American Journal of
EPIDEMIOLOGY

Volume 149
Number 8
April 15, 1999

Sponsored by the Society for Epidemiologic Research

COMMENTARY

Power and Sample Size Calculations in Case-Control Studies of Gene-Environment Interactions: Comments on Different Approaches

Montserrat García-Closas and Jay H. Lubin

Power and sample size considerations are critical for the design of epidemiologic studies of gene-environment interactions. Hwang et al. (*Am J Epidemiol* 1994;140:1029–37) and Foppa and Spiegelman (*Am J Epidemiol* 1997;146:596–604) have presented power and sample size calculations for case-control studies of gene-environment interactions. Comparisons of calculations using these approaches and an approach for general multivariate regression models for the odds ratio previously published by Lubin and Gail (*Am J Epidemiol* 1990;131:552–66) have revealed substantial differences under some scenarios. These differences are the result of a highly restrictive characterization of the null hypothesis in Hwang et al. and Foppa and Spiegelman, which results in an underestimation of sample size and overestimation of power for the test of a gene-environment interaction. A computer program to perform sample size and power calculations to detect additive or multiplicative models of gene-environment interactions using the Lubin and Gail approach will be available free of charge in the near future from the National Cancer Institute. *Am J Epidemiol* 1999;149:689–92.

case-control studies; epidemiologic methods; sample size; statistical power

The evaluation of gene-environment interactions is becoming a central theme in epidemiologic studies of complex diseases with both genetic and environmental determinants (1). Khoury et al. (2) and Ottman et al. (3) have proposed several biologically plausible models to describe the relation between genetic and environmental determinants of disease. Power and sample size considerations are critical for the statistical evaluation of these models of interaction. Two recent papers have presented power and sample size calculations for such studies. Hwang et al. (4) presented calculations for binary genetic and environmental factors based on a previously published formulae (5). Foppa and Spiegelman (6) presented formulae for a binary genetic factor and an environmental exposure with multiple categories. Most of the sample size calculations presented by Hwang et al. (4) correspond to a model of interaction where the environmental factor influences the risk of disease, whereas the genetic factor exacerbates the effect of the environmental factor but does not have an effect on disease risk in the absence of the environmental factor (pattern 2 in Khoury et al. (2) or model B in Ottman et al. (3)). On the other hand, calculations presented by Foppa and Spiegelman (6) correspond to a model of interaction where both the genetic and environmental factors increase the risk of disease, and the combined effect of both factors is larger than would be expected under a multiplicative model (pattern 4 in Khoury et al. (2) or model E in Ottman et al. (3)).

Lubin and Gail (7) have presented power and sample size formulae for general multivariate regression models for the odds ratio, which can be used for calculations to detect interactions in logistic risk models. In particular, the Lubin and Gail formulae subsume the

Received for publication August 20, 1998, and accepted for publication November 13, 1998.

Abbreviations: OR, odds ratio; FP, Foppa and Spiegelman; LG, Lubin and Gail; MLE, maximum likelihood estimate.

From the Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD.

Reprint requests to Dr. Montserrat García-Closas, Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6130 Executive Blvd., EPN 443, Bethesda, MD 20892-7374.

design situations discussed by Hwang et al. and Foppa and Spiegelman in the context of gene-environment interactions.

Comparisons of calculations using the Lubin and Gail formulae and the formulae by Foppa and Spiegelman have revealed substantial differences. Here, we illustrate these differences for an example presented in Foppa and Spiegelman's paper, explain the source of the differences, and conclude with some remarks.

FOPPA AND SPIEGELMAN'S APPROACH

For comparison purposes, we use the same notation and assumptions as Foppa and Spiegelman. Assume D and G are binary indicators of disease status and genetic susceptibility, respectively, taking values 1 or 0 for presence or absence of the characteristic, and the categorical exposure E has Q levels taking the quantitative values $0,1,\ldots,Q-1$. The odds ratio (OR) for the gene effect in the lowest exposure category is denoted by $OR_{G=1|E=0}$. The exposure effect for level 1 in non-susceptible subjects is denoted by $OR_{E=1|G=0}$. For increasing exposure levels, the exposure odds ratio is a power of $OR_{E=1|G=0}$, i.e., the log odds ratio of disease is a linear function of exposure. The "top-to-bottom quantile" contrast of the exposure effect in non-susceptible subjects ($OR^{tb}_{E|G=0}$) is defined as $(OR_{E=1|G=0})^{Q-1}$. The gene-environment interaction effect at exposure level 1 is denoted by θ and the top-to-bottom quantile interaction effect (θ^{tb}) is defined as $(\theta)^{Q-1}$. The Foppa and Spiegelman approach (denoted as the FS approach) can be embedded within the general framework of sample size and power based on a standard multivariate logistic regression model:

$$\text{Logit}\,[P(D=1|E,G)] = \beta_0 + \beta_E E + \beta_G G + \beta_{EG} EG, \quad (1)$$

where $\beta_G = \ln(OR_{G=1|E=0})$, $\beta_E = \ln(OR_{E=1|G=0}) = \ln(OR^{tb}_{E|G=0})/(Q-1)$ and $\beta_{EG} = \ln(\theta) = \ln(\theta^{tb})/(Q-1)$.

Table 1 depicts the odds ratios for the association between disease risk and a binary genetic factor and an environmental exposure categorized in quintiles, when non-susceptible unexposed subjects are the reference category.

Table 2 shows sample sizes required to achieve 80 percent power (two-sided test with 5 percent Type I error) to detect an interaction between a genetic factor with a 50 percent frequency in the population and an environmental exposure categorized in quintiles (Q = 5). In this example, the genetic and environmental factors are independent in the population, $OR_{G=1|E=0}$ is 1.5 and the control to case ratio is one. Table 2 shows that the study sizes calculated with the FS approach are lower than the study sizes calculated with the Lubin and Gail approach (denoted as the LG approach) and these differences increase with the size of the interaction effect, θ^{tb}, and, for a given θ^{tb}, with the size of the exposure effect among non-susceptible subjects, $OR^{tb}_{E|G=0}$.

The differences in the required sample sizes are the result of a highly restrictive characterization of the null hypothesis in the FS approach, which, we assert, is generally not the null hypothesis of primary interest in most case-control studies.

As indicated in Lubin and Gail (7), the design of a case-control study to test for an interaction begins by specifying an alternative hypothesis, denoted as H_A, as the "true state of nature." This implies the specification of all parameters in model 1, including the magnitude of the interaction to detect, β_{EG} (or equivalently θ^{tb}), as well as the odds ratios for the main effects for G, $\exp(\beta_E)$ (or $OR_{G=1|E=0}$), and for E, $\exp(\beta_E)$ (or $OR^{tb}_{E|G=0} = OR_{E=Q-1|G=0}$). In addition, β_0 must be specified, since model 1 is a prospective model of disease outcome. Note that for rare diseases, variations in β_0 have little effect on required sample size. For the examples in table 2, we set $\beta_0 = \ln(0.001)$, which corresponds to a disease rate for the baseline categories of $0.001 \cong 0.001/(1 + 0.001)$. Thus, the alternative hypothesis for line 1 of table 2 is H_A: $\beta_0 = -6.9$, $\beta_G = \ln(1.5)$, $\beta_E = \ln(1.5)/4$, $\beta_{EG} = \ln(1.5)/4$.

The next step is for the investigator to specify the null hypothesis, denoted H_0. The null hypothesis for the test of no multiplicative gene-environment interaction, i.e., the odds ratio for exposed susceptibles is the

TABLE 1. Odds ratios (OR) for the association between disease risk and a dichotomous genetic factor and an environmental exposure classified in quintiles, when non-susceptible unexposed subjects are the reference group. Odds ratios are based on model 1

	$E = 0$	$E = 1$	$E = 2$	$E = 3$	$E = 4$
$G = 0$	1.0	$OR_{E=1\|G=0}$	$OR_{E=2\|G=0} = (OR_{E=1\|G=0})^2$	$OR_{E=3\|G=0} = (OR_{E=1\|G=0})^3$	$OR^{tb}_{E\|G=0} = (OR_{E=1\|G=0})^4$
$G = 1$	$OR_{G=1\|E=0}$	$OR_{G=1\|E=0} \times OR_{E=1\|G=0} \times \theta$	$OR_{G=1\|E=0} \times OR_{E=2\|G=0} \times \theta^2$	$OR_{G=1\|E=0} \times OR_{E=3\|G=0} \times \theta^3$	$OR_{G=1\|E=0} \times OR^{tb}_{E\|G=0} \times \theta^4$

TABLE 2. Comparison between the formulae of Foppa and Spiegelman (6) and Lubin and Gail (7) to calculate the minimum sample size to achieve an 80% power to detect a particular top-to-bottom quartile interaction effect (θ^{tb}), for a particular exposure effect among non-susceptibles ($OR^{tb}_{E|G=0}$), with genetic effect $OR_{G=1|E=0}$ = 1.5 and $P(G)$ = 50%*

θ^{tb}	$OR^{tb}_{E\|G=0}$	Total sample size		% Difference
		Foppa and Spiegelman	Lubin and Gail	
1.50	1.50	6,386	6,580	2.9
3.00	1.50	906	1,020	11.2
6.00	1.50	366	472	22.5
1.50	3.00	6,858	7,172	4.4
3.00	3.00	986	1,158	14.9
6.00	3.00	404	561	28.0
1.50	6.00	7,798	8,267	5.7
3.00	6.00	1,134	1,385	18.1
6.00	6.00	470	696	32.5

* Other parameters were fixed at the values specified in the text.

product of the individual factor-specific odds ratios, is specified by H_0: $\beta_{EG} = 0$, or equivalently H_0: $\theta^{tb} = 1$. The null hypothesis of the test for no interaction in model 1 does not specify values of the main effect parameters β_G and β_E. H_0 is an example of a composite null hypothesis (8). The alternative hypothesis arises from a 4-dimensional parameter space defined by (β_0, β_G, β_E, β_{EG}), while the null hypothesis arises from a 3-dimensional parameter subspace defined by (β_0, β_G, β_E, $\beta_{EG} = 0$). The null hypothesis only specifies $\beta_{EG} = 0$; however, sample size and power formulae depend on the covariance matrix under H_0, which is a function of β_0, β_G, and β_E.

The difference between the LG and the FS approaches arises from the specification of β_G and β_E under the null hypothesis. The LG formulae use the maximum likelihood estimates (MLE's) of β_G and β_E when $\beta_{EG} = 0$ and the alternative is true, while the FS formulae use the values for β_G and β_E specified in the alternative hypothesis. The need to estimate the parameters β_G and β_E when $\beta_{EG} = 0$ can be explained by thinking about the definition of power. Power is the probability of rejecting H_0, *given that* H_A *is true*. The test statistic involves the covariance matrix for the parameters which is a function of β_G and β_E. Thus, to obtain the correct sample size and power, the procedure for testing $\beta_{EG} = 0$ must use those values of β_G and β_E that are most likely to be observed *when* H_A *is true*.

The MLE's of β_G and β_E when $\beta_{EG} = 0$ (i.e., under H_0) and the H_A is true will not generally be equal to their values under H_A. In practical terms, if we fit model 1 with $\beta_{EG} = 0$ using data generated under H_A, the MLE of $\exp(\beta_G)$ will be an "average" of the category-specific odds ratios for the genetic effect, i.e., ($OR_{G|E=0}$), ($OR_{G|E=0} \times \theta^1$),...,($OR_{G|E=0} \times \theta^{Q-1}$) (see table 1). Similarly, the MLE of $\exp(\beta_E)$ will be an "average" of the category-specific odds ratios for the exposure effect among susceptibles and non-susceptibles. Table 3 shows the values of the odds ratios for the genetic and environmental factors under H_0 when H_A is true. The larger the interaction, the greater the difference in the odds ratios for the main effects under the null, the greater the misspecification of the covariance matrix under H_0 in the FS approach, and thus the greater the difference in the sample sizes between the LG approach and the FS approach.

Specifying that β_G and β_E are the same under the null and alternative hypotheses, as in the FS approach, leads to a restricted definition of the null hypothesis, namely, no gene-environment interaction *and* specific values for the main effects odds ratios, i.e., H_0: $\beta_{EG} = 0$, $\beta_G = \beta_G^*$, and $\beta_E = \beta_E^*$, where "*" denotes specific values, which in the FS formulation equals the values under the H_A. The null is more specific and thus fewer subjects are needed to reject H_0.

THE HWANG ET AL. APPROACH

The approach of Hwang et al. (4) for sample size and power for assessing gene-environment interaction was defined only for a binary genetic susceptibility factor and a binary exposure factor, and suffers from the same limitation as the FS approach. As previously indicated, Hwang et al. present sample size estimates corresponding to a model of interaction where the genetic factor does not have an effect on disease risk in the absence of the environmental factor. This assumption adds an additional condition to the alternative hypothesis, namely, $OR_{G=0|E=0} = OR_{G=1|E=0} = 1$ or in the

TABLE 3. Maximum likelihood estimates (MLE) of $OR_{G=1|E=0}$ and $OR^{tb}_{E|G=0}$ under the null hypothesis of $\theta^{tb} = 1.0$ for different alternative hypotheses

Alternative hypotheses			Under the null hypothesis ($\theta^{tb} = 1.0$)	
θ^{tb}	$OR^{tb}_{E\|G=0}$	$OR_{G=1\|E=0}$	$OR^{tb}_{E\|G=0}$	$OR_{G=1\|E=0}$
1.50	1.50	1.50	1.90	1.87
3.00	1.50	1.50	2.99	2.78
6.00	1.50	1.50	5.09	4.29
1.50	3.00	1.50	3.80	1.90
3.00	3.00	1.50	6.02	2.49
6.00	3.00	1.50	10.38	4.75
1.50	6.00	1.50	7.61	1.94
3.00	6.00	1.50	12.14	3.10
6.00	6.00	1.50	21.15	5.23

logistic formulation $\beta_G = 0$. Model 1 indicates that the magnitude of the interaction is determined by the comparison of the logit in exposed susceptibles with exposed non-susceptibles. Thus, the estimation of the effect of exposure in the non-susceptible population is not informative for the estimation of β_{EG}. The test of the null hypothesis $\beta_{EG} = 0$ is equivalent to the test of the null hypothesis of no genetic effect among the exposed. To test for this pattern of gene-environment interaction (pattern 2 in Khoury et al. (2) and pattern *E* in Ottman et al. (3)), subjects not exposed to the environmental factor are not required and the sample size formula for a single binary variable is applicable. We should emphasize that the estimation of the odds ratio for exposure in non-susceptibles will often be of interest; however, it does not provide information on the nature of the interaction, given this particular alternative hypothesis.

CONCLUSION

The approaches used by Hwang et al. (4) and Foppa and Spiegelman (6) result in an underestimation of sample size for the test of a gene-environment interaction. The results from these approaches will approximate the correct power and sample size estimates only when the gene-environment interaction specified under the alternative hypothesis is small, or when the odds ratios for the genetic and exposure effects are small. Otherwise, the approaches of Hwang et al. and Foppa and Spiegelman can lead to a substantial underestimation of sample size and overestimation of power.

The approaches of Hwang et al. (4) and Foppa and Spiegelman (6) can only be used to calculate power and sample size needed to detect an interaction when the null joint effect of the environmental and genetic factors is multiplicative. However, other types of statistical interactions such as when the null joint effect is additive might also be of interest (9). Calculations for sample size and power to detect either additive or multiplicative interactions between genetic and environmental risk factors (continuous or categorical) can be performed using the general formulae developed in Lubin and Gail (5). These calculations can be carried out using POWER included in the computer program EPITOME (National Cancer Institute, Bethesda, Maryland). A new version of this program tailored to perform calculations of interest in studies of gene-environment interactions is being developed in the Division of Cancer Epidemiology and Genetics of the National Cancer Institute, and will be available free of charge in the near future. A copy of the program can be obtained by e-mail by sending a message to brownh@exchange.nih.gov, or by mail by sending a diskette to the corresponding author, Dr. García-Closas.

ACKNOWLEDGMENTS

The authors thank Dr. Nathaniel Rothman for his helpful comments on the manuscript.

REFERENCES

1. Yang Q, Khoury MJ. Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research. Epidemiol Rev 1997;19:33–43.
2. Khoury MJ, Beaty TH, Cohen BH. Fundamentals of genetic epidemiology. New York: Oxford University Press, 1993.
3. Ottman R, Pike MC, King M-C, et al. Familial breast cancer in a population-based series. Am J Epidemiol 1986;123:15–21.
4. Hwang S-J, Beaty T, Liang K-Y, et al. Minimum sample size estimation to detect gene-environment interaction in case-control designs. Am J Epidemiol 1994;140:1029–37.
5. Smith PG, Day NE. The design of case-control studies: the influence of confounding and interaction effects. Int J Epidemiol 1984;13:356–65.
6. 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.
7. Lubin JH, Gail MH. On power and sample size for studying features of the relative odds of disease. Am J Epidemiol 1990;131:552–66.
8. Cox DR, Hinkley DV. Theoretical statistics. London: Chapman and Hall, 1974.
9. Greenland S, Rothman KJ. Concepts of interaction. In: Rothman KJ, ed. Modern epidemiology. Philadelphia, PA: Lippincott-Raven, 1998:329–42.