American Journal of Epidemiology
Vol. 155, No. 5
Printed in U.S.A.

Sample Size Requirements for Association Studies of Gene-Gene Interaction

W. James Gauderman

In the study of complex diseases, it may be important to test hypotheses related to gene-gene ($G \times G$) interaction. The success of such studies depends critically on obtaining adequate sample sizes. In this paper, the author investigates sample size requirements for studies of $G \times G$ interaction, focusing on four study designs: the matched-case-control design, the case-sibling design, the case-parent design, and the case-only design. All four designs provide an estimate of interaction on a multiplicative scale, which is used as a unifying theme in the comparison of sample size requirements. Across a variety of genetic models, the case-only and case-parent designs require fewer sampling units (cases and case-parent trios, respectively) than the case-control (pairs) or case-sibling (pairs) design. For example, the author describes an asthma study of two common recessive genes for which 270 matched case-control pairs would be required to detect a $G \times G$ interaction of moderate magnitude with 80% power. By comparison, the same study would require 319 case-sibling pairs but only 146 trios in the case-parent design or 116 cases in the case-only design. A software program that computes sample size for studies of $G \times G$ interaction and for studies of gene-environment ($G \times E$) interaction is freely available (http://hydra.usc.edu/gxe). *Am J Epidemiol* 2002;155:478–84.

association; case-control studies; genetics; interaction; research design; sample size

In designing an epidemiologic study of genetic risk factors for a complex disease (e.g., cancer, diabetes), an investigator is likely to focus on hypotheses regarding the main effects of specific candidate genes. However, the proposal of candidate genes in the first place is often due to an underlying hypothesis regarding some pathway that leads to disease, a pathway that probably involves more than one gene. For this reason, it may also be important to develop and test hypotheses related to gene-gene ($G \times G$) interaction.

Case-control studies are widely used in epidemiology for studying associations between disease and potential risk factors. It is critical to the success of such studies that an adequate-sized sample be recruited. For case-control studies of gene-environment ($G \times E$) interaction, several authors have developed methods for estimating required sample sizes in both unmatched (1–4) and matched (5, 6) designs. In this paper, I describe a general method of computing required sample size for tests of $G \times G$ interaction in the context of four designs: the matched case-control design (7), the case-sibling design (8–10), the case-parent design (11, 12), and the case-only design (13–15). For a range of genetic models, I provide estimates of sample size needed for detecting $G \times G$ interaction, with the primary goal of comparing requirements across designs to infer their efficiencies relative to one another.

Received for publication April 18, 2001, and accepted for publication August 23, 2001.

Abbreviations: $G \times G$, gene-gene [interaction]; $G \times E$, gene-environment [interaction]; GST, glutathione *S*-transferase.

From the Department of Preventive Medicine, School of Medicine, University of Southern California, 1540 Alcazar Street, Suite 220, Los Angeles, CA 90089 (e-mail: jimg@usc.edu). (Reprint requests to Dr. W. James Gauderman at this address).

METHODS

Notation

Let D be an indicator of disease, g a genotype at a disease-susceptibility locus with alleles "A" and "a," and h a genotype at a second disease-susceptibility locus with alleles "B" and "b." The prevalences of the high-risk alleles at the g and h loci are denoted by q_A and q_B, respectively. I assume that the two candidate loci are unlinked, are in Hardy-Weinberg equilibrium, and are independently distributed in the population. Under these assumptions, the distribution of genotypes g in the population is given by $\Pr(g|q_A) = q_A^2$, $2q_A(1 - q_A)$, and $(1 - q_A)^2$ for g = AA, Aa, and aa, respectively, with analogous definition of $\Pr(h|q_B)$. I define a covariate $G(g) = 0$ for g = aa, $G(g) = 1$ for g = AA, and $G(g) = \delta$ for g = Aa, where at its extremes $\delta = 0$ corresponds to a recessive model and $\delta = 1$ to a dominant model. For notational convenience, I simply use G to denote the function $G(g)$ and H to denote the analogous function $H(h)$ at the second locus. Although I assume either a dominant model or a recessive model in sample-size calculations, one might want to increase flexibility in practice by using a set of two covariates for each locus—for example, G_1 as an indicator of g = Aa and G_2 as an indicator of g = AA.

Sampling designs and likelihood formation

Below I describe the design and analytical approach for each of the four case-based designs considered in this paper.

In each design, cases are subjects affected by the disease of interest, perhaps with restrictions on age of onset, disease subtype, etc.

Matched case-control design. In the matched case-control design, controls are subjects who are unaffected by the disease of interest and are assumed to be genetically unrelated to cases. They should be selected from the same source population as the cases. Since genotype frequencies may vary across ethnic groups, controls will typically be matched to cases according to ethnicity in order to avoid confounding, also known as population stratification bias. Age is also likely to be a matching factor for complex diseases, since disease risk often varies substantially by age.

In the simpler situation of testing for an association between a disease and a single gene, McNemar's χ^2 test and the associated matched odds ratio could be used for analysis. The natural extension of this method to allow modeling of two genes and their interaction is conditional logistic regression (7). The corresponding conditional likelihood in a sample of N matched sets has the form

$$L(\beta_g, \beta_h, \beta_{gh}) = \Pi_{i=1}^{N} \frac{e^{\beta_g G_{i1} + \beta_h H_{i1} + \beta_{gh} G_{i1} H_{i1}}}{\sum_{j \in S(i)} e^{\beta_g G_{ij} + \beta_h H_{ij} + \beta_{gh} G_{ij} H_{ij}}}, \quad (1)$$

where the index "1" refers to the case and the set $S(i)$ includes all subjects in matched set i. The quantities $R_g = \exp(\beta_g)$ and $R_h = \exp(\beta_h)$ are the odds ratios for G when $H = 0$ and H when $G = 0$, respectively. When both $G > 0$ and $H > 0$, $R_{gh} = \exp(\beta_{gh})$ measures the departure from a purely multiplicative odds ratio model—i.e., from a model in which the joint odds ratio for G and H is simply $R_g \times R_h$. I use the term "interaction" in this paper to denote the situation in which $R_{gh} \neq 1$, recognizing that there are other scales of measurement on which one might prefer to assess interaction (16–18). If the matching criteria are relatively coarse, so that several case-control pairs fall into the same matching class, a stratified analysis rather than a pair-matched analysis may be considered (19).

Case-sibling design. In the case-sibling design, controls are selected from unaffected siblings of the case. For a disease with variation in age of onset, an eligible sibling should have attained the age of the case free of disease, which often will restrict the sample to older siblings. While this restriction can be problematic in the analysis of environmental factors if there are secular changes in exposure levels (10), it should not bias a study of two genes and their interaction. The conditional likelihood in equation 1 can be used to estimate odds ratios R_g and R_h and the odds-ratio ratio R_{gh} (9, 10).

Case-parent design. In the case-parent design, genotypes are measured in the parents of the case, but parental disease status is neither required nor used in the analysis. The most commonly used approach to the analysis of a single gene is the transmission disequilibrium test (11), which is equivalent to McNemar's χ^2 test comparing the distributions of alleles transmitted and nontransmitted from parents to the case. As in the case-control settings, this approach can be generalized to the analysis of two or more genes and their interaction(s) using conditional logistic regression (5, 9, 10, 12, 20). The likelihood is the same as that shown in equation 1, where the denominator now includes a contribution from the case and from 15 "pseudosiblings" of the case, the latter formed as the 15 possible joint genotypes that the case could have inherited from the parents but did not. For example, if the father's genotype was Aa/Bb (at loci G/H, respectively), the mother's was aa/bb, and the case's was Aa/Bb, the 15 pseudosibling genotypes would include three copies of Aa/Bb, four of Aa/bb, four of aa/Bb, and four of aa/bb. The $\exp(\beta)$ quantities based on equation 1 represent genetic relative risks (R_g and R_h) and the relative-risk ratio (R_{gh}), rather than odds-ratio parameters as in the case-control design (21). Of course, these will be equivalent to the odds-ratio parameters provided that the disease is rare in all genetic subcategories. Some researchers have described the application of Poisson regression to the case-parent design (22, 23), an approach that can be extended to allow for maternally mediated effects (24) and imprinting (25) and can be used when there are missing parental data (26).

Case-only design. In the case-only design, no controls are selected. Such a sample cannot be used to estimate genetic main effects, but it can be used to test and estimate $G \times E$ (14, 15) or $G \times G$ interaction effects (13). The analysis is a standard χ^2 test of association between the genes, or, equivalently, it can be based on unconditional logistic regression with the likelihood form

$$L(\alpha, \beta_{gh}) = \Pi_{i=1}^{N} \frac{(e^{\alpha + \beta_{gh} H_i})^{G_i}}{1 + e^{\alpha + \beta_{gh} H_i}}. \quad (2)$$

Here, G_i and H_i represent indicators of genetic susceptibility for case i, and the interaction quantity $R_{gh} = \exp(\beta_{gh})$ estimates the relative-risk ratio (13), as in the case-parent design. A key assumption in this design is that there is no association between G and H in the general population.

Calculation of sample size

For each of the four designs, I will provide examples of the minimum number (N) of sampling units that will provide a given power for detecting a gene-gene interaction. Depending on the design, a sampling unit is defined as a case-control pair (design 1), a case-sibling pair (design 2), a case-parent trio (design 3), or simply a case (design 4). The null hypothesis (H_0) is $\beta_{gh} = 0$, i.e., that there is no $G \times H$ interaction on a multiplicative scale. In all models, I assume that the disease is rare enough that the test of the odds-ratio parameter in the case-control and case-sibling designs is equivalent to the test of the relative-risk parameter in the case-parent and case-only designs. I adopt an approach to sample-size determination that has been previously described (6, 27); it is summarized in the Appendix. In all calculations, I assume a significance level of 5% and a power of 80%, and I allow for a two-sided alternative hypothesis. For comparative purposes, I compute the ratio of N for the case-control design to N for each of the other three designs, which provides a measure of asymptotic

relative efficiency per sampling unit of the latter to the former.

Computer software

A colleague (John Morrison) and I have developed a user-friendly Windows-based software program called QUANTO for computing either sample size or power in studies of $G \times G$ or $G \times E$ interaction (28). Inputs to the program include the design (case-control, case-sibling, case-parent, case-only), true model parameters, and the significance level. Required sample size will be computed for a given power or vice versa. The program is available at no charge and may be downloaded from a University of Southern California website (http://hydra.usc.edu/gxe).

EXAMPLES

Study of $G \times G$ interaction for asthma

Gilliland et al. (29) have proposed several candidate genes that may be associated with risk of asthma, including genes in the glutathione *S*-transferase (GST) family, myeloperoxidase, and tumor necrosis factor-α. Although Gilliland et al. proposed these genes for their role in response to air pollution, the possibility of $G \times G$ interaction also exists to the extent that these genes are involved in common pathways to disease.

Suppose one wants to conduct a study that has as one of its aims a test of whether there is an interaction between the genes *GSTM1* and *GSTT1*. For some specific values of model parameters, I will demonstrate sample sizes that would be required by each of the four designs to address this hypothesis. For simplicity, I assume that subjects will be selected from a single population. For both loci, it is the "null/null" genotype that is suspected of increasing asthma risk, indicating use of the recessive model at both loci. I assume that the prevalence of the null/null genotype is 40 percent for *GSTM1* and 25 percent for *GSTT1* (30). Letting $g = GSTM1$ and $h = GSTT1$, the frequencies of the corresponding null alleles in the population are $q_A = 0.63$ (i.e., $\sqrt{0.40}$) and $q_B = 0.5$, respectively. I also assume a pure interaction model, in which neither *GSTM1* nor *GSTT1* increases risk by itself (i.e., $R_g = 1.0$ and $R_h = 1.0$) but risk is increased in subjects with the null/null genotype at both loci ($R_{gh} > 1.0$).

For the four designs and a range of values for R_{gh}, table 1 shows the number of required sampling units. The required sample size in a case-control study exceeds 2,000 pairs when $R_{gh} \leq 1.5$ but declines sharply with increasing magnitude of the interaction strength. The case-sibling design requires a larger sample size than the case-control design, by a factor of approximately 20 percent. The case-parent and case-only designs, however, require substantially fewer sampling units than the other two designs. For instance, when $R_{gh} = 3.0$, the case-control and case-sibling designs would require 270 and 319 matched pairs, respectively, while the case-parent design would require 146 trios and the case-only design 116 cases.

General design comparisons

I compare the four designs across a variety of genetic models to determine whether the relations observed in the asthma example hold more generally. I consider recessive and dominant models, as well as rare and common susceptibility alleles. For the rare gene, I assume that the proportion of susceptible individuals is 1 percent (i.e., $\Pr(G = 1) = 0.01$), so that the underlying susceptibility-allele prevalence is 0.005 under a dominant model and 0.10 under a recessive model. For the common gene, I assume that 25 percent of the population is susceptible, yielding susceptibility-allele prevalences of 0.134 and 0.5 for dominant and recessive models, respectively. I also consider several relative-risk models of the type described by Ottman (31, 32) and Khoury (33). These include the "pure-interaction" model described above, in which risk is only increased when one carries the susceptibility genotype at both loci

TABLE 1. Number (*N*) of sampling units* required for 80% power to detect an interaction of magnitude R_{gh} between the glutathione *S*-transferase genes *GSTM1* and *GSTT1* under four different study designs

GSTM1 × *GSTT1* interaction (R_{gh})†	Study design						
	Case-control (*N*)	Case-sibling		Case-parent		Case-only	
		N	Ratio‡	*N*	Ratio‡	*N*	Ratio‡
1.5	2,008	2,428	0.83	1,174	1.71	947	2.12
2.0	674	807	0.84	381	1.77	305	2.21
2.5	385	458	0.84	213	1.81	169	2.28
3.0	270	319	0.85	146	1.85	116	2.33
3.5	210	248	0.85	112	1.88	89	2.36
4.0	174	205	0.85	92	1.89	73	2.38
5.0	134	158	0.85	70	1.91	55	2.44
10.0	81	95	0.85	41	1.98	32	2.53

* A sampling unit is a pair for the case-control and case-sibling designs, a trio for the case-parent design, and a case for the case-only design.

† Assumptions: a recessive model at both loci, with $\Pr(G = 1) = 0.40$ at *GSTM1* and $\Pr(H = 1) = 0.25$ at *GSTT1* and main-effect relative risks $R_g = 1$ and $R_h = 1$.

‡ Compared with the case-control design; ratios above (below) 1.0 indicate greater (lesser) efficiency.

($R_g = 1$, $R_h = 1$, $R_{gh} > 1$), and various "risk-modification" models in which each locus alone might increase risk but their combined effect is more than multiplicative ($R_g \geq 1$, $R_h \geq 1$, $R_{gh} > 1$).

Table 2 shows the sample sizes required to detect an interaction effect of magnitude $R_{gh} = 3.0$ in each design, assuming a pure-interaction model and various combinations of susceptible-genotype prevalence and dominance model. When genetic susceptibility is rare (0.01) at both loci, the required number of sampling units is impractically large (more than 30,000) regardless of the design. Sample size requirements are substantially less (in the range of 130–400) if susceptibility is common (0.25) at both loci. For the case-control and case-only designs, sample size requirements do not depend on the dominance model, since there is no familial relationship among subjects. The case-sibling design requires more pairs than the case-control design in all models, with asymptotic relative efficiencies ranging from about 0.75 to 0.90. The case-parent and case-only designs are more efficient than the case-control design, with asymptotic relative efficiencies ranging from 1.7 to 2.2 in the former and from 2.5 to 2.7 in the latter. Sample size requirements in the case-sibling and case-parent designs are lower if at least one locus is recessive than if they are both dominant.

Table 3 shows sample sizes required to detect an interaction effect of magnitude $R_{gh} = 3.0$ in various risk-modification models, i.e., for different values of the main-effect relative risks R_g and R_h. In most models considered, the case-parent design requires less than half the number of sampling units as the case-control design. The case-only design is clearly most efficient, with asymptotic relative efficiencies that range from 2.4 to 5.1 and that increase with increasing magnitudes of R_g and R_h. The case-sibling design ranges from being slightly less efficient to slightly more efficient than the case-control design.

DISCUSSION

With recent advances in technology, large-scale epidemiologic studies of genes are now possible. These studies can be quite expensive to conduct, and it is therefore essential to use study designs that make the most efficient use of available resources. For testing hypotheses about $G \times G$ interaction, the results presented in this paper indicate that the case-only and case-parent designs can be substantially more efficient than the case-sibling design or the standard matched case-control design. For a variety of genetic models, the case-only and case-parent designs typically require less than half the number of sampling units as the other two designs to achieve the same statistical power. However, nonstatistical issues, including the costs of recruiting control subjects, measuring genotypes, and verifying phenotypes, will also be factors in choosing a preferred design. For example, genotyping costs per matched set will be 50 percent higher for the case-parent design than for the case-control or case-sibling design. This increased cost may be partially offset by the fact that no phenotyping on parents is required. For a childhood disease, recruiting parents or siblings may require less effort than obtaining an unrelated control subject, while for an adult-onset disease the reverse may hold if relatives are geographically distant from cases.

For characterizing genetic and interactive effects, designs alternative to those considered in this paper have been suggested. These include hybrid designs such as those combining the case-parent and case-sibling designs (26, 34), the case-parent and case-control designs (35), and the case-sibling and case-control designs (36). Other investigators have developed methods for the analysis of whole pedigrees, allowing for collection of genetic data on only a subset of family members (37, 38). In the context of the case-control design, Andrieu et al. (39) have proposed counter-matching of cases to controls as a technique for using available data at the time of sampling (e.g., family

TABLE 2. Number (*N*) of sampling units* required for 80% power to detect a gene-gene interaction with magnitude $R_{gh} = 3.0$, assuming $R_g = 1.0$, $R_h = 1.0$, and various genetic-susceptibility prevalences and dominance models

Proportion susceptible		Dominance model		Study design						
				Case-control	Case-sibling		Case-parent		Case-only	
Pr($G = 1$)	Pr($H = 1$)	*G*	*H*	(*N*)	*N*	Ratio†	*N*	Ratio†	*N*	Ratio†
0.01	0.01	Dominant	Dominant	77,914	103,193	0.76	46,877	1.66	31,244	2.49
		Dominant	Recessive	77,914	90,932	0.86	39,368	1.98	31,244	2.49
		Recessive	Recessive	77,914	85,026	0.92	35,751	2.18	31,244	2.49
0.01	0.25	Dominant	Dominant	4,964	6,093	0.81	2,743	1.81	1,860	2.67
		Dominant	Recessive	4,964	5,811	0.85	2,557	1.94	1,860	2.67
		Recessive	Dominant	4,964	5,572	0.89	2,342	2.12	1,860	2.67
		Recessive	Recessive	4,964	5,430	0.91	2,255	2.20	1,860	2.67
0.25	0.25	Dominant	Dominant	312	389	0.80	183	1.70	128	2.44
		Dominant	Recessive	312	374	0.83	171	1.82	128	2.44
		Recessive	Recessive	312	362	0.86	162	1.93	128	2.44

* A sampling unit is a pair for the case-control and case-sibling designs, a trio for the case-parent design, and a case for the case-only design.

† Compared with the case-control design; ratios above (below) 1.0 indicate greater (lesser) efficiency.

TABLE 3. Number (*N*) of sampling units* required for 80% power to detect a gene-gene interaction with magnitude $R_{gh} = 3.0$, assuming various genetic-susceptibility prevalences and main-effect relative risks

Proportion susceptible†		Main-effect relative risk		Study design						
				Case-control	Case-sibling		Case-parent		Case-only	
Pr($G = 1$)	Pr($H = 1$)	R_g	R_h	(*N*)	*N*	Ratio‡	*N*	Ratio‡	*N*	Ratio‡
0.25	0.25	1	1	312	374	0.83	171	1.82	128	2.44
		1	2	321	373	0.86	161	1.99	116	2.77
		2	1	321	367	0.87	163	1.97	116	2.77
		2	2	341	381	0.90	161	2.12	109	3.13
0.01	0.25	1	1	4,964	5,811	0.85	2,557	1.94	1,860	2.67
		1	2	4,892	5,513	0.89	2,249	2.18	1,548	3.16
		2	1	4,061	3,826	1.06	1,448	2.80	960	4.23
		2	2	4,066	3,684	1.10	1,296	3.14	802	5.07
0.25	0.01	1	1	4,964	5,572	0.89	2,342	2.12	1,860	2.67
		1	2	4,061	4,063	1.00	1,352	3.00	960	4.23
		2	1	4,892	5,166	0.95	2,063	2.37	1,548	3.16
		2	2	4,066	3,821	1.06	1,227	3.31	802	5.07

* A sampling unit is a pair for the case-control and case-sibling designs, a trio for the case-parent design, and a case for the case-only design.

† Assumed dominance models are dominant for locus *G* and recessive for locus *H.*

‡ Compared with the case-control design; ratios above (below) 1.0 indicate greater (lesser) efficiency.

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 with the designs considered in this paper.

The investigator planning a new study is likely to have parameter choices that differ from the specific values used in this paper. For this reason, my colleague and I distribute software that investigators may use to compute power or required sample size for their particular design parameters. For unmatched case-control studies of $G \times E$ interaction, Garcia-Closas and Lubin (1) also distribute a software program for computing sample size or power. Their program could be used to obtain sample size for an unmatched case-control study of $G \times G$ interaction, with their *"E"* being replaced by the second gene and exposure prevalence being replaced by the corresponding prevalence of the susceptibility genotype. However, their program is not directly applicable to a matched case-control study, and it will not provide calculations for the case-sibling, case-parent, or case-only design.

In the sample size comparisons, I assumed that the loci *G* and *H* were independently transmitted from parents to offspring (unlinked) and that they were independently distributed in the population (no disequilibrium). The linkage assumption will be violated if the two genes are in close physical proximity to one another. The disequilibrium assumption can be violated for this same reason, or because of other mechanisms that cause correlation among alleles in the population (e.g., admixture or selective forces that favor or discourage specific alleles at both loci). For each design, I describe below how the validity of the test of $G \times G$ interaction is affected by deviations from these assumptions.

Case-control and case-sibling designs. The case-control and case-sibling designs are valid in the presence of linkage and/or disequilibrium.

Case-parent design. The case-parent design is valid when there is disequilibrium between *G* and *H* but invalid if there is linkage. The problem if there is linkage is that the 16 possible pseudosibling genotypes are not equally likely under the null hypothesis of no genetic effects; rather, the distribution of genotypes depends on the recombination fraction (θ) between *G* and *H*. If θ were known, which may be possible given that *G* and *H* have known chromosomal locations, a valid test could be recovered by including it as an offset to the pseudosibling genotype distribution. Determining the sensitivity of a $G \times G$ interaction test to misspecification of θ requires further investigation.

Case-only design. The case-only design is valid if there is linkage between *G* and *H* but invalid if there is disequilibrium. The problem with the latter is that a population-level association between *G* and *H* will also be reflected in a case series, in the absence of any interaction.

In practice, the linkage assumption is easily assessed, since the investigator will know (approximately) the chromosomal locations of *G* and *H*. If *G* and *H* are on different chromosomes, for example, the two loci are unlinked with certainty. However, it will be difficult to evaluate the disequilibrium assumption unless one has genotypic data for *G* and *H* on a random sample of persons drawn from the same population as the people under study. Finally, it is possible that one will not observe *G* and *H* directly but rather markers M_1 and M_2 that are in linkage disequilibrium with *G* and *H*, respectively. The same conditions for test validity in each design apply to analysis of $M_1 \times M_2$ interaction, although one will suffer a loss in power relative to a study of *G* and *H* directly. For a test of association between disease and *G* using M_1 (or *H* using M_2), it is known that one must modify the analytical approach in the following two situations: 1) when there is *m*:*n* matching (with $m \neq 1$

and/or $n \neq 1$) in the case-sibling design (40) and 2) when there are two or more affected offspring in the case-parent design (41). In these two situations, similar corrections will be required for valid analysis of $M_1 \times M_2$ interaction.

In the example calculations, I assumed that all subjects were obtained from a single population. However, there may be variations across subgroups of the population (e.g., ethnic groups) in overall disease prevalence and in susceptible-allele prevalences (q_A, q_B). In fact, this may be the reason one chooses to use a matched design. The sample size and power calculation approach described above may be modified to account for population stratification, by including in equation A1 (see Appendix) the stratum-specific parameters and the frequency of each stratum in the population (6). Although the absolute sample sizes depend on these additional parameters, the relative efficiencies among designs are similar (calculations not shown). However, one should be prepared to assume that the relative risks R_g, R_h, and R_{gh} are the same in all population subgroups. If this cannot be assumed, one should conduct separate sampling and estimation within each stratum, since there is no single set of parameters to estimate.

Previous papers have focused on design comparisons for case-control studies of genetic main effects (9, 10) and $G \times E$ interactions (5, 6). For testing of genetic main effects, the case-parent design is typically more efficient than the matched case-control design, and both are more efficient than the case-sibling design. For testing of $G \times E$ interaction, the case-parent design is also more efficient than the case-control design, but the case-sibling design can be the most efficient provided that there is not a high degree of sharing of the environmental exposure between siblings (6). These findings, in addition to those presented in this paper, indicate that the case-parent design is a good choice for studies of genetic main and interaction effects. The case-only design might best be viewed as a screening tool with which to identify promising interactions, with follow-up by one of the other three designs to rule out the possibility of population association between genes.

ACKNOWLEDGMENTS

This work was supported in part by grants ES10421 and 5P30 ES07048-03 from the National Institute of Environmental Health Sciences and grant CA52862 from the National Cancer Institute.

REFERENCES

1. Garcia-Closas M, Lubin J. Power and sample size calculations in case-control studies of gene-environment interactions: comments on different approaches. Am J Epidemiol 1999;149:689–92.
2. Foppa I, Spiegelman D. Power and sample size calculations for case-control studies of gene-environment interactions with a polytomous exposure variable. Am J Epidemiol 1997;146:596–604.
3. Goldstein AM, Falk RT, Korczak JF, et al. Detecting gene-environment interactions using a case-control design. Genet Epidemiol 1997;14:1085–9.
4. Hwang S, Beaty T, Liang K, et al. Minimum sample size estimation to detect gene-environment interaction in case-control designs. Am J Epidemiol 1994;140:1029–37.
5. Schaid DJ. Case-parents design for gene-environment interaction. Genet Epidemiol 1999;16:261–73.
6. Gauderman W. Sample size calculations for matched case-control studies of gene-environment interaction. Stat Med 2002;21:35–50.
7. Breslow N, Day N. Statistical methods in cancer research. I. The analysis of case-control studies. (IARC scientific publication no. 32). Lyon, France: International Agency for Research on Cancer, 1980.
8. Curtis D. Use of siblings as controls in case-control association studies. Ann Hum Genet 1997;61:319–33.
9. Gauderman WJ, Witte JS, Thomas DC. Family-based association studies. J Natl Cancer Inst Monogr 1999;(26):31–7.
10. Witte JS, Gauderman WJ, Thomas DC. Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. Am J Epidemiol 1999;149:693–705.
11. Spielman R, McGinnis R, Ewens W. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506–16.
12. Schaid D. General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 1996;13:423–49.
13. Yang Q, Khoury M, Sun F, et al. Case-only design to measure gene-gene interaction. Epidemiology 1999;10:167–70.
14. Khoury M, Flanders D. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! Am J Epidemiol 1996;144:207–13.
15. Piegorsch W, Weinberg C, Taylor J. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med 1994;13:153–62.
16. Rothman K, Greenland S, Walker A. Concepts of interaction. Am J Epidemiol 1980;112:467–70.
17. Siemiatycki J, Thomas D. Biological models and statistical interactions: an example from multistage carcinogenesis. Int J Epidemiol 1981;10:383–7.
18. Rothman KJ, Greenland S. Modern epidemiology. 2nd ed. Philadelphia, PA: Lippincott Williams and Wilkins, 1998.
19. Brookmeyer R, Liang K, Linet M. Matched case-control designs and overmatched analyses. Am J Epidemiol 1986;124:693–701.
20. Self S, Longton G, Kopecky K, et al. On estimating HLA/disease association with application to a study of aplastic anemia. Biometrics 1991;47:53–61.
21. Weinberg C. Re: "Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interaction: basic family designs." (Letter). Am J Epidemiol 2000;152:689–90.
22. Weinberg CR, Wilcox AJ, Lie RT. A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. Am J Hum Genet 1998;62:969–78.
23. Umbach D, Weinberg CR. The use of case-parent triads to study joint effects of genotype and exposure. Am J Hum Genet 2000;66:251–61.
24. Wilcox AJ, Weinberg CR, Lie RT. Distinguishing the effects of maternal and offspring genes through studies of "case-parent-triads." Am J Epidemiol 1998;148:893–901.
25. Weinberg CR. Methods for detection of parent-of-origin effects in genetic studies of case-parent triads. Am J Hum Genet 1999;65:229–35.
26. Weinberg CR. Allowing for missing parents in genetic studies of case-parent triads. Am J Hum Genet 1999;64:1186–93.

27. Longmate J. Complexity and power in case-control association studies. Am J Hum Genet 2001;68:1229–37.
28. Gauderman W, Morrison J. QUANTO documentation. (Technical report no. 157). Los Angeles, CA: Department of Preventive Medicine, University of Southern California, 2001.
29. Gilliland F, McConnell R, Peters J, et al. A theoretical basis for investigating ambient air pollution and children's respiratory health. Environ Health Perspect 1999;107:403–7.
30. Cotton S, Sharp L, Little J, et al. Glutathione *S*-transferase polymorphisms and colorectal cancer: a HuGE review. Am J Epidemiol 2000;151:7–32.
31. Ottman R. Epidemiologic approach to gene-environment interaction. Genet Epidemiol 1990;7:177–85.
32. Ottman R. Gene-environment interaction: definitions and study designs. Prev Med 1996;25:764–70.
33. Khoury MJ. Genetic and epidemiologic approaches to the search for gene-environment interaction: the case of osteoporosis. (Editorial). Am J Epidemiol 1998;147:1–2.
34. Spielman R, Ewens W. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 1998;62:450–8.
35. Martin E, Kaplan N. A Monte Carlo procedure for two-stage tests with correlated data. Genet Epidemiol 2000;18:48–62.
36. Andrieu N, Goldstein A. A case-combined design using both population-based and related controls: a potential alternative for increasing power in gene-environment interaction detection. Genet Epidemiol 2000;19:235–6.
37. Martin E, Monks S, Warren L, et al. A test for linkage and association in general pedigrees: the pedigree disequilibrium test. Am J Hum Genet 2000;67:146–54.
38. Gail M, Pee D, Benichou J, et al. Designing studies to estimate the penetrance of an identified autosomal dominant mutation: cohort, case-control, and genotype-proband designs. Genet Epidemiol 1999;16:15–39.
39. Andrieu N, Goldstein A, Thomas D, et al. Counter-matching in studies of gene-environment interaction: efficiency and feasibility. Am J Epidemiol 2001;153:265–74.
40. Siegmund K, Langholz B, Kraft P, et al. Testing linkage disequilibrium in sibships. Am J Hum Genet 2000;67:244–8.
41. Martin E, Kaplan N, Weir B. Tests for linkage and association in nuclear families. Am J Hum Genet 1997;61:439–48.
42. Greenland S. Power, sample size, and smallest detectable effect determination for multivariate studies. Stat Med 1985; 4:117–27.

APPENDIX

Approach to Computation of Required Sample Sizes

For a single sampling unit, the expected value of the likelihood in equation 1 (designs 1–3) or equation 2 (design 4) is computed as

$$E(\ln[L(\boldsymbol{\beta})]) = \sum_{\underline{g}} \sum_{\underline{h}} \ln[L(\boldsymbol{\beta}; \underline{G}, \underline{H})] \Pr(\underline{g}, \underline{h}|\underline{D}, \alpha^*, \boldsymbol{\beta}^*, q_A, q_B). \quad \text{(A1)}$$

The summation is over all possible observable genotypes $\underline{g}$ and $\underline{h}$ in a sampling unit, and the first factor $L(\boldsymbol{\beta}; \underline{G}, \underline{H})$ is the contribution to the likelihood for a single sampling unit with a specific realization of $\underline{g}$ and $\underline{h}$. The factor $\Pr(\underline{g}, \underline{h}|\underline{D}, \alpha^*, \boldsymbol{\beta}^*, q_A, q_B)$ is the probability distribution for the observable genotypes, conditional on the disease-based ascertainment rule and the true baseline disease rate parameter α^* and genetic-effect parameters $\boldsymbol{\beta}^* = \{\beta_g^*, \beta_h^*, \beta_{gh}^*\}$. By Bayes' rule, this second factor is proportional to $\Pr(\underline{D}|\underline{G}, \underline{H}, \alpha^*, \boldsymbol{\beta}^*) \Pr(\underline{g}, \underline{h}|q_A, q_B)$. For the true disease model $\Pr(\underline{D}|\cdot)$, I assume a logistic form in the case-control and case-sibling designs and a log-linear form in the case-parent and case-only designs. This is done so that the $\boldsymbol{\beta}$ parameters estimated in the likelihoods of equation 1 or equation 2 are unbiased estimates of the corresponding $\boldsymbol{\beta}^*$ parameters, even if the disease is not rare. However, to facilitate comparisons among designs, I assume a rare disease in the calculations (population disease prevalence 1/10,000) so that the odds ratio parameters in designs 1 and 2 are equivalent to the relative risk parameters in designs 3 and 4. For the case-only design, $\Pr(\underline{g}, \underline{h}|q_A, q_B) = \Pr(\underline{g}|q_A) \Pr(\underline{h}|q_B)$ under the assumption of no disequilibrium between loci. The form of the joint probability of $\underline{g}$ and $\underline{h}$ for the other three designs is given elsewhere (6).

For each design, I maximize the expected log-likelihood in equation A1 twice, once letting β_{gh} be a free parameter and once fixing $\beta_{gh} = 0$ (i.e., under H_0). I let $\hat{L}^1$ and $\hat{L}^0$ denote the maximum values of the corresponding log-likelihoods. In both maximizations, β_g and β_h are free parameters. The quantity $\Lambda = 2(\hat{L}^1 - \hat{L}^0)$ is the expected likelihood ratio test statistic for a single sampling unit, and $N\Lambda$ is the noncentrality parameter of the χ^2 distribution under the alternative hypothesis (27, 42). Since I assume that each gene can be coded by a single covariate, sample size may be computed as

$$N = (z_{a/2} + z_b)^2/\Lambda, \quad \text{(A2)}$$

where z_u denotes the $(1 - u)$th percentile of the standard normal distribution and a and $1 - b$ are the significance level and power, respectively. One could also use equation A2 to estimate power for a given N—for example, to determine whether a completed study that failed to find a significant effect had reasonable power to detect an interaction of a specified magnitude. In the context of a case-parent study, one might also want to condition power estimation on an observed distribution of genotypes in the parents, which would eliminate the need for the Hardy-Weinberg equilibrium assumption.