

Detecting Gene-Environment Interactions Using a Case-control Design

Alisa M. Goldstein, Roni T. Falk, Jeannette F. Korczak, and Jay H. Lubin

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

We assessed the sample size required for detecting gene-environment ($G \times E$) interactions in a case-control study of complex diseases. The results suggest that large numbers of cases and controls will be required to detect $G \times E$ interaction for some odds ratio and exposure frequency combinations. These and other results suggest that alternative study designs may be needed to detect $G \times E$ interaction particularly with rare genes or uncommon environmental exposures.

© 1997 Wiley-Liss, Inc.

Key words: case-control studies, environmental exposures, genes, interactions

INTRODUCTION

The study of diseases with complex origins involves many approaches to identify the genetic and environmental contributions to disease risk in populations. For instance, using genetic marker and phenotypic information from families, linkage analyses localize disease gene(s); the gene(s) may then be cloned and mutations responsible for functional abnormalities identified. While gene mutations identified in this manner are likely to be very strong predictors of disease in these families, they may be quite rare in the population, accounting for only a small portion of the disease incidence. Alternatively, genes which are polymorphic, relatively common, and in a biologically plausible pathway for disease (e.g., in cancer, susceptibility genes), may be candidates for study of disease in populations, particularly when known environmental exposures are suspected of acting via the same pathway. In either case, population-based studies must be designed to maximize the power to study both genetic and environmental risk factors for disease. The purpose of this report is to assess the power/sample size requirements for a case-control study of gene-environment ($G \times E$) interactions in complex diseases.

Address reprint requests to Dr. Alisa M. Goldstein, Genetic Epidemiology Branch, EPN 439, 6130 Executive Blvd., Bethesda, MD 20892-7372.

© 1997 Wiley-Liss, Inc.

METHODS

To examine the gene-exposure-disease relationship, we wish to design a case-control study with a 1:1 ratio of cases to controls. We define the parameters for modeling the exposure-disease relation as: $P(g)$ = frequency of the genetic factor g in the population; $P(e)$ = frequency of the environmental factor e in the population; $P(D)$ = frequency of the disease D in the population.

Assume OR_{eg} is the odds ratio for exposure e and genetic status g , with $e = 1$ denoting **exposed** and $e = 0$ nonexposed, and with $g = 1$ denoting the presence of the genetic factor and $g = 0$ the absence of the factor. By definition, $OR_{00} = 1$. Without loss of generality, assume the exposure and the genetic factor are "harmful," i.e., $OR_{10} > 1$ and $OR_{01} > 1$.

A fundamental component of the sample size/power problem is the specification of the joint odds ratio, OR_{11} , under the null and alternative hypotheses. Suppose I_{00} , I_{10} , I_{01} , and I_{11} are the disease rates for nonexposed without the genetic factor ($e = 0, g = 0$), exposed without the genetic factor ($e = 1, g = 0$), nonexposed with the genetic factor ($e = 0, g = 1$), and exposed with the genetic factor ($e = 1, g = 1$), respectively. We wish to define I_{11} in terms of the other rates. One characterization of the joint association is additive, where I_{11} is the sum of the background disease rate and the excess rates for the exposure and for the genetic factor, that is,

$$I_{11} = I_{00} + (I_{10} - I_{00}) + (I_{01} - I_{00}) = I_{10} + I_{01} - I_{00} \quad (1)$$

Dividing (1) by I_{00} gives an expression of relative risks, which, for rare diseases is approximated by odds ratios, namely,

$$OR_{11} = OR_{10} + OR_{01} - 1 \quad (2)$$

An alternate characterization for the joint association is multiplicative, where I_{11} is the product of the risks for the individual factors, i.e., $I_{11} = I_{10} \times I_{01}$. Again, dividing by I_{00} and assuming rare diseases, the multiplicative association is approximated by

$$OR_{11} = OR_{10} \times OR_{01} \quad (3)$$

The precise characterization of the joint association has important consequences, because in a multiplicative relationship, the effect of exposure on risk depends on the gene status, while in an additive model the effect of exposure is unrelated to gene status. For addressing public health concerns regarding disease frequency reduction, deviations from additivity appear to have the most relevance. On the other hand, multiplicative models are more often used in studies of disease etiology [Kleinbaum et al., 1982].

In designing an epidemiologic study to detect a "gene-environment interaction," the meaning of interaction must be specified explicitly, because interaction is a model-dependent concept. For purposes of illustration, we create two different test situations [Lubin and Gail, 1990]. In the first, we assume that the multiplicative model (3) is the alternative hypothesis and defines the true state. The null hypothesis is the

additive model (2). In the second, we let the multiplicative model (3) be the null hypothesis, with an alternative which is greater or less than multiplicative, namely,

$$OR_{11} = OR_{10} \times OR_{01} \times \Psi \tag{4}$$

where Ψ is given a specific value that defines the precise alternative hypothesis. The factor Ψ has been referred to as the odds ratio of interaction.

To estimate the power/sample size, we use equation 17 from Lubin and Gail [1990]. We assume e and g are independent and the two-sided type I error is 5%. We use the two test situations (Table I) described above to assess sample size using the following for illustration: P(g) = 0.01, 0.10, 0.50; P(e) = 0.30; P(D) = 0.07; $\Psi = 3.0$.

RESULTS

Panels A-F in Figure 1 show the power/sample size needed to detect a multiplicative interaction between g and e when the null association is additive (panels A-C) or greater than a multiplicative interaction (panels D-F) with an interaction coefficient = 3 when the null association is multiplicative. Larger sample sizes are required to discriminate a multiplicative interaction (H_A) from an additive one (H_0) when OR_{10} and OR_{01} are smaller, i.e., 2 vs 10 (Panels A-C). In contrast, larger sample sizes are needed to differentiate a greater than multiplicative interaction (H_A) from a multiplicative association (H_0) when OR_{10} and OR_{01} are larger, i.e., 10 vs 2. As can be seen from the panels, certain combinations of odds ratios, and frequencies of g and e make study requirements for a case-control study prohibitive. For instance, with a rare gene (P(g) = 0.01) and 80% power, many more than 10,000 study subjects are needed except when $OR_{10} = OR_{01} \geq 10$ (panel A). With higher gene frequencies, sample size requirements are more reasonable. We note, however, that all panels presented were calculated assuming a 1:1 control:case ratio. Sample size estimation for detecting G × E interactions is very complicated when there is an unequal control:case ratio and may require larger or smaller numbers of total subjects. Regardless of the control:case ratio, studies of rare factors would still require prohibitive sample sizes (results not shown).

DISCUSSION

We examined the sample size and power required to detect 1) a multiplicative interaction between a dichotomous genetic factor g and environmental factor e when the null association is assumed additive and 2) greater than multiplicative interaction when the null association is assumed multiplicative using specific odds ratios and frequencies

TABLE I. Odds Ratios Under the Alternative (H_A) Hypotheses to Assess Sample Size/Power in a Case-control Study for the Two Design Settings. In the Absence of Both e and g, $OR_{00} = 1$

OR ₁₀	OR ₀₁	Design setting 1	Design setting 2
		$H_A:OR_{11}=OR_{10} \times OR_{01}$	$H_A:OR_{11}=OR_{10} \times OR_{01} \times 3$
2	2	4	12
10	2	20	60
2	10	20	60
10	10	100	300

Fig. 1. Sample size and power needed to detect a multiplicative interaction between g and e when the null association is additive (panels A-C) or greater than a multiplicative interaction with an interaction coefficient $\Psi = 3$ when the null association is multiplicative (panels D-F). Three gene frequencies are used for illustrative purposes: 0.01, 0.10 and 0.50. For all panels, exposure prevalence = 0.30 and the control:case ratio is 1:1.

of these exposures. The results suggest that large numbers of cases and controls will be required to detect $G \times E$ interaction for some combinations of odds ratios and exposure frequencies. For example (from panels A and D), assuming a rare gene with a frequency of 0.01 such as BRCA1 185delAG among Ashkenazi Jews [Struewing et al., 1995], $OR_{01} = 2$ and $OR_{10} = 2$ which is consistent with BRCA1 risk [Fitzgerald et al., 1996] and some environmental exposure effects for breast cancer (e.g., alcohol use, age at menopause) [Kelsey, 1993], would require approximately 60,000 cases and an equal number of controls to detect a multiplicative interaction between g and e . Detecting a greater than multiplicative interaction with $\Psi = 3$ when the null association is multiplicative would require approximately 12,000 subjects.

Hwang et al. [1994] also examined the minimum sample size estimate to detect $G \times E$ interaction in case-control designs, using a control:case ratio of 2:1 and the multiplicative model (2) as the null hypothesis, with the alternative hypothesis the more complex $\Psi \neq 1$ and OR_g (in the nonexposed subgroup) = $OR_{01} = 1$. The results were similar to what was shown here under different conditions, suggesting that case-control designs may be used to detect greater than or less than multiplicative $G \times E$ interaction only when there are both a common environmental factor and a common genetic factor.

These results and others suggest that alternative approaches to traditional case-control studies may be needed to detect $G \times E$ interaction, particularly with rare genes or uncommon environmental exposures. These alternative approaches may include cohort designs, 2-tier sampling strategies [Weinberg and Wacholder, 1990; Weinberg and Sandler, 1991], case-only designs [Khoury and Flanders, 1996], and family- or sibling-based designs [Andrieu and Goldstein, 1996]. Additional studies to assess sample size/power issues using these alternative design approaches will be conducted.

REFERENCES

- Andrieu N, Goldstein AM (1996): Use of relatives of cases as controls to identify risk factors when an interaction between environmental and genetic factors exists. *Int J Epidemiol* 25:649-657.
- Fitzgerald MG, MacDonald DJ, Krainer M, et al. (1996): Germ-line BRCA1 mutations in Jewish and non-Jewish women with early-onset breast cancer. *N Engl J Med* 334:143-149.
- Hwang S-J, Beaty TH, Liang K-Y, Coresh J, Khoury MJ (1994): Minimum sample size estimation to detect gene-environment interaction in case-control designs. *Am J Epidemiol* 140:1029-1037.
- Kelsey JL (ed) (1993): *Breast Cancer*. *Epidemiol Rev* 15:1-263.
- Khoury MJ, Flanders WD (1996): Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! *Am J Epidemiol* 144:207-213.
- Kleinbaum DG, Kupper LL, Morgenstern H (1982): "Epidemiologic Research. Principles and Quantitative Methods." Belmont, CA: Lifetime Learning Publications.
- Lubin JH, Gail MH (1990): On power and sample size for studying features of the relative odds of disease. *Am J Epidemiol* 131:552-566.
- Struewing JP, Abeliovich D, Peretz T, et al. (1995): The carrier frequency of the BRCA1 185delAG mutation is approximately 1 percent in Ashkenazi Jewish individuals. *Nat Genet* 11:198-200.
- Weinberg CR, Sandler DP (1991): Randomized recruitment in case-control studies. *Am J Epidemiol* 134:421-432.
- Weinberg CR, Wacholder S (1990): The design and analysis of case-control studies with biased sampling. *Biometrics* 46:963-975.