STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:35–50 (DOI: 10.1002/sim.973)

Sample size requirements for matched case-control studies of gene–environment interaction

W. James Gauderman*,†

Department of Preventive Medicine, University of Southern California, CA, U.S.A.

SUMMARY

Consideration of gene–environment ($G\times E$) interaction is becoming increasingly important in the design of new epidemiologic studies. We present a method for computing required sample size or power to detect $G\times E$ interaction in the context of three specific designs: the standard matched case-control; the case-sibling, and the case-parent designs. The method is based on computation of the expected value of the likelihood ratio test statistic, assuming that the data will be analysed using conditional logistic regression. Comparisons of required sample sizes indicate that the family-based designs (case-sibling and case-parent) generally require fewer matched sets than the case-control design to achieve the same power for detecting a $G\times E$ interaction. The case-sibling design is most efficient when studying a dominant gene, while the case-parent design is preferred for a recessive gene. Methods are also presented for computing sample size when matched sets are obtained from a stratified population, for example, when the population consists of multiple ethnic groups. A software program that implements the method is freely available, and may be downloaded from the website http://hydra.usc.edu/gxe. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: candidate genes; sample size; power; interaction

1. INTRODUCTION

With increasing frequency, epidemiologic studies are addressing hypotheses related to gene–environment ($G\times E$) interaction. These studies are facilitated by the increasing availability of candidate genes and genetic markers, along with technologies that make the utilization of genetic data in large-scale epidemiologic studies affordable. Case-control studies are widely used in epidemiology for studying associations between disease and potential environmental and/or genetic risk factors. In the context of unmatched case-control studies, several authors have described methods for estimating power and sample size in studies of $G\times E$ interaction [1–4]. However, estimates of genetic effects in unmatched case-control studies may suffer from bias due to confounding by ethnicity, also known as population stratification [5], and thus

*Correspondence to: W. James Gauderman, Department of Preventive Medicine, University of Southern California, 1540 Alcazar St, Suite 220, Los Angeles, CA 90089, U.S.A.
†E-mail: jimg@usc.edu

Contract/grant sponsor: NIEHS; contract/grant numbers: ES10421, SP30 ES07048-03
Contract/grant sponsor: NCI; contract/grant numbers: CA52862, CA78296

Received October 2000
Accepted January 2001

one may want to match cases to controls on ethnic background. Matching cases and controls on age as well will be preferable for diseases that show a strong age-trend in incidence and for exposures that exhibit secular trends (for example, the specific composition of oral contraceptives).

For the study of genetic factors, three basic types of matched case-control designs have been proposed. In the first design, which we denote 'case-control', each case is matched to one or more controls selected from the source population of the case. Controls are selected using one of several protocols, including random digit dialling or neighbourhood walking, subject to the matching criteria (for example, age, similar ethnic background). In the second, denoted the 'case-sib' design, each case is matched to one or more unaffected siblings [5–7]. Compared to the case-control design, this design has the advantage that cases and controls are perfectly matched on ethnic background. In the third, denoted the 'case-parent' design, genotypes are collected from the case and his/her two parents, while environmental data are required only from the case [8]. As with the case-sib design, this design provides perfect control for ethnic confounding. The main effect of environmental factors cannot be assessed in the case-parent design, but analysis of genetic main effects and $G \times E$ interactions can be conducted.

Witte *et al.* [5] and Gauderman *et al.* [7] have compared the relative efficiency of the case-sib and case-parent designs to the matched case-control design for estimation of genetic main effects. They also provided some comparisons of efficiency for estimation of a $G \times E$ interaction effect, but only for a limited number of disease models. Neither paper provided calculations of required sample sizes. Schaid [8] described a method for computing power and sample size to detect $G \times E$ interaction in the case-parent design, and provided numerical calculations for some specific examples. He also compared required sample sizes for detecting $G \times E$ interaction using the case-parent to that using the unmatched case-control design, but no comparisons were provided with the matched case-control or case-sib design.

In this paper we describe a general framework for the calculation of power or required sample size for studies of $G \times E$ interaction in the context of the matched case-control, case-sib and case-parent designs. For a range of models, sample size estimates are provided and can be compared across designs to infer their efficiencies relative to one another. We also describe a freely-available, user friendly computer program that can be downloaded from our website and used to compute power or sample size for any settings of the model parameters.

2. METHODS

2.1. Notation and models

Let D be an indicator of disease, E an indicator of exposure, and g a genotype at a candidate locus with susceptibility allele 'A' and normal allele 'a'. The population prevalence of exposure will be denoted by p_{E}, and the prevalence of the high risk allele by q_{A}. We assume Hardy–Weinberg equilibrium so that the distribution of genotypes g in the population is given by $\Pr(g|q_{\mathrm{A}}) = q_{\mathrm{A}}^2$, $2q_{\mathrm{A}}(1-q_{\mathrm{A}})$, and $(1-q_{\mathrm{A}})^2$ for $g = \mathrm{AA}$, Aa and aa, respectively. We furthermore assume independence of genotypes and exposure in the population. Although we assume E is binary in the development of the methods, we will describe generalizations to allow for a continuous exposure factor. For the case-sib design, it is important to allow for possible *a priori* correlation in exposure between siblings. For binary E, we let ϕ denote the odds ratio for exposure in sib 1 when sib 2 is exposed, compared to when sib 2 is unexposed, so that when $\phi > 1$ (<1) there is positive (negative) correlation in exposure between sibs.

In order to utilize the candidate genes in a statistical model, it is necessary to map the three possible genotypes into a genetic covariate. For example, if we assume genetic susceptibility follows a dominant pattern of inheritance, then persons with genotype $g = \mathrm{AA}$ or Aa are genetically susceptible, that is, they are at either increased or decreased risk compared to the baseline group ($g = \mathrm{aa}$). This structure can be captured by defining the genetic covariate $G(g) = 0$ for $g = \mathrm{aa}$, and $G(g) = 1$ for $g = \mathrm{AA}$ or $g = \mathrm{Aa}$. We also consider in this paper the recessive model, in which $G(g) = 1$ for $g = \mathrm{AA}$ and $G(g) = 0$ otherwise. The proportion of subjects in the population that carry a susceptibility genotype is $q_{\mathrm{A}}^2 + 2q_{\mathrm{A}}(1 - q_{\mathrm{A}})$ for a dominant model and q_{A}^2 for a recessive model. For notational convenience, we will simply use G to denote the function $G(g)$.

We consider two models for relating disease to the genetic and environmental covariates: the logistic model and the log-linear model. The reason for considering these two models will be made clear in the development of the case-control likelihood function below. The logistic model is given by

$$\Pr(D = 1 \mid G, E) = \frac{\mathrm{e}^{\alpha + \beta_{\mathrm{g}} G + \beta_{\mathrm{e}} E + \beta_{\mathrm{ge}} GE}}{1 + \mathrm{e}^{\alpha + \beta_{\mathrm{g}} G + \beta_{\mathrm{e}} E + \beta_{\mathrm{ge}} GE}} \tag{1}$$

where the baseline probability of disease in the population is given by $\mathrm{e}^{\alpha}/(1 + \mathrm{e}^{\alpha})$, and the quantities $\mathrm{OR_g} = \mathrm{e}^{\beta_{\mathrm{g}}}$, $\mathrm{OR_e} = \mathrm{e}^{\beta_{\mathrm{e}}}$, and $\mathrm{OR_{ge}} = \mathrm{e}^{\beta_{\mathrm{ge}}}$ are the genetic, environmental and interaction *odds ratios*, respectively. Although we will be forming a conditional likelihood for parameter estimation that will not be a function of α, estimates of power and sample size will depend on this parameter. Note that $\mathrm{OR_g}$ is odds ratio for the gene in the absence of exposure ($E = 0$) and $\mathrm{OR_e}$ is the exposure odds ratio in the genetically non-susceptible ($G = 0$). When $\mathrm{OR_{ge}} = 1$, the odds ratio in exposed, genetically susceptible subjects is $\mathrm{OR_g OR_e}$, that is, the product of the genetic and exposure odds ratios. Thus, $\mathrm{OR_{ge}}$ is a measure of the departure from a purely multiplicative odds-ratio model. When $\mathrm{OR_{ge}} > 1$ ($\mathrm{OR_{ge}} < 1$) exposure stimulates (suppresses) adverse genetic expression, or equivalently genotype confers sensitivity to (protection from) the effects of exposure. The corresponding log-linear model is given by

$$\Pr(D = 1 \mid G, E) = \mathrm{e}^{\alpha + \beta_{\mathrm{g}} G + \beta_{\mathrm{e}} E + \beta_{\mathrm{ge}} GE} \tag{2}$$

where now the baseline probability of disease is simply e^{α}, and $\mathrm{RR_g} = \mathrm{e}^{\beta_{\mathrm{g}}}$, $\mathrm{RR_e} = \mathrm{e}^{\beta_{\mathrm{e}}}$ and $\mathrm{RR_{ge}} = \mathrm{e}^{\beta_{\mathrm{ge}}}$ are the genetic, environmental and interaction *relative risks*, respectively. The interpretation of these relative risks is analogous to those for the odds ratios described above.

Specification of values for the above parameters will be necessary in computing power or sample size. Some of these values may be difficult to specify without pilot data, while the three quantities described below may be more readily available at the stage of planning a study. The first is the population disease prevalence, defined as

$$K_{\mathrm{p}} = \Pr(D = 1) = \sum_{G} \sum_{E} \Pr(D = 1 \mid G, E) \Pr(G \mid q_{\mathrm{A}}) \Pr(E \mid p_{\mathrm{E}}) \tag{3}$$

The second measure is the population-average exposure relative risk, a quantity that would be estimated in an epidemiological study of the exposure factor alone, and is defined as

$$\overline{\mathrm{RR}}_{\mathrm{e}} = \frac{\Pr(D=1\,|\,E=1)}{\Pr(D=1\,|\,E=0)} = \frac{\sum_G \Pr(D=1\,|\,E=1)\Pr(G\,|\,q_{\mathrm{A}})}{\sum_G \Pr(D=1\,|\,E=0)\Pr(G\,|\,q_{\mathrm{A}})} \tag{4}$$

The third measure is the population-average genetic relative risk, which would be estimated in an epidemiological study of the genetic factor alone. It is given by

$$\overline{\mathrm{RR}}_{\mathrm{g}} = \frac{\Pr(D=1\,|\,G=1)}{\Pr(D=1\,|\,G=0)} = \frac{\sum_E \Pr(D=1\,|\,G=1)\Pr(E\,|\,p_{\mathrm{E}})}{\sum_E \Pr(D=1\,|\,G=0)\Pr(E\,|\,p_{\mathrm{E}})} \tag{5}$$

Equations (4) and (5) can be modified accordingly to express the population-average exposure and genetic odds ratios, respectively. The summary quantities in equations (3)–(5) can be used to solve for up to three of the unknown underlying model parameters. In our experience, one is most likely to have estimates (or educated guesses) of q_{A}, p_{E}, ϕ, and the above three summary measures when planning a study. With these in hand, and for specified values of the interaction parameter β_{ge}, one can simultaneously solve equations (3)–(5) to obtain estimates of the parameters α, β_{g} and β_{e}. Alternatively, one may wish to estimate power or sample size for specific values of β_{g} and β_{e}, and thus only an estimate of K_{p} and equation (3) would be needed to obtain the baseline parameter α.

2.2. Sampling designs and likelihood formation

We assume that both genotypic and exposure data will be collected from cases. Matched controls are selected either from the source population (case-control design), from among unaffected siblings of the case (case-sib design), or from parents of the case (case-parent design). Both genotypic and environmental data are collected from controls in the case-control or case-sib design, while only genotypic data are required from parents in the case-parent design. In the development of the general approach to power and sample size calculation, we assume that a single population is under study. Modifications of the approach to allow for two or more subpopulations are described in Section 2.4.

For all three designs, we assume that conditional logistic regression will be utilized in the analysis of the data. Conditional logistic regression for estimation and testing of interactions is routinely used in matched case-control studies [9], and has been described for the analysis of case-sib [5] and case-parent [8] data. The conditional likelihood for a sample of N matched sets has the form:

$$L(\beta_{\mathrm{g}},\beta_{\mathrm{e}},\beta_{\mathrm{ge}}) = \prod_{i=1}^{N} \frac{\mathrm{e}^{\beta_{\mathrm{g}}G_{i1}+\beta_{\mathrm{e}}E_{i1}+\beta_{\mathrm{ge}}G_{i1}E_{i1}}}{\sum_{j\in M(i)}\mathrm{e}^{\beta_{\mathrm{g}}G_{ij}+\beta_{\mathrm{e}}E_{ij}+\beta_{\mathrm{ge}}G_{ij}E_{ij}}} \tag{6}$$

where the index ‘1’ refers to the case and the set $M(i)$ includes all subjects in matched set i. For the case-control and case-sib designs, the terms in the denominator of equation (6) include a contribution from the case and each of his/her matched controls [9]. For the case-parent design, the denominator includes a contribution from the case and from the three ‘pseudo-siblings’ of the case, the latter formed as the three possible genotypes the case could have inherited from the parents but did not. For example, if both the parents have genotype $g = Aa$ and the case has genotype AA, then the genotypes of the three pseudosiblings are Aa (paternal A, maternal a), aA (paternal a, maternal A), and aa (paternal a, maternal a).

Maximum likelihood estimates (MLEs) derived from equation (6) are consistent estimates of the corresponding log-odds ratios from the logistic model (equation (1)) in the case-control or case-sib designs, while they are consistent for the log-relative risks from the log-linear model in the case-parent design [8, 10] (equation (2)). Of course, these will be nearly equivalent for a rare disease [9], but for more common diseases, one should understand that the parameter being estimated by the above likelihood is different in the case-parent design than in the other two designs.

2.3. Calculation of power and sample size

We assume that the null hypothesis (H_0) of interest is that $\beta_{\text{ge}} = 0$, that is, that there is no $G \times E$ interaction. The alternative hypothesis (H_1) may either be two-sided ($B_{\text{ge}} \neq 0$) or one-sided (for example, $B_{\text{ge}} > 0$). We let α and $\beta^* = \{\beta^*_{\text{g}}, \beta^*_{\text{e}}, \beta^*_{\text{ge}}\}$ denote the true parameter values for the logistic (equation (1)) or log-linear (equation (2)) model and let $\Omega = \{q_{\text{A}}, p_{\text{E}}, \phi\}$ denote the true values of the other model parameters. To estimate power or sample size, we use the following approach:

1. Maximize the expected log-likelihood $L^1 = \ln[L(\beta_{\text{g}}, \beta_{\text{e}}, \beta_{\text{ge}})]$ with respect to the distribution of observable genotype and exposure data conditional on the true parameters α, β^* and Ω. We denote the expected MLEs by $\hat{\beta}^1 = \{\hat{\beta}^1_{\text{g}}, \hat{\beta}^1_{\text{e}}, \hat{\beta}^1_{\text{ge}}\}$ and the corresponding maximum of the expected log-likelihood by $\hat{L}^1$. Details of the computation of these expected values are provided in Appendix A.
2. In an analogous fashion, maximize the expected log-likelihood $L^0 = \ln[L(\beta_{\text{g}}, \beta_{\text{e}})]$, that is, the expectation of the main-effect parameters of the conditional logistic regression in which β_{ge} is fixed at its null value of zero. The expected MLEs from this model are denoted $\hat{\beta}^0 = \{\hat{\beta}^0_{\text{g}}, \hat{\beta}^0_{\text{e}}\}$ and the corresponding maximum of the expected log-likelihood is denoted $\hat{L}^0$.
3. Define $\Lambda = 2(\hat{L}^1 - \hat{L}^0)$, that is, the likelihood ratio test statistic for a single matched set based on the expected maximum log-likelihoods under H_1 and H_0. For a given number N of matched sets, the quantity $N\Lambda$ is the non-centrality parameter of the chi-squared distribution under the alternative hypothesis.
4. To compute required sample size, we can first use a non-central χ^2 table [11] to find the non-centrality C corresponding to the desired power $(1 - b)$ and significance level (a), and then equate C to our expectation $N\Lambda$ and solve for N [12]. In the case of a one-degree of freedom test of H_0, which results when G and E are each represented by a single covariate, $C = (z_a + z_b)^2$, where z_u denotes the $(1 - u)$th percentile of the standard normal distribution. Thus, for a one-sided alternative hypothesis, the required sample size can be computed as

$$N = (z_a + z_b)^2/\Lambda \tag{7}$$

   For a two-sided alternative hypothesis, z_a is replaced by $z_{a/2}$.
5. Alternatively, for a given N, compute power for a one-sided alternative hypothesis as $1 - b = \Phi(\sqrt{(N\Lambda)} - z_a)$ where $\Phi(u)$ is the cumulative standard normal distribution evaluated at u. For a two-sided alternative, $1 - b = \Phi(\sqrt{(N\Lambda)} - z_{a/2}) + \Phi(-\sqrt{(N\Lambda)} - z_{a/2})$.

2.4. Allowance for matching variables

Assume that the population consists of S strata, with $s=1,2,\ldots,S$ used to denote specific groups. If there are variations across strata in both disease prevalence (K_p) and genotype and/or exposure parameters (q_A, p_E, Φ), then one should match cases to controls within stratum to avoid biases in the relative risk parameters. Examples of strata in which such variations are likely to be present include racial/ethnic groups or age groups. We assume that the relative risk parameters β_g^*, β_e^* and β_{ge}^* are the same in all strata. If this is known not to be the case prior to conducting the study, one should conduct separate sampling and estimation within each stratum, as relative risk estimates from a mixture sample will not be meaningful.

The case-parent design provides implicit control for any stratification factor, and while the case-sib design also provides control for race/ethnicity, one may still have to actively match on other factors (for example, age) in this design. However, regardless of the design, stratification in the population will have consequences for the statistical power and required sample size. The sample size and power calculation approach described above is easily modified to account for population stratification, by simply including stratum-specific parameters in the computation of the expected log-likelihood. Details are described in Appendix B.

2.5. Computer software

We have developed a user-friendly Windows-based software program that implements the approach described above for computing either power or sample size. The program was written in C++ and utilizes Microsoft foundation classes to create menus and dialogues. Inputs to the program include the design (case-control, case-sib or case-parent), true model parameters (α,β^*,Ω), significance level, and either power or sample size. Optionally, the population prevalence K_p may be specified instead of baseline disease prevalence parameter α, and the population-average effect measures in equations (4) and (5) may be input instead of the model parameters β_g^* and β_e^*. The program is available at no charge and may be downloaded from our website http://hydra.usc.edu/gxe.

2.6. Design comparisons

We compare required sample sizes for the three matched designs, under four specific conceptual models of $G\times E$ interaction of the type described by Ottman [13, 14] and Khoury *et al.* [15]. In all models, we assume that $\beta_{ge}^*>0$ so that there is a positive interaction between G and E. In the first model, which we denote the *pure interaction* model $\beta_e^*=0$ and $\beta_g^*=0$, that is, disease risk is increased only in persons that are both exposed and genetically susceptible. In the second model, denoted the *G-modification* model, $\beta_e^*=0$ and $\beta_g^*=\ln(2)$ so that there is no E effect in the absence of genetic susceptibility, but there is a G effect in the absence of the environment and that effect is modified in the presence of E. The third model, the *E-modification* model, is analogous to model 2 with $\beta_e^*=\ln(2)$ and $\beta_g^*=0$. Finally, in the fourth model, the *GE-modification* model, $\beta_e^*=\ln(2)$ and $\beta_g^*=\ln(2)$ so that both G and E have independent effects in the absence of each other. For all models, we fix the population disease prevalence at $K_p=0.01$ and use this to solve for the underlying baseline disease parameter α.

We assume a two-sided alternative hypothesis and compute the required number of matched sets to achieve 80 per cent power for rejecting the null hypothesis that $\beta_{ge}=0$. We also compute the ratio of N for the case control design to N for each of the other two designs, which provides a measure of asymptotic relative efficiency (ARE) of the family-based designs to the case-control design.

3. RESULTS

3.1. Design comparisons

Under the pure interaction model, required sample sizes are quite large when $p_{\mathrm{E}}=0.20$ and the gene is uncommon ($\Pr[G=1]=0.01$, Table I). For a dominant gene, one would have to sample approximately 16000, 2350 or 1000 matched case-control sets to detect an interaction of magnitude $R_{\mathrm{ge}}=2.0, 5.0$ or 10.0, respectively. The case-parent design is slightly more efficient for all R_{ge}, while the case-sib design is slightly more efficient when $R_{\mathrm{ge}}=2$, but increases in relative efficiency as R_{ge} increases (for example, $\mathrm{ARE}=1.28$ when $R_{\mathrm{ge}}=10.0$). For a recessive gene, identical numbers are required in the case-control design since the joint distribution of case and control genotypes does not depend on Mendelian transmission. The case-sib design is still more efficient than the case-control design for a recessive gene, although less so than for a dominant gene, but the case-parent design is substantially more efficient with AREs that range from approximately 1.5 to 1.9.

When the gene is more common ($\Pr[G=1]=0.20$) and $p_{\mathrm{E}}=0.20$, substantially fewer case-control pairs are required (approximately 1100, 180 and 90 for $R_{\mathrm{ge}}=2, 5$ and 10, respectively, Table I). Here, the case-parent design provides the greatest efficiency in all situations considered. When the exposure factor is more common ($p_{\mathrm{E}}=0.50$), there is again a reduction in required sample sizes for all designs. The case-sib design now provides the greatest efficiency for a dominant gene, regardless of the prevalence of genetic susceptibility, while the case-parent design is again most efficient for a recessive gene. Taken as a whole, the results in Table I demonstrate that the family-based designs provide greater efficiency compared to the standard matched case-control design, with the most efficient design depending on whether the gene is dominant (case-sib optimal) or recessive (case-parent optimal).

Table II provides a comparisons of required sample sizes for the four conceptual models of interaction described in Section 2.6. Holding $R_{\mathrm{ge}}=5$, required sample sizes in the matched case-control design are lowest for the E-modification model ($R_{\mathrm{e}}=2$, $R_{\mathrm{g}}=1$) and highest for the G-modification model ($R_{\mathrm{e}}=1$, $R_{\mathrm{g}}=2$). Both of the case-sib and case-parent designs provide greater efficiency than the case-control design for most models. Again, the greatest efficiency is generally achieved with the case-sib design for a dominant gene and with the case-parent design for a recessive gene. For the case-sib design, the ARE is greater when $R_{\mathrm{e}}=2$ than when $R_{\mathrm{e}}=1$, regardless of p_{E} and the mode of inheritance. The ARE for the case-parent design, on the other hand, depends most strongly on the values of p_{E} and R_{g}, with the highest AREs for $R_{\mathrm{g}}=2$ when $p_{\mathrm{E}}=0.2$, and for $R_{\mathrm{g}}=1$ when $p_{\mathrm{E}}=0.5$.

The efficiency of the case-sib design relative to the matched case-control design declines as the sibling exposure-sharing odds ratio (ϕ) increases (Table III). Provided ϕ is between 1 and 5, the case-sib design still provides increased efficiency relative to the case-control design for most models. However, in the extreme case of complete sharing of exposure between sibs ($\phi=\infty$), the case-sib design can require more than twice the sample size to achieve the same power as the case-control design. Large exposure-sharing odds ratios are

Table I. Number (N) of matched sets required for 80 per cent power to detect a $G \times E$ interaction of magnitude R_{ge}. Pure interaction model ($R_g = 1,\ R_e = 1$), disease prevalence= 0.01, significance level = 0.05, 2-sided alternative.

Exposure	Gene			Case-Control	Case-Sibling		Case-Parent	
$\Pr(E=1)$	$\Pr(G=1)$	Inheritance	R_{ge}	N	N	(Ratio*)	N	(Ratio*)
0.20	0.01	Dom	2	16119	15444	(1.04)†	15950	(1.01)
			5	2347	2044	(1.15)†	2279	(1.03)
			10	1030	804	(1.28)†	982	(1.05)
		Rec	2	16119	15714	(1.03)	10605	(1.52)†
			5	2347	2158	(1.09)	1370	(1.71)†
			10	1030	883	(1.17)	546	(1.89)†
	0.20	Dom	2	1078	1054	(1.02)	1003	(1.07)†
			5	181	172	(1.05)	160	(1.13)†
			10	92	86	(1.07)	80	(1.15)†
		Rec	2	1078	1058	(1.02)	830	(1.30)†
			5	181	173	(1.05)	124	(1.46)†
			10	92	87	(1.06)	59	(1.56)†
0.50	0.01	Dom	2	11547	10432	(1.11)†	11499	(1.00)
			5	1988	1472	(1.35)†	1976	(1.01)
			10	985	598	(1.65)†	981	(1.00)
		Rec	2	11547	10861	(1.06)	7514	(1.54)†
			5	1988	1646	(1.21)	1140	(1.74)†
			10	985	711	(1.39)	519	(1.90)†
	0.20	Dom	2	773	732	(1.06)†	763	(1.01)
			5	157	138	(1.14)†	166	(0.95)
			10	94	79	(1.19)†	109	(0.86)
		Rec	2	773	739	(1.05)	624	(1.24)†
			5	157	140	(1.12)	125	(1.26)†
			10	94	81	(1.16)	78	(1.21)†

*Compared to the case-control design; ratios above one indicate greater efficiency.
†The most efficient design.

most likely to occur when study subjects are children and the exposure is family-specific, for example E = parental smoking. On the other hand, if there is discordance in exposure between sibs (for example $\phi = 0.5$), the ARE of the case-sib design is larger than when exposures are independent (Table III, last column). Although such negative correlation in exposure is unlikely in sibs, it could conceivably be a by-product of sibling rivalry for some exposures.

3.2. Real study examples

We provide three example calculations of required sample size in the context of testing specific hypotheses, to demonstrate aspects of the design comparisons presented above. In all computations, we assumed a two-sided alternative hypothesis and 0.05 significance level.

Table II. Number (N) of matched sets required for 80 per cent power to detect a $G\times E$ interaction of magnitude $R_{\mathrm{ge}}=5.0$. Varying R_{e} and R_{g}, with $\Pr(\mathrm{G}=1)=0.05$, $K_{\mathrm{p}}=0.01$, significance level$=0.05$, and a 2-sided alternative.

Exposure	Gene			Case-Control	Case-Sibling		Case-Parent	
$\Pr(E=1)$	Inheritance	R_{e}	R_{g}	N	N	(Ratio*)	N	(Ratio*)
0.2	Dom	1	1	514	457	(1.12)	489	(1.05)
		1	2	517	464	(1.11)	416	(1.24)
		2	1	465	328	(1.42)	442	(1.05)
		2	2	485	343	(1.41)	389	(1.25)
	Rec	1	1	514	473	(1.09)	324	(1.59)
		1	2	517	479	(1.08)	271	(1.91)
		2	1	465	359	(1.30)	257	(1.81)
		2	2	485	376	(1.29)	223	(2.17)
0.5	Dom	1	1	437	338	(1.29)	443	(0.99)
		1	2	507	413	(1.23)	582	(0.87)
		2	1	406	243	(1.67)	429	(0.95)
		2	2	468	304	(1.54)	580	(0.81)
	Rec	1	1	437	363	(1.20)	283	(1.54)
		1	2	507	438	(1.16)	367	(1.38)
		2	1	406	276	(1.47)	242	(1.68)
		2	2	468	339	(1.38)	324	(1.44)

*Compared to the case-control design; ratios above one indicate greater efficiency.

Table III. Number (N) of matched sets required for 80 per cent power to detect a $G\times E$ interaction of magnitude $R_{\mathrm{ge}}=5.0$. Varying the sibling exposure-sharing odds ratio, with $\Pr(G=1)=0.05$, $K_{\mathrm{p}}=0.01$, significance level$=0.05$, and a 2-sided alternative.

				Case-Control	N for case-sib design by magnitude of sibling exposure-sharing odds ratio (ϕ)							
$\Pr(E=1)$	Gene	R_{e}	R_{g}	N	$\phi=1$	Ratio*	$\phi=5$	Ratio*	$\phi=\infty$	Ratio*	$\phi=0.5$	Ratio*
0.2	Dom	1	1	514	457	(1.12)	546	(0.94)	1003	(0.51)	433	(1.19)
		2	2	485	343	(1.41)	424	(1.14)	821	(0.59)	328	(1.48)
	Rec	1	1	514	473	(1.09)	533	(0.96)	766	(0.67)	456	(1.13)
		2	2	485	376	(1.29)	431	(1.13)	628	(0.77)	366	(1.33)
0.5	Dom	1	1	437	338	(1.29)	442	(0.99)	884	(0.49)	305	(1.43)
		2	2	468	304	(1.54)	418	(1.12)	1126	(0.42)	271	(1.73)
	Rec	1	1	437	363	(1.20)	441	(0.99)	677	(0.65)	336	(1.30)
		2	2	468	339	(1.38)	437	(1.07)	865	(0.54)	309	(1.51)

*Compared to the case-control design; ratios above one indicate greater efficiency.

3.2.1. BRCA1 × Oral contraceptive use in breast cancer.

An interaction between BRCA1 and oral contraceptive (OC) use in the risk of breast cancer has been reported by Ursin *et al.* [16] using a case-only design [17]. Assume that one is interested in replicating this finding using a matched case-control design. We assume that case selection will be restricted

Table IV. Number (N) of matched sets required for 80 per cent power to detect a BRCA1 × OC interaction of size R_{ge}.

BRCA1 × OC Interaction (R_{ge})	Case-Control	Case-Sibling		Case-Parent	
	N	N	(Ratio*)	N	(Ratio*)
2	9242	4778	(1.93)	8841	(1.05)
3	3603	1701	(2.12)	3402	(1.06)
4	2260	997	(2.27)	2116	(1.07)
5	1688	706	(2.39)	1571	(1.07)
10	878	312	(2.81)	811	(1.08)
15	676	219	(3.09)	630	(1.07)
20	583	179	(3.26)	553	(1.05)
30	494	141	(3.50)	490	(1.01)

*Compared to the case-control design; ratios above one indicate greater efficiency.
Assumptions: $\Pr(G=1)=0.01$, dominant inheritance, $\Pr(E=1)=0.3$, $\phi=2$, $R_{\mathrm{e}}=1$, $R_{\mathrm{g}}=5$, $K_{\mathrm{p}}=0.01$. Significance level $=0.05$, and a 2-sided alternative hypothesis.

to women under age 50 years, to maximize the yield of BRCA1-related disease, and that controls will be matched to cases on age at case-diagnosis. The BRCA1 mutation frequency is assumed to be $q_{\mathrm{A}}=0.005$ so that, under dominant inheritance, approximately 1 per cent of the population are genetically susceptible. We also assume that the prevalence of OC use is $p_{\mathrm{E}}=0.3$ and that the sib of an OC-user has twice the odds of using OCs (that is, $\phi=2$). For demonstration purposes, we consider a model in which OCs have no effect in BRCA1 normals ($R_{\mathrm{e}}=1$) but carrying a BRCA1 mutation increases risk five-fold in OC non-users ($R_{\mathrm{g}}=5$). Finally, we assume the cumulative risk of breast cancer by age 50 is 1 per cent ($K_{\mathrm{p}}=0.01$). Table IV shows the required number of matched sets to achieve 80 per cent power for detecting various magnitudes of the BRCA1 × OC interaction (R_{ge}) for each of the three matched designs. To detect a BRCA1 × OC interaction relative risk of 5.0, one would need to sample 1688 case-controls pairs, with substantially larger sample sizes required to detect a lesser effect. Even for a very strong interaction effect ($R_{\mathrm{ge}}=30$), nearly 500 matched pairs would be required. Sample sizes for the case-parent design are slightly less, while for the case-sib design they are substantially less. For example, 706 case-sib pairs are required when $R_{\mathrm{ge}}=5$, while only 141 are needed if R_{ge} is 30.

3.2.2. GSTM1 × air pollution in asthma. Gilliland *et al.* [18] have hypothesized that the risk of childhood respiratory symptoms (for example, asthma attack) is influenced by exposure to air pollution (AP). Furthermore, they have reasoned that AP-associated risk is likely to be further increased in children that have reduced oxidant defences, for example in children with the null/null genotype at the GSTM1 locus. One might consider a case-control study to test this hypothesis. To demonstrate the associated sample size calculations, we assume a recessive gene and set q_{A} to 0.5, so that 25 per cent of the population are assumed to carry the null/null genotype. We furthermore assume that AP exposure can be dichotomized into high or low, with the prevalence of high exposure set at $p_{\mathrm{E}}=40$ per cent and a high degree of sharing among sibs ($\phi=10$). We assume that both GSTM1 and AP have independent effects on disease risk ($R_{\mathrm{g}}=2$ and $R_{\mathrm{e}}=2$, respectively), and set the population prevalence of asthma

Table V. Number (N) of matched sets for 80 per cent power to detect a GSTM1 × ETS interaction of size R_{ge}.

GSTM1 × ETS Interaction (R_{ge})	Case-Control	Case-Sibling		Case-Parent	
	N	N	(Ratio*)	N	(Ratio*)
2	766	930	(0.82)	583	(1.31)
3	333	402	(0.83)	266	(1.25)
4	227	272	(0.83)	191	(1.19)
5	180	215	(0.84)	160	(1.13)
10	114	133	(0.86)	126	(0.90)
15	98	113	(0.87)	128	(0.77)
20	92	105	(0.88)	136	(0.68)
30	86	97	(0.89)	155	(0.55)
50	85	94	(0.90)	197	(0.43)
100	89	97	(0.92)	296	(0.30)

*Compared to the case-control design; ratios above one indicate greater efficiency.
Assumptions: $\Pr(G=1)=0.25$, recessive inheritance, $\Pr(E=1)=0.4$, $\phi=10$, $R_e=2$, $R_g=2$, $K_p=0.10$. Significance level $=0.05$, and a 2-sided alternative hypothesis.

to $K_p=10$ per cent. Table V gives the required sample sizes for the three designs. Since genetic susceptibility is more common than in example 3.2.1, sample sizes are much lower for this study. For example, if $R_{ge}=5$, only 180 case-control pairs are required. The case-parent design requires the fewest matched sets ($N=160$), while the case-sib design requires the most ($N=215$), due to the high degree of concordance in exposure between sibs.

3.2.3. GSTM1 × air pollution in asthma, sampling from a multi-ethnic cohort. The Children's Health Study (CHS) [19, 20] is a 10-year prospective study to investigate the effects of air pollution on childrens' respiratory health in southern California. The sample includes school-aged children from four racial/ethnic groups: African-Americans; Asian-Americans; Hispanic-Americans; European-Americans. One might consider obtaining a nested case-control sample from within the CHS to study a possible GSTM1 × air pollution interaction for asthma. In this situation, it would be important to match cases to controls on ethnic background, since this factor is known *a priori* to be related to allele frequency, exposure frequency (for some pollutants) and to prevalence of asthma.

Suppose the study described in Section 3.2.2 will focus sampling on two of the four ethnic groups: African-Americans (AA) and European-Americans (EA). We assume that parameter values in the EA group are those described in Section 3.2.2. In the AA group, we assume that the null allele is less frequent ($q_A=0.30$), air pollution exposure is more common ($p_E=0.60$), sibling exposure sharing is equivalent to the EA group ($\phi=10$), and asthma is more prevalent ($K_p=15$ per cent). Required sample sizes to detect a GSTM1 × air pollution interaction with magnitude $R_{ge}=5$ are shown in Table VI, varying the proportion (π) of subjects in each stratum. When $\pi_{EA}=1.0$, all subjects in the sample are EA and the required sample sizes are the same as those shown in Table V for $R_{ge}=5$. As the sampling proportion of AA subjects increases, the total required sample size increases, primarily due to the lower assumed genetic susceptibility fraction in AA compared to EA. For a sample of 50 per cent EA and 50 per cent AA, 222 case-control pairs are required, 111 from each ethnic group. Note that this sample size

Table VI. Number (N) of matched sets for 80 per cent power to detect a GSTM1 × ETS interaction of size $R_{\mathrm{ge}} = 5$ When the sample includes subjects from two sub-populations with differing prevalence parameter values.**

Sampling mixture+		Case-Control	Case-Sibling		Case-Parent	
Stratum 1	Stratum 2	N	N	(Ratio*)	N	(Ratio*)
100%	0%	180	215	(0.84)	160	(1.13)
80%	20%	191	219	(0.87)	175	(1.09)
60%	40%	209	232	(0.90)	195	(1.07)
50%	50%	222	243	(0.91)	207	(1.07)
40%	60%	238	259	(0.92)	221	(1.08)
20%	80%	286	311	(0.92)	257	(1.11)
0%	100%	377	426	(0.88)	312	(1.21)

*Compared to the case-control design; ratios above one indicate greater efficiency.

**Stratum 1: $\Pr(G=1)=0.25$, $\Pr(E=1)=0.4$, $K_{\mathrm{p}}=0.10$.

Stratum 2: $\Pr(G=1)=0.09$, $\Pr(E=1)=0.6$, $K_{\mathrm{p}}=0.15$.

Additional assumptions: Recessive inheritance, $R_{\mathrm{e}}=2$, $R_{\mathrm{g}}=2$, $\phi=10$, significance level $=0.05$, and a 2-sided alternative hypothesis.

is not simply the average of sample sizes when $\pi_{\mathrm{EA}}=1.0$ ($N=180$) and $\pi_{\mathrm{AA}}=1.0$ ($N=377$), demonstrating the need to allow for the sample mixture and stratum-specific parameter values directly in the computation of sample size and power.

4. DISCUSSION

We presented a unified framework for computing power or required sample size for matched case-control, case-sibling and case-parent studies of gene–environment interaction. Sample size comparisons indicate that one of the two family-based designs (case-sibling or case-parent) typically requires fewer matched sets than the standard case-control design, with the case-sibling design preferred when genetic susceptibility is rare, and the case-parent design when susceptibility is more common. Although we presented required sample sizes for a number of parameter combinations, it is likely that our choices will differ in some way from most studies that are being planned. It is therefore important for the investigators planning a new study to compute sample sizes specific to their particular design parameters. For this reason, we provide, free of charge, a user-friendly Windows-based software program that may be downloaded from our website. For *unmatched* case-control studies of $G \times E$ interaction, Garcia-Closas and Lubin (Garcia-Closas, 1999 #177) also distribute a user-friendly software program for computing sample size or power.

The methods presented in this paper are applicable to very specific designs, that is, one-to-one (1 : 1) matching of cases to controls in the case-control and case-sib designs, and sets of one case with both parents in the case-parent design. If multiple cases and/or controls per matched set were to be studied, the methods described could be extended naturally provided a candidate gene was being studied. However, if a marker gene, believed to be in disequilibrium with the disease gene, was being studied, sample size methods would have to be modified to allow for the residual familial correlation due to linkage [21–25]. Even if a study is planned with 1 : 1 matching as the standard, there are likely to be deviations in the actual sample availability. For example, if a case-sibling study is planned, many cases will not have a sibling that meets the necessary criteria for being a control [5], while in a case-parent study, some cases may have only one parent available [26]. Some have suggested a hybrid design, in which case-parent triads are recruited if possible, and unaffected sibling controls sampled when parents are unavailable [27]. Sample size requirements in such a hybrid design should be intermediate between the requirements for each design alone. Andrieu *et al.* [28] have proposed counter matching cases to controls as a technique for using available data at the time of sampling (for example, family history of disease) to enrich the sample for informative matched sets. Additional work is required to compare the sample size requirements of these alternative designs to the more standard designs considered in this paper.

Complex diseases are likely to depend on gene–gene ($G \times G$) as well as ($G \times E$) interactions. For the case-control design, one could use the methods described in this paper if the second gene is assumed to follow a dominant or recessive susceptibility pattern. Here the exposure prevalence would be replaced by the corresponding prevalence of susceptible individuals, with remaining calculations unchanged. This, however, does not work for the case-sibling design, since the correlation in the second gene is not easily described by a single ‘exposure-sharing’ odds ratio. For the case-parent design, one would have to create a binary covariate based on susceptibility to one of the two genes (making this gene the equivalent of a binary exposure factor), and then study parent-to-case transmission of the remaining gene. This would be valid if a causative gene (for example, BRCA1) had already been identified, and one wanted to test for interaction between that gene and a second, more speculative gene. However, this approach is less satisfying if neither gene has previously been established as a risk factor. More appropriate methodology tailored to computing sample size and power for studies of $G \times G$ interaction will be the subject of a separate paper.

APPENDIX A: CALCULATION OF THE EXPECTED LOG-LIKELIHOOD

The key step in the calculation of power or required sample size is the computation of the expected log-likelihood. Regardless of the specific design being considered, the form of the expected log-likelihood is

$$E(\ln[L(\boldsymbol{\beta})]) = \sum_{\underline{g}} \sum_{\underline{E}} \ln[L(\boldsymbol{\beta}; \underline{G}, \underline{E})] f(\underline{g}, \underline{E} \mid \underline{D}, \alpha, \boldsymbol{\beta}^*, \Omega) \tag{A1}$$

where the summation is over all possible observable genotypes ($\underline{g}$) and exposures ($\underline{E}$) in a matched case-control set. For a continuous environmental covariate, the summation over $\underline{E}$ is replaced with an integral. The factor $L(\boldsymbol{\beta}; \underline{G}, \underline{E})$ is the contribution to the likelihood in equation (6) for a matched set with specific genotypes $\underline{g}$ and exposures $\underline{E}$. The factor $f(\underline{g}, \underline{E} \mid \underline{D}, \alpha, \boldsymbol{\beta}^*, \Omega)$ is the probability distribution for the observable covariate data in a given matched set conditional on the disease-based ascertainment rule and the true model parameters. For a 1 : 1 matched case-control or case-sib study, $\underline{D} = \{D_1, D_2\}$ with disease status $D_1 = 1$ in the case and $D_2 = 0$ in the control. In the case-parent design, $\underline{D} = D_1$, that is, only the disease status of the case.

The joint density of genotypes and exposures in a matched set has the form

$$f(\underline{g},\underline{E}\mid\underline{D},\alpha,\boldsymbol{\beta}^*,\Omega)=\frac{f(\underline{D}\mid\underline{G},\underline{E},\alpha,\boldsymbol{\beta}^*)f(\underline{g}\mid\Omega)f(\underline{E}\mid\Omega)}{\sum_{\underline{\underline{g}}}\sum_{\underline{\underline{E}}}f(\underline{D}|\underline{\underline{G}},\underline{\underline{E}},\alpha,\boldsymbol{\beta}^*)f(\underline{\underline{g}}\mid\Omega)f(\underline{\underline{E}}\mid\Omega)} \tag{A2}$$

The first factor in the numerator of equation (A2) is the penetrance function, which we assume is based on the logistic model in equation (1) for the case-control and case-sib designs, and on the log-linear model in equation (2) in the case-parent design. We make the assumption that disease status is independent within each matched set, conditional on the genotypes and exposures, so that $f(\underline{D}\mid\underline{G},\underline{E},\alpha,\boldsymbol{\beta}^*)=\prod_i f(D_i\mid G_i,E_i,\alpha,\boldsymbol{\beta}^*)$. This is reasonable for the case-control and case-parent designs, but the assumption would be violated in the case-sib design if there was residual correlation between sibling disease status due to unmeasured factors. Substitution of equation (A2) into equation (A1) yields the following expression for the expected log-likelihood:

$$E(\ln[L(\boldsymbol{\beta})])=\frac{\sum_{\underline{g}}\sum_{\underline{E}}\ln[L(\boldsymbol{\beta};\underline{G},\underline{E})]f(\underline{D}\mid\underline{G},\underline{E},\alpha,\boldsymbol{\beta}^*)f(\underline{g}\mid\Omega)f(\underline{E}\mid\Omega)}{\sum_{\underline{\underline{g}}}\sum_{\underline{\underline{E}}}f(\underline{D}\mid\underline{\underline{G}},\underline{\underline{E}},\alpha,\boldsymbol{\beta}^*)f(\underline{\underline{g}}\mid\Omega)f(\underline{\underline{E}}\mid\Omega)} \tag{A3}$$

Additional components of equation (A3) are now described for each of the three matched designs. In all of the following, we let the subscripts 1, 2, m and f denote the case, control, mother and father, respectively, and we assume the environmental factor is a binary indicator of exposure.

1. *Matched Case-control*. For a 1 : 1 matched set, the random variables include $\underline{g}=\{g_1,g_2\}$ and $\underline{E}=\{E_1,E_2\}$, so that there are $3\times3\times2\times2=36$ possible joint covariate profiles. Since cases and controls are assumed to be unrelated, $f(\underline{g}|\Omega)=\Pr(g_1\mid q_{\rm A})\Pr(g_2\mid q_{\rm A})$ and $f(\underline{E}\mid\Omega)=\Pr(E_1\mid p_{\rm E})\Pr(E_2|p_{\rm E})$.
2. *Case-sib*. For a 1:1 matched set, the random variables again include $\underline{g}=\{g_1,g_2\}$ and $\underline{E}=\{E_1,E_2\}$, and there are $3\times3\times2\times2=36$ possible joint covariate profiles. The joint distributions of genotypes, $f(\underline{g}|\Omega)=\Sigma_{g_{\rm m}}\Sigma_{g_{\rm f}}\Pr(g_1|g_{\rm m},g_{\rm f})\Pr(g_2|g_{\rm m},g_{\rm f})\Pr(g_{\rm m}|q_{\rm A})\Pr(g_{\rm f}|q_{\rm A})$, and of exposures, $f(E|\Omega)=\Pr(E_1|E_2,\phi)\Pr(E_2|p_{\rm E})$, are the determinants of power differences relative to the case-control design.
3. *Case-parent*. The random variables include three genotypes $\underline{g}=\{g_1,g_{\rm m},g_{\rm f}\}$ and one exposure $\underline{E}=\{E_1\}$, so that in theory there would be $3\times3\times3\times2=54$ possible profiles. However, the joint distribution $f(\underline{g}|\Omega)=\Pr(g_1|g_{\rm m},g_{\rm f})\Pr(g_{\rm m}|q_{\rm A})\Pr(g_{\rm f}|q_{\rm A})$ is positive for only 10 of the 27 combinations of joint genotypes due to Mendelian rules of gene transmission (see Table 1 in Schaid [8] for a listing of these ten). There are thus only 20 observable joint genotype and exposure profiles in this design. The exposure density for this design is simply $f(E|\Omega)=\Pr(E_1|p_{\rm E})$.

For a continuous exposure factor, the summations over E in equation (A3) are replaced by integrals and the probability functions for E_1 and E_2 above are replaced by the appropriate density functions.

APPENDIX B: THE EXPECTED LOG-LIKELIHOOD IN A STRATIFIED POPULATION

We describe how the expected log-likelihood in Appendix A can be modified to allow for variations in $q_{\mathrm{A}}, p_{\mathrm{E}}, \phi$, and K_{p} across S strata in the population. Let K_{p}^s denote the disease prevalence in stratum s, $s=1,\ldots,S$, and let $\Omega^s=\{q_{\mathrm{A}}^s, p_{\mathrm{E}}^s, \phi^s\}$ denote the corresponding stratum specific genotype and exposure frequency parameters. Let π^s be the proportion of subjects sampled from stratum s, where $\Sigma_s \pi^s = 1.0$. The baseline prevalence parameters α^s, $s=1,\ldots,S$, can be computed separately for each stratum using the corresponding value of K_{p}^s and equation (3).

For a stratified sample, the expected log-likelihood shown in Appendix A (equation (A1)) is replaced by

$$E(\ln[L(\boldsymbol{\beta})]) = \sum_{\underline{g}} \sum_{\underline{E}} \ln[L(\boldsymbol{\beta}; \underline{G}, \underline{E})] \left(\sum_s f(\underline{g}, \underline{E} \,|\, \underline{D}, \alpha^s, \boldsymbol{\beta}^*, \Omega^s) \pi^s \right)$$

Since the conditional log-likelihood does not depend on s, this equation can be rewritten as

$$E(\ln[L(\boldsymbol{\beta})]) = \sum_s \sum_{\underline{g}} \sum_{\underline{E}} \ln[L(\boldsymbol{\beta}; \underline{G}, \underline{E})] f(\underline{g}, \underline{E} \,|\, \underline{D}, \alpha^s, \boldsymbol{\beta}^*, \Omega^s) \pi^s$$

Thus, the expected log-likelihood depends on the distribution of genes and exposures within each stratum, weighted by the proportion of each stratum in the sample. This likelihood is maximized and the approach described in Section 2.3 is used to compute power or sample size. For the latter, the total requirement N is multiplied by π^s, $s=1,\ldots,S$ to obtain the stratum-specific sample sizes.

ACKNOWLEDGEMENTS

This work was supported in part by NIEHS grants ES10421 and 5P30 ES07048-03, and NCI grants CA52862 and CA78296.

REFERENCES

1. Garcia-Closas M, Lubin J. Power and sample size calculations in case-control studies of gene–environment interactions: comments on different approaches. *American Journal of Epidemiology* 1999; **149**:689–692.
2. Foppa I, Spiegelman D. Power and sample size calculations for case-control studies of gene–environment interactions with a polytomous exposure variable. *American Journal of Epidemiology* 1997; **146**(7):596–604.
3. Goldstein A, Falk R, Korczak J, Lubin J. Detecting gene–environment interactions using a case-control design. *Genetic Epidemiology* 1997; **14**:1085–1089.
4. Hwang S, Beaty T, Liang K, Coresh J, Khoury M. Minimum sample size estimation to detect gene–environment interaction in case-control designs. *American Journal of Epidemiology* 1994; **140**:1029–1037.
5. Witte J, Gauderman W, Thomas D. Asymptotic bias and efficiency in case-control studies of candidate genes and gene–environment interactions: basic family designs. *American Journal of Epidemiology* 1999; **149**:693–705.
6. Curtis D. Use of siblings as controls in case-control association studies. *Annals of Human Genetics* 1997; **61**:319–333.
7. Gauderman W, Witte J, Thomas D. Family-based association studies. *Journal of the National Cancer Institute Monographs* 1999; **26**:31–39.
8. Schaid D. Case-parents design for gene–environment interaction. *Genetic Epidemiology* 1999; **16**:261–273.
9. Breslow N, Day N. *Statistical Methods in Cancer Research: Volume I—The Analysis of Case-control Studies.* IARC Scientific publications: Lyon, 1980.
10. Weinberg C, Umbach D. Choosing a retrospective design to assess joint genetic and environmental contributions to risk. *American Journal of Epidemiology* 2000; **152**:197–203.

11. Pearson E, Hartley H. *Biometrika Tables for Statisticians*, Vol. II. Biometrika Trust: London, 1976, p. 240.
12. Greenland S. Power, sample size, and smallest detectable effect determination for multivariate studies. *Statistics in Medicine* 1985; **4**:117–127.
13. Ottman R. Epidemiologic approach to gene–environment interaction. *Genetic Epidemiology* 1990; **7**: 177–185.
14. Ottman R. Gene–environment interaction: Definitions and study designs. *Preventive Medicine* 1996; **25**: 764–770.
15. Khoury M, Beaty T, Cohen B. *Fundamentals of Genetic Epidemiology*, *Vol. 19.* Oxford University Press: Oxford, 1993.
16. Ursin G, Henderson B, Haile R, Zhou N, Diep A, Bernstein L. Is oral contraceptive use more common in women with BRCA1/BRCA2 mutations than in other women with breast cancer? *Cancer Research* 1997; **57**:3678–3681.
17. Piegorsch W, Weinberg C, Taylor J. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine* 1994; **13**:153–162.
18. Gilliland F, McConnell R, Peters J, Gong Jr H. A theoretical basis for investigating ambient air pollution and children's respiratory health. *Environmental Health Perspectives* 1999; **107**:403–407.
19. Peters J, Avol E, Navidi W, London SJ, Gauderman WJ, Lurmann F, Linn WS, Margolis H, Rappaport E, Gong H Jr., Thomas DC. A study of twelve southern California communities with differing levels and types of air pollution I. Prevalence of respiratory morbidity. *American Journal of Respiratory and Critical Care Medicine* 1999; **159**:760–767.
20. Peters J, Avol E, Gauderman WJ, Linn WS, Navidi W, London SJ, Margolis H, Rappaport E, Vora H, Gong H Jr., Thomas DL. A study of twelve southern California communities with differing levels and types of air pollution II. Effects on pulmonary function. *American Journal of Respiratory and Critical Care Medicine* 1999; **159**:768–775.
21. Cleves M, Olson J, Jacobs K. Exact transmission-disequilibrium tests with multiallelic markers. *Genetic Epidemiology* 1997; **14**:337–347.
22. Lazzeroni L, Lange K. A conditional inference framework for extending the transmission/disequilibrium test. *Human Heredity* 1998; **48**:67–81.
23. Martin E, Kaplan N, Weir B. Tests for linkage and association in nuclear families. *American Journal of Human Genetics* 1997; **61**:439–448.
24. Tu I, Whittemore A. Power of association and linkage tests when the disease alleles are unobserved. *American Journal of Human Genetics* 1999; **64**:641–649.
25. Siegmund K, Langholz B, Kraft P, Thomas D. Testing linkage disequilibrium in sibships. *American Journal of Human Genetics* 2000; **67**: 244–248.
26. Sun F, Cheng R, Flanders W, Yang Q, Khoury M. Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT. *American Journal of Epidemiology* 1999; **150**:97–104.
27. Spielman R, Ewens W. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *American Journal of Human Genetics* 1998; **62**:450–458.
28. Andrieu N, Goldstein A, Thomas D, Langholz B. Counter-matching in gene–environment interaction studies: efficiency and feasibility. *American Journal of Epidemiology* 2000; **153**:265–274.