.04 | 4.04 | [c] | 941 |
| | | add | 4 | 7 | 16 | 28 | 1.42 | 1.08 | 1 | 1 | 6048.55 | 4.03 | 4.04 | 1388.33 | 938 |
| | | dom | 4 | 4 | 16 | 16 | 1.42 | 1.08 | 1 | 1 | 8571.46 | 4.01 | 4.04 | 924.83 | 938 |
| | | rec | 1 | 4 | 1 | 16 | 3.59 | 3.59 | 4.70 | 53.79 | 1 | [c] | [c] | 12.62 | 304,958 |
| 0.1 | 2 | mul | 2 | 4 | 8 | 16 | 1.52 | 1.39 | 1.01 | 1 | 291.15 | 1.23 | 1.23 | 41.33 | 204 |
| | | add | 2 | 3 | 8 | 12 | 1.52 | 1.40 | 1 | 1 | 346.68 | 1.22 | 1.23 | 33.45 | 201 |
| | | dom | 2 | 2 | 8 | 8 | 1.54 | 1.42 | 1 | 1 | 462.18 | 1.20 | 1.24 | 25.65 | 197 |
| | | rec | 1 | 2 | 1 | 8 | 1.59 | 1.58 | 2.95 | 20.92 | 1 | 57.73 | 250.71 | 4.73 | 2,817 |
| | 4 | mul | 4 | 16 | 16 | 64 | 1.94 | 1.74 | 1.01 | 1 | 417.08 | 1.23 | 1.24 | 75.40 | 301 |
| | | add | 4 | 7 | 16 | 28 | 1.97 | 1.79 | 1 | 1 | 624.74 | 1.26 | 1.31 | 38.43 | 278 |
| | | dom | 4 | 4 | 16 | 16 | 1.99 | 1.81 | 1 | 1 | 910.81 | 1.24 | 1.32 | 24.91 | 270 |
| | | rec | 1 | 4 | 1 | 16 | 1.03 | 1 | 2.36 | 10.45 | 1.28 | 813.84 | [c] | 12.25 | 1,459 |
| 0.5 | 2 | mul | 2 | 4 | 8 | 16 | 11.78 | 11.56 | 3.33 | 3.07 | 308.63 | 1.04 | 1 | 6.95 | 491 |
| | | add | 2 | 3 | 8 | 12 | 11.57 | 11.41 | 3.14 | 2.86 | 359.13 | 1 | 1 | 6.26 | 439 |
| | | dom | 2 | 2 | 8 | 8 | 11.63 | 11.54 | 2.95 | 2.68 | 474.63 | 1 | 1.03 | 5.54 | 383 |
| | | rec | 1 | 2 | 1 | 8 | 1.04 | 1 | 2.59 | 13.31 | 1.05 | 3.49 | 24.84 | 1.47 | 334 |
| | 4 | mul | 4 | 16 | 16 | 64 | 44.30 | 43.47 | 7.16 | 6.98 | 948.74 | 1 | 1.28 | 7.85 | 1,215 |
| | | add | 4 | 7 | 16 | 28 | 48.50 | 47.92 | 7.00 | 6.74 | 1543.99 | 1 | 1.53 | 6.65 | 786 |
| | | dom | 4 | 4 | 16 | 16 | 48.75 | 48.39 | 6.27 | 6.14 | 2255.75 | 1 | 1.57 | 5.45 | 625 |
| | | rec | 1 | 4 | 1 | 16 | 1.10 | 1 | 3.07 | 9.01 | 1.26 | 9.53 | 94.96 | 1.83 | 442 |

[a] Model: Multiplicative (mul), Additive (add), Dominant (dom), and Recessive (rec) represent genotype relative risks among the nonexposed; relative risks for exposed are 4 times those for nonexposed (i.e., relative risk for interaction = 4).

[b] Case-parental control statistics: *TDT*, $LR_{Mul}$, $LR_{Add}$, $LR_{Dom}$, $LR_{Rec}$. Nonrelated case-control logistic regression models: $Logit_{mul}$, $Logit_{Dom}$, $Logit_{Rec}$.

[c] Sample size estimates are large, impractical, and numerically unreliable.

in Table III: (1) Parental controls can be more efficient than nonrelated controls when the allele frequency is small ($P \leq .10$). Although not shown in Table III, the amount of efficiency depends on $OR_{Env}$; when $OR_{Env}$ increases, parental controls become more efficient than nonrelated controls. (2) When using parental controls, the $LR_{Add}$ and $LR_{Dom}$ statistics tended to be the most efficient methods for when the true genetic models were multiplicative, additive, and dominant. Although the efficiency of $LR_{Dom}$ was slightly better than $LR_{Add}$ for these genetic models, when the true genetic model was recessive, the efficiency of $LR_{Add}$ was much better than that for $LR_{Dom}$. (3) The efficiencies of *TDT* and $LR_{Mul}$ were often similar, and were often much worse than those for $LR_{Add}$ and $LR_{Dom}$. (4) When the true genetic model was recessive, the most efficient method was often $LR_{Rec}$, although $LR_{Mul}$ and *TDT* tended to be somewhat more efficient when the allele frequency was $p = .50$. Although not shown in Table III, similar trends occurred for when $OR_{Env} = 1$ or 4, and for when $r_{\text{int}} = 2$. Because of points 2 and 3 above, it appears that when the genetic mechanism is unknown, the most reasonable statistic to assess GxE when analyzing parental controls was $LR_{Add}$ - it was often close to the most efficient method for multiplicative, additive, and dominant effects, and was much more efficient than $LR_{Dom}$ for recessive effects.

Note that these conclusions are based on the relative efficiencies determined by total sample sizes, which may be most relevant when considering costs related to obtaining and genetically analyzing samples. However, if cases are much more difficult to obtain than controls (parents or nonrelated controls), it may be more informative to consider relative efficiencies as determined by the number of diseased cases. To consider this type of relative efficiency based on the results presented in Table III, let *RE* denote a value of the relative efficiency presented in Table III, and note that when the case-parents design was the most efficient,

$$RE = \frac{2N_{CC}}{3N_{CP}}\cdot$$

So, to consider the relative efficiency in terms of the ratio of required cases ($N_{CC}/N_{CP}$), multiply the *RE* values in Table III for the case-control design by 3/2, which illustrates that in terms of required cases, the case-parents design will be even more efficient. In contrast, when the case-control design is indicated in Table III to be more efficient than the case-parents design, the *RE* values in Table III for the case-parents design should be multiplied by 2/3. Hence, in terms of diseased cases, the *RE* values in Table III for the case-parents design can be as high as 1.5 and still be as efficient as a case-control design that has an efficiency of 1.

To gain insights to the sample sizes required to detect GxE using the case-parents design, we present in Table III the number of cases ($N_{CP}$) required to achieve sufficient power when using the $LR_{Add}$ statistic. This demonstrates the large sample sizes needed to detect even a fairly large GxE, as demonstrated by the genotype relative risks for the various models in Table III. When the interaction parameter $r_{\text{int}}$ was reduced from 4 (Table III) to 2 (results not shown), the sample sizes for $LR_{Add}$ increased by a factor of 3–4 for the multiplicative, additive, and dominant models, and by a factor of 3–8 for the recessive model (larger sample sizes needed when $p$ is small).

## DISCUSSION

We have presented statistical methods, both likelihood-based and in the spirit of the *TDT* method, to assess GxE when using the case-parents design. Methods to compute sample size and power are presented, and numerical comparisons with the traditional case-control design have indicated that the case-parents design can sometimes be more efficient to detect GxE than the case-control design, especially when the susceptible allele is rare and the environmental risk factor has a large effect in the absence of the susceptible genotype. However, the validity of the case-parents design to assess GxE requires independence of the genotype and exposure in the general population (Weinberg CR, pers. comm.), yet the traditional case-control design does not require this independence. When considering how to test for GxE for the case-parents design, our numerical results suggest that the $LR_{Add}$ statistic may be the most robust method when the genetic mode of inheritance is unknown. However, this conclusion depends on the relative risks that we used for sample size calculations. When planning a study, it would be best to consider a specific model of relative risks, and then determine the power of all statistics in order to determine which test would offer the greatest power for the alternative of interest. If the mode of inheritance is known, a likelihood ratio statistic sensitive to the mode of inheritance can be used as the most powerful method. Although our results give promise that the case-parents design can be more efficient than the case-control design, our results demonstrate the need for large sample sizes when attempting to detect GxE.

Up to this point we have considered methods that allow for only two alleles. This may be adequate for sample size and power calculations, but in practice, multiple alleles often occur. Furthermore, the environmental risk factor may not be binary, in which case greater power to assess interaction may be achieved for continuous environmental factors. To allow for multiple alleles and continuous environmental risk factors when analyzing the case-parents design, conditional logistic regression can be used. As discussed in Schaid [1997], this can be performed by matching the diseased case with three pseudo-controls, where the controls are the other three genotypes of children that the parents could have produced. This approach is very general, allowing flexible modeling of the influence of alleles and their interaction with environmental covariates. For an example of this type of application, see Schaid [1995]. Note that this approach is more flexible than using logistic regression to model the transmission status of alleles, because models that resemble genetic effects of dominance can be fit, as well as more flexible assessment of the interaction of risk factors with genotypes. Note that conditional logistic regression is most often applied assuming multiplicative effects of risk factors, so that interaction implies deviation from the multiplicative model. We have discussed the additive effects of alleles on the relative risk (i.e., $r_1 = r$, $r_2 = 2r - 1$), and suggested that a test for interaction with a binary exposure could be performed by comparing $r^+$ with $r^-$. To generalize this approach to a continuous environmental covariate $x$, a regression model could be developed, such as $r = \alpha + \beta x$, where $\beta$ represents the interaction parameter ($\beta = 0$ implies no interaction). Then, the parameters $\alpha$ and $\beta$ can be estimated by likelihood methods, although software would need to be developed to implement this approach. As a note of caution, care must be taken when both genetic and environmental factors are not binary, because interaction (modeled as the product of two factors) and the functional form of the dose-response curve tend to be confounded [Greenland, 1993; Thomas, 1981] so that a misspecified dose-response curve can lead to biased estimates of interaction.

## ACKNOWLEDGMENTS

This research was supported by the United States NIH grant GM51256.

## REFERENCES

Goldstein AM, Falk RT, Korczak JF, Lubin JH. Detecting gene-environment interactions using a case-control design. Genet Epidemiol (submitted).

Greenland S. 1983. Tests for interaction in epidemiologic studies: a review and a study of power. Stat Med 2:243–251.

Greenland S. 1993. Basic problems in interaction assessment. Environ Health Persp Suppl 101(4):59–66.

Hwang S-J, Beaty TH, Liang K-Y, Coresh J, Khoury MJ. 1994. Minimum sample size estimation to detect gene-environment interaction in case-control designs. Am J Epidemiol 140:1029–1037.

Khoury MJ. 1994. Case-parental control method in the search for disease-susceptibility genes. Am J Hum Genet 55:414–415.

Khoury MJ, Flanders D. 1996. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! Am J Epidemiol 144:207–213.

Khoury MJ, Beaty TH, Hwang S-J. 1995. Detection of genotype-environment interaction in case-control studies of birth defects: how big a sample size? Teratology 51:336–343.

Knudson AG, Hino O. 1997. Commentary: Meeting report: genetic environmental interactions in cancer susceptibility in animal models. J Nat Canc Inst 89:1669–1672.

MacMahon B. 1968. Gene-environment interaction in human disease. J. Psychiatr Res (6 Suppl):393–402.

Maestri NE, Beaty TH, Hetmanski J, Smith EA, McIntosh I, Wyszynski DF, Liang K-Y, et al. Application of transmission disequilibrium tests to nonsyndromic oral clefts: including candidate genes and environmental exposures in the models. Am J Med Genet (in press).

Ottman R. 1990. An epidemiologic approach to gene-environment interaction. Genet Epidemiol 7:177–185.

Perera FP. 1997. Environment and cancer: who are susceptible? Science 278:1068–1073.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 1992. Numerical recipes in C. New York: Cambridge University Press.

Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273:1516–1517.

Rothman KJ, Greenland S, Walker AM. 1980. Concepts of interaction. Am J Epidemiol 112(4):467–470.

Schaid DJ. 1995. Relative-risk regression models using cases and their parents. Genet Epidemiol 12:813–818.

Schaid DJ. 1996. General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 13:423–449.

Schaid DJ. 1999. Likelihoods and *TDT* for the case-parents design. Genet Epidemiol 16:250–260.

Schaid DJ, Sommer SS. 1993. Genotype relative risks: methods for design and analysis of candidate-gene association studies. Am J Hum Genet 53:1114–1126.

Schaid DJ, Sommer SS. 1994. Comparison of statistics for candidate-gene association studies using cases and parents. Am J Hum Genet 55:402–409.

Self SG, Longton G, Kopecky KJ, Liang KY. 1991. On estimating HLA/disease association with application to a study of aplastic anemia. Biometrics 47:53–61.

Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516.

Thomas DC. 1981. Are dose-response, synergy, and latency confounded? Alexandria, VA: American Statistical Association.