Power and sample size for case-control studies of gene-environment interactions: a new method with comparison to the old

Donna Spiegelman[1,2] and Roger Logan[2]

January 7, 2002

Abstract

A new method for power and sample size calculations for studies of gene-environment interactions of a binary genotype and ordinal exposure is proposed, and compared to previous methods, including those of Foppa and Spiegelman (1997), Lubin and Gail (1990) and Greenland (1983). These methods differ in the assumptions that are made about the values of the main effects of exposure and genotype under the null and alternative hypotheses. In the new method, the null values are set to the values obtained or expected from data not mutually adjusted for gene and environmental effects, and the alternative values of the parameters are solved for as a function of the other design parameters specified. This procedure for fixing assumptions about these nuisance parameters most accurately utilizes the information available at the planning stage of such studies. In addition, the new method gives smaller sample sizes and higher power in some realistic examples. A fully-documented, user-friendly

[1]Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston MA. stdls@channing.harvard.edu

[2]Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue,Boston MA.

Acknowledgements: This work was supported, in part, by grants NIH P01 CA55075 and P30 CA06516

program implementing the new method can be downloaded from the first author's web-site, `http://www.hsph.harvard.edu/faculty/spiegelman/ge_trend.html`.

1 Introduction

With the mapping of the human genome, the ability to study the possible interaction of genetic and environmental risk factors for disease becomes more important. A recent search of the Pub Med web site, at the National Library of Medicine for the terms "(gene AND environment AND interaction) OR "gene-environment interaction OR gene-gene interaction" published between 1995–2001 listed over 1532 articles. When the gene-environment interaction is the parameter of interest, power and sample size calculations based on main effects only can lead to overestimation of the power and underestimation of the required sample size.

A recent paper by García-Closas & Lubin (1) compare the power and sample size calculations for case-control studies of gene-environment interactions with a binary genotype and an ordinal exposure given by two different methods. In this paper, they showed by several examples that the method of Foppa & Spiegelman (2) underestimated the sample size for detecting a gene-environment interaction. In this paper, we wish to investigate the extent to which this phenomena is true more generally. In fact, we will show that there are regions of the design space that quite reasonably could occur in practice in which García-Closas & Lubin's claim is reversed, and the alternative method of Lubin & Gail (3) considered by García-Closas & Lubin gives a smaller sample size. In addition, we present a new method that eases some of the restrictions that were placed on the Foppa & Spiegelman approach for calculating power and sample size, and more appropriately utilizes the information available when such calculations are made than perhaps has been done by previously proposed methods.

In section 2, we will review the approaches used by Lubin & Gail and Foppa & Spiegelman for power and sample size calculations. We will also present the new method for these same calculations.

In section 3, we will present a comparison of the various methods similar to that given in García-Closas & Lubin. In section 4, a summary and recommendations are given.

2 Methods

2.1 Notation and assumptions

Assume D and G are binary indicators of disease status and the genotype. As in Foppa & Spiegelman (2), let E denote ordinal exposure levels $0, 1, \ldots, q-1$, for some $q \geq 2$, where $q-1$ is the highest exposure level. We assume the log odds ratio of disease to be a linear function of the ordinal scores for exposure. That is, the log odds ratio comparing adjacent exposure levels is assumed to be constant for those with the genotype, $OR(E|G=1)$ and for those without the genotype $OR(E|G=0)$. Denote the top-to-bottom quantile exposure effect in those without the genotype as $\mathrm{OR}^{tb}(\mathrm{E}|\mathrm{G}{=}0) = \mathrm{OR}(\mathrm{E}|\mathrm{G}{=}0)^{q-1}$. Let θ denote the gene-environment interaction effect, where $\theta = \dfrac{OR(E|G=1)}{OR(E|G=0)}$, and let the top-to-bottom odds ratio for the gene-environment interaction to be $\theta^{tb} = \theta^{q-1}$. These definitions imply the prospective logistic model for the case-control study

$$\mathrm{logit}\left[\mathrm{P}(\mathrm{D}=1|\mathrm{E},\mathrm{G})\right] = \beta_0 + \beta_{\mathrm{g}}\mathrm{G} + \beta_{\mathrm{e}}\mathrm{E} + \beta_{\mathrm{eg}}\mathrm{EG}, \tag{1}$$

where $\beta_g = \log\left[OR(G|E=0)\right]$, $\beta_e = \log\left[OR(E|G=0)\right]$, and $\beta_{eg} = \log(\theta)$.

We are interested in finding the minimum sample size, N, in a case-control study for which the hypothesis $H_0 : \beta_{eg} = 0$ versus the test of $H_A : \beta_{eg} \neq 0$ has power ψ, using the power function $\psi(N, \beta_{eg}^A)$, where β_0, β_g and β_e are nuisance parameters and β_v^u is the value of β_v under H_u. The four methods that we consider in this paper differ in the manner in which they handle the nuisance parameters in calculating the power functions.

2.2 Foppa and Spiegelman

Foppa & Spiegelman (2), hereafter denoted FS, assumed that the power function $\psi(N, \beta_{eg}^A)$ has the following form

$$\psi_{FS}(N, \beta_{eg}^A) = 1 - \Phi\left(\frac{z_{\alpha/2}\sigma_{GE}^0 - \beta_{EG}^A}{\sigma_{GE}^A}\right), \tag{2}$$

where Φ is the cumulative standard normal function, $z_{\alpha/2}$ is the value of a standard normal variable, Z, such that $P(Z \geq z) = 1 - \alpha/2$, α is the size of the test, and σ_{GE}^0 and σ_{GE}^A are the asymptotic standard errors of the estimates of β_{eg} under H_0 and H_A, respectively. For information on how the standard errors are calculated, we refer the reader to the Appendix of Foppa and Spiegelman (2). In calculating the standard errors, Foppa and Spiegelman require that the analyst provides the prevalence of the variant genotype in the study base which produced the cases, which can be estimated from the controls if the disease is rare or if incidence density sampling (4, p. 94, 99) was used to obtain the controls. These authors assume that the nuisance parameters β_e and β_g have the same values under the null and alternative hypotheses and are known or have been obtained from a prior study. The null and alternative values of β_0 which preserves the case-control matching ratio, C, is then solved for. In particular, as given by FS, the intercept term, β_0, is the solution of

$$e^{\beta_0} = \frac{1}{C}\left[\frac{1}{\sum_{e,g} e^{\beta_g g + \beta_e e + \beta_{eg} eg} P(E = e, G = g | d = 0)}\right], \tag{3}$$

where $\beta_{eg} = 0$ under the null hypothesis, and $P(E = e,\ G = g \mid d = 0)$ is the joint prevalence of e and g in the population from which the cases will arise.

To summarize, the required inputs for implementing the FS method are:

1. The distribution of gene prevalence and exposure in the population from which the cases will arise. Under the common assumption that the distribution of genotype and exposure are independent in the population from which the cases will arise, when the exposure is a quantile, then only then number of quantiles and the marginal prevalance of the genotype is needed.

2. The control-case ratio, C.

3. The ratio of top-to-bottom quantile odds ratios, θ^{tb}, for gene-environment interaction under the alternative hypothesis.

4. The assumed odds ratio for genotype in the lowest exposure level, $OR(G|E=0)$.

5. The top-to-bottom quantile odds ratio for exposure, $OR^{tb}(E|G=0)$ in the reference genotype level.

Greenland (6) and Smith and Day (7) discussed a variant of the FS method, which we will call the Wald method, where the null variance term σ^0_{GE} is replaced by the alternative variance σ^A_{GE}, giving the power function

$$\psi_W(N, \beta^A_{eg}) = 1 - \Phi\left(\frac{z_{\alpha/2}\sigma^A_{GE} - \beta^A_{EG}}{\sigma^A_{EG}}\right) = 1 - \Phi\left(z_{\alpha/2} - \frac{\beta^A_{EG}}{\sigma^A_{EG}}\right). \tag{4}$$

The Wald method requires similar inputs as the FS method. However, the top-to- bottom quantile odds ratio for exposure, $OR(E|G=0)$, is taken to be the value assumed under the alternative, H_A, and no assumption is needed at all about the value of this parameter under H_0. Similarly, the value input for $OR(E|G=0)$ is taken to be the value assumed under H_A, with no assumption needed for the value of this parameter under H_0. Since the null variance of $\hat{\theta^{tb}}$ is likely to be smaller than the variance under the alternative, we might generally expect for the FS method to give smaller sample sizes for the same inputs.

Although the software implementation of the FS method assumes that $P(E,G|D=0) = P(E|D=0)P(G|D=0)$, in the appendix of FS the condition of independence of the gene prevalence and exposure is not required. In practice, rarely will there be good prior knowledge of the joint distribution of E and G, so this additional generalization has little practical importance.

2.3 Lubin and Gail

Lubin and Gail (3), hereafter denoted LG, base their power function on the score statistic. Since the score statistic is asymptotically equivalent to the Wald statistic (9), given sufficient sample size this in itself should not be an important difference between the FS and LG methods. All other things being equal, this could however account for small differences between the two approaches. Then,

$$\psi_L \ \ (N, \beta_{eg}^A) = 1 - \Phi\left(\frac{z_{\alpha/2}\sigma_U^0 - E_{H_A}(U)}{\sigma_U^A}\right), \tag{5}$$

where U is the score function of the log-likelihood based upon model (1), that is, the vector of first derivatives of the log-likelihood of the case-control study with respect to the parameters $(\beta_0, \beta_e, \beta_g, \beta_{eg})$, $E_{H_A}(X)$ is the expectation of some statistic, say X, under the assumption that the data were generated under the alternative. Lubin and Gail require that the values of β_0, β_e and β_g be provided under the alternative. Then, β_0^A is obtained by solving an equation similar to (3).

The values of β_0, β_e and β_g under the null hypothesis are determined by solving the equation $E_{H_A}(U) = 0$ for β_0, β_g, and β_e when β_{eg} is set to 0. The expected values of the $2 \times 2 \times q$ table under H_A are calculated, and the maximum likelihood estimates that would be obtained if the mis-specified model assuming $\beta_{eg} = 0$ were fit to these pseudo-data are then computed. LG also request from the user the baseline incidence rate for the outcome in the study base, which is often unknown, and, if known, typically varies over time and with age. Under the rare disease assumption, or under incidence-density sampling of controls (4,p. 94, 99), the baseline incidence rate disappears from the expression used to calculate the exposure distribution among cases and controls (approximately so, under the rare disease assumption).

To summarize, the input requirements for use of the LG method are the same as items 1-3 of the FS method, but items 4 and 5 are a bit different:

4. The assumed odds ratio for genotype in the lowest exposure level, $OR(G|E = 0)$, under H_A.

5. The top-to-bottom quantile odds ratio for exposure, $OR^{tb}(E|G=0)$, in the reference genotype level, under H_A.

2.4 Spiegelman and Logan

We propose a new method for power and sample size calculations, denoted SL, that follows the same general approach as the FS method, except we alter how the assumptions about β_e and β_g under H_A are framed. Before conducting the proposed study where cases and controls are to be genotyped, information about values of the marginal odds ratios OR(E) and OR(G) are available. Under the null and when E and G are independent or approximately so, these marginal odds ratios are approximately equal to the null values, i.e. under H_0, when $Corr(E,G) \approx 0$, $OR(E) \approx OR(E|G=0)$ and $OR(G) \approx OR(G|E=0)$. These relations do not hold under the alternative. That is, under H_A, OR(E)$\not\approx$OR(E| G=0) and OR(G)$\not\approx$OR(G| E=0).

This new method exploits the information typically available before a case-control study of a gene-environment interaction is conducted in a manner which we believe to be most appropriate. Usually, values for these marginal quantities have been published previously and, in addition, pilot data may be available for OR(E). Much less is known about the corresponding values that these parameters might take on under the alternative, even when $P(G,E) \approx P(G)P(E)$. Hence, it makes sense to specify a reasonable range of values that these parameters are likely to fall within, which could but not necessarily overlap the published marginal values.

With these considerations in mind, the input requirements for use of the SL method are the same as those for the FS method, except that items 4 and 5 of the FS method are replaced with new items 4-6 given below:

4. The odds ratio for genotype in the lowest exposure level, $OR(G|E=0)$, assumed under H_0.

5. An interval for $OR^{tb}(E|G=0)$ in which it is assumed that this parameter will lie under H_A.

6. An interval for $OR(G|E=0)$ in which it is assumed that this parameter will lie under H_A.

The SL method uses the power function given in equation (2). From equation (3), it can be shown that β_0 is a function of β_e, β_g, so that for fixed β_{eg}, equation (2) can be assumed to be a function of N, β_e and β_g. For a fixed sample size, N, we calculate the power by first finding the smallest value of equation (2) when β_e and β_g are restricted to be in the intervals provided by requirements 5 and 6. That is, equation (2) is minimized with respect to β_e and β_g, within the rectangular region formed by the Cartesian product of the intervals specified by the user under H_A. This is a standard non-linear optimization problem, and the publicly available DMNGB subroutine, obtained from `http://www.netlib.org`, which applies a quasi-Newton method (for example, see 5) was used to solve it. In addition to the smallest possible power, the program also calculates the largest possible power for the given sample size over the allowable intervals for β_g and β_e.

To find the smallest sample size so that the power of testing the above hypothesis is at least the desired power for any pair of β_g and β_e in the provided intervals, we use the derivation of the asymptotic standard errors of $(\beta_0, \beta_g, \beta_e, \beta_{eg})$ from the appendix of Foppa & Spiegelman (2), and note that the expected cell counts under the null and alternative hypothesis can be expressed as a multiple of the number of cases, m_1. Hence, we can rewrite the standard errors in Foppa and Spiegelman's equation (2) in terms of a new standard error and the square root of the number of cases, i.e. under H_0, $\sigma^0_{GE} = \tilde{\sigma}^0_{GE}/\sqrt{m_1}$, where $\tilde{\sigma}^0_{GE}$ is the (4,4)-entry of the inverse of the expected information matrix of the log-likelihood of the data. Using the same notation as in the appendix of Foppa & Spiegelman (2), the the (u,v)-entry of the expected information matrix is

$$E\left(\left.\frac{\partial^2 \ell(\beta|\mathbf{X})}{\partial\beta_u \partial\beta_v}\right|_{H_0}\right) = -\sum_{j=1}^{2\times Q} z_{ju} z_{jv} \times [\mathrm{E}(m_{0,j}) + \mathrm{E}(m_{1,j}|H_0)] \times \frac{\exp(\beta^T_{H_0}\mathbf{z}_j)}{\left[1+\exp(\beta^T_{H_0}\mathbf{z}_j)\right]^2},$$

where we now have simplified $[\mathrm{E}(m_{0,j}) + \mathrm{E}(m_{1,j}|H_0)]$ to

$$\mathrm{E}(m_{0,j}) + \mathrm{E}(m_{1,j}|H_0) = m_1 \left[\frac{\exp(\beta_{H_0}^T \mathbf{z}_j) \times \Pr(E = z_{j2}, G = z_{j3}|D = 0)}{\sum_{k=1}^{2\times Q} \exp(\beta_{H_0}^T \mathbf{z}_k) \times \Pr(E = z_{k2}, G = z_{k3}|D = 0)} + \Pr(E = z_{j3}|D = 0) \times \Pr(G = z_{j2}|D = 0) \times C \right].$$

We then solve equation (2) for m_1 to obtain

$$m_1 = \left(\frac{z_{\alpha/2}\tilde{\sigma}_{GE}^0 - z_{1-\psi}\tilde{\sigma}_{GE}^A}{\beta_{eg}^A} \right)^2, \tag{6}$$

where $z_{1-\psi}$ is the value of a standard normal variable, z, such that $P(Z > z) = 1 - (1 - \psi) = \psi$. To find the minimum sample size required to achieve a specified power, ψ, we then maximize the right hand side of (6) with respect to β_g and β_e, using the quasi-Newton method (ref 5) implemented by the subroutine DMNGB publicly available from from `http://www.netlib.org`. A sample of size $N = m_1(1 + C)$ then gives power of at least ψ for any value of β_g and β_e in the specified intervals.

Given $OR(G)$, $OR(E)$, ROR^{tb}, $P(G)$, and $P(E)$, under the assumption of independence of G and E, and all levels of $P(G, E)$ otherwise, we can solve equations (A.1) and (A.2) of the Appendix explicitly for $OR(G|E = 0)$ and $OR(E|G = 0)$ under the alternative. Further details on how this is done are given in the Appendix. Tables 1 and 2 gives these values when $q = 2$ for $P(E) = 0.05$, $P(G) = 0.25$ and $P(E = 1|G = 1)/P(E = 1|G = 0) = 1$ and 1.5, indicating no and a modest positive association between G and E, respectively. Recall that, as always, even when G is not an effect modifier of the association of E with the outcome, it will be a confounder as long as G and E are associated in the controls. This is why the marginal odds ratios are not equal to the conditional odds ratios even when there is no interaction between E and G, when E and G are associated. With P(E)=0.05, P(G)=0.5, no association of G with E, an ROR^{tb} of 4 and marginal odds ratios for genotype and exposure of 2 and 1.5, respectively, the conditional values of the two odds ratio under H_A are 1.8 and 0.7 (Table 1). When P(E) increases to 0.5 and P(G) to 0.5, the two conditional

odds ratios under H_A for genotype and exposure, respectively, are 0.61 and 0.93. Depending on a complicated, non-linear relationship between all these parameters, when there is a gene-environment interaction, the conditional odds ratios for E and G can be in the opposite direction of the marginal ones. Tables 1 and 2 demonstrate how difficult it is to 'guess' what the conditional odds ratios under the alternative might be, given the information about the marginal odds ratios, which is typically all that is available at the planning phase of such a study. Our new computer program allows the user to explore the likely values of $OR^{tb}(E|G = 0)$ and $OR(G|E = 0)$ under H_A, so a realistic range of these can be specified as inputs for the power and sample size calculations to be conducted subsequently.

3 Comparison

In this section, we compare the methods used for calculating the required sample size needed for testing the null hypothesis of no gene-environment interaction. As can be seen in table 3, the required inputs for the various methods are similar except for the required information about $OR(G|E = 0)$ and $OR(E|G = 0)$. In the FS and SL methods, the values of these parameters under H_0 are needed, while for the LG method, the values under H_A are needed. In addition to the values under the null, SL requires the user to provide a range of possible values of $OR^{tb}(E|G = 0)$ and $OR(G|E = 0)$ to be used under H_A. In the FS method, the null values are assumed equal to their corresponding values under the alternative. Because in binomial data the variance of the model parameters depend on their assumed values, these distinctions have some importance. In contrast, when power or sample size for a test for gene-environment interaction for a continuous outcome in a linear model is considered, as in the recent paper by Luan et al. (8) this issue is not relevant.

For all the required sample size comparisons presented in this paper, we fixed the desired power at 80%, α at 0.05, and assumed that G and E were independent. First, we expand upon table 2 of García–Closas & Lubin. In that comparison, $OR(G|E = 0) = 1.5$, $C = 1.0$, $P(G = 1|D = 0) = 0.5$,

Table 3: Required inputs for the various methods for calculating sample size. Y=required, N=not required, H_0 = used under null hypothesis, and H_A= used under alternative hypothesis.

	Methods			
Required Input	FS	Wald	LG	SL
Desired power, ψ	Y	Y	Y	Y
Type I error, α	Y	Y	Y	Y
Control-case ratio, C	Y	Y	Y	Y
Number of quantiles for E, q	Y	Y	Y	Y
Genotype prevalence	Y	Y	Y	Y
θ^{tb}	Y	Y	Y	Y
$OR(G\|E=0)$	H_0=H_A	H_A	H_A	H_0, range under H_A
$OR^{tb}(E\|G=0)$	H_0=H_A	H_A	H_A	H_0, range under H_A

$q = 5$, and G is assumed independent of E in the controls. We add the Wald and SL methods to this table, and for the SL method we used a suitable range of values for $OR(G|E = 0)$ and $OR^{tb}(E|G = 0)$ under the alternative H_A, as suggested by methods given in the previous section. The results for this comparison are presented in table 4. In these examples, the new method gave sample sizes smaller than those given by all the other methods. This appears to be because the range of values assumed under H_A for $OR(G|E = 0)$ and $OR^{tb}(E|G = 0)$ often seem to produce in expectation a set of 2×q tables which are more balanced and thus yield parameter estimates with smaller variances. For example, consider the last row of Table 4, where $\theta^{tb} = 6$. In this case, under H_0, $OR(G|E = 0)$ was 1.5 under SL and 5.2 under LG, and $OR^{tb}(E|G = 0)$ was 6 under SL and 18.4 under LG. Clearly, these extreme odds ratios will lead to highly unbalanced tables in expectation under H_0 and a large null variance for the model parameters. Under, H_A, the situation is somewhat similar. The alternative value for $OR(G|E = 0)$ is 1.5 for LG and 0.41 for SL, and the alternative value for $OR^{tb}(E|G = 0)$ is 6 for LG and 3.3 for SL. Inspection of the expected cell counts for these four sets of tables (not shown) confirms these observations, and it is clear that, at least in this example, the tables implied by the LG method are more unbalanced than those implied by the SL method.

The Wald method gave results very similar to the LG method.

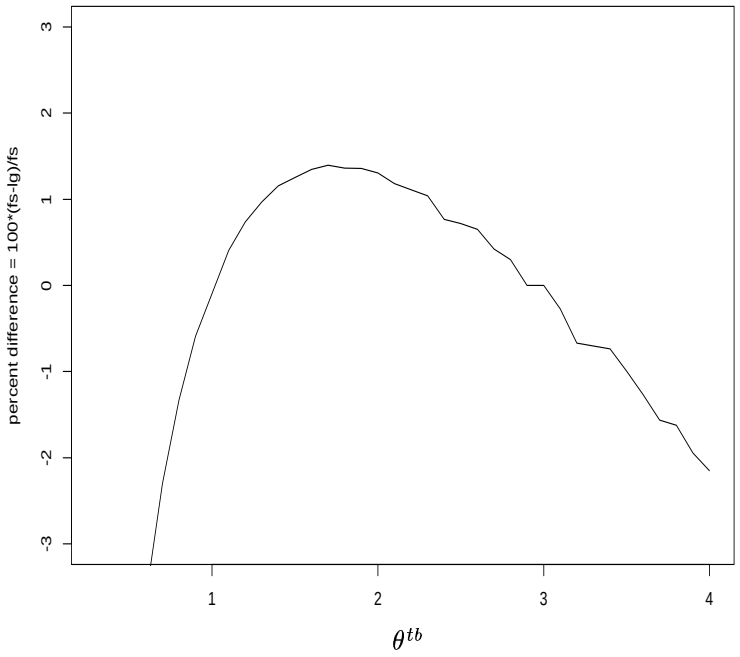
Table 4: Comparison of sample size calculations using the methods of Foppa & Spiegelman (FS), Wald, Spiegelman & Logan (SL), and Lubin & Gail (LG) to achieve 80% power when OR(G|E = 0) = 1.5, $C = 1$, and $P(G = 1) = 0.5$.

θ^{tb}	OR^{tb}(E|G=0)	FS	Wald	LG		SL	
						$OR^{tb,A}(E\|G=0)$*	$OR^{A}(G\|E=0)$*
1.5	1.5	6386	6580	6580	6395	1.2(1.0,1.5)	1.2(1.0,1.5)
3.0	1.5	906	1020	1020	885	0.89(0.75,1.2)	0.82(0.7,1.1)
6.0	1.5	366	472	472	345	0.72(0.5, 1.0)	0.54(0.25,0.75)
1.5	3.0	6858	7162	7172	6835	2.4(1.5,2.75)	1.2(0.8,1.5)
3.0	3.0	986	1152	1158	945	1.8(1.25, 2.25)	0.75(0.5,1.0)
6.0	3.0	404	554	561	365	1.6(1.0,2.0)	0.46(0.25,0.75)
1.5	6.0	7798	8248	8267	7755	4.8(3.5,5.5)	1.1(0.75, 1.5)
3.0	6.0	1134	1374	1385	1075	3.7(3.0,4.5)	0.69(0.4, 0.9)
6.0	6.0	470	684	696	415	3.3(2.5,4.0)	0.41(0.2,0.6)

* Value of parameter fixed under the assumption of independence of E and G, and interval given to produce sample sizes shown for the SL method.

In a second comparison, we explore sample sizes given by the FS and LG methods over a range of values for all the required parameters for these methods to investigate more generally the assertion made by García-Clossas and Lubin that the FS method leads to an underestimation of the sample size, unless the gene-environment interaction is small or if the odds ratio for the genetic and exposure effects are small (1, p. 692) . We compared the relative difference in the sample size obtained by inverting the power functions given by equations (2) and (5) over a grid which allowed the exposure to have between two and five quantiles, $P(G = 1) = 0.25$, $OR(G|E = 0) = 2.0$, and $C = 2.0$. The odds ratio for $OR^{tb}(E|G = 0)$ took on values 2.0, 2.5, and 3.0, while θ^{tb} took on values between 0.1 and 4.0 with increments of size 0.1 (removing the case where $\theta^{tb} = 1.0$). The differences between the two methods generally followed the conclusions of García-Clossas and Lubin, but not consistently (figure 1). We have graphed the relative difference between the FS and LG methods for the same parameters when θ^{tb} varies between 0.1 and 4.0. As can be seen, for values of θ^{tb} between 1 and 3 the FS method gives a greater required sample size, in contradiction to García-Closas and Lubin's claim,

Figure 1: Relative differences in sample size given by the FS and LG methods verses θ^{tb} when $OR(E|G=0) = 2.0$, $OR(G|E=0) = 1.5$, $q = 5$, $C = 2$, and $P(G=1) = 0.25$.



and for values of θ^{tb} outside of this range, the LG gives a smaller required sample size, consistent with García-Closas and Lubin's claim. However, in these cases the difference in not large, either way.

4 Discussion

The new method (SL) resembles most closely the situation encountered in practice at the planning stage of a study of a gene-environment interaction. The investigator is most likely to have greatest knowledge about the null values to be assumed for $OR^{tb}(E|G)$ and $OR(G|E)$, and least likely to have knowledge about the conditional values of these parameters under the alternative. In realistic examples, it appears that this new method gives sample sizes smaller than all of the other methods

considered. Given the importance of these power and sample size calculations in current epidemiologic research, this new method is likely to be widely used. There is, of course, no limit of this methodology to studies of gene-environment interactions. Studies of interaction between any binary and ordinal variable can be planned using the methodology presented in this paper.

A fully documented, user-friendly program implementing the new method can be downloaded from the first author's web-site `http://www.hsph.harvard.edu/faculty/spiegelman/ge_trend.html`.

5 References

1. García–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;149:689-692.

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. Lubin J, Gail M. On power and sample size for studying features of the relative odds of disease. *Am J Epidemiol* 1990;131:552-66

4. Rothman K, Greenland G. Modern Epidemiology, Second Edition. Philadelphia: Lippincott-Raven, 1998.

5. Dennis, JE, Schnabel RB. Numerical methods for unconstrained optimization and nonlinear equations. Philadelphia, PA: SIAM, 1996.

6. Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. *Stat Med* 1983;2:243-51.

7. Smith PG, Day NE. The design of case-control studies: the influence of confounding and interaction effects. *Int J Epidemiol* 1984;13:356-365.

8. Luan JA, Wong MY, Day NE, Wareham NJ. Sample size determination for studies of gene-environment interaction. *Int J Epidemiol* 2001; 30:1035-1040.

9. Cox D.R., Hinkley D.V. Theoretical statistics. London: Chapman Hall Ltd. 1974.

10. Mor J. J., Sorensen D. C., Hillstrom K. E., and Garbow B. S.. The MINPACK Project, in Sources and Development of Mathematical Software, W. R. Cowell, ed. Englewood Cliffs, NJ: Prentice-Hall, 1984.

11. Powell M.J.D. "A Hybrid Method for Nonlinear Equations" (Chap 6, p 87-114) in Numerical Methods for Nonlinear Algebraic Equations, P. Rabinowitz, editor. New York, New York: Gordon and Breach, 1970.

6 Appendix: Calculation of $OR^{tb}(E|G=0)$ and $OR(G|E=0)$ under H_A

Suppose that disease incidence in the study base which produced the cases follows the proportional hazards model given by $P(D=1|E,G,t,T\geq t)=h_0(t)e^{\beta_e E+\beta_g \ +\beta_{eg} E G}$, where $h_0(t)$ is the baseline incidence rate, and that the probability of being selected as a control at the time, t, that a case occurs is given by $f(t)$, then the expected number of cases and controls with exposure level E and genotype G are given by $h(t)e^{\beta_e E+\beta_g \ +\beta_{eg} E G)}P(E,G)$ and $f(t)\,P(E,G)$, respectively. To find the marginal odds ratio for E, we need to collapse the cells over the two levels of G, from the $2\times 2\times q$ given in table 5, for the case when $q=2$, to obtain the 2×2 table given in table 6.

Then the expected cell counts in the collapsed table are proportional to a fixed constant times

Table 5: Layout for a Case-Control study of a gene-environment interaction with a binary exposure ($q = 2$)

$G = 0$			$G = 1$		
	E			E	
D	1	0	D	1	0
1	a_0	b_0	1	a_1	b_1
0	c_0	d_0	0	c_1	d_1

Table 6: 2×2 table for E and D when $q = 2$, collapsed over levels of G.

	E	
D	1	0
1	a_e	b_e
0	c_e	d_e

the following expressions:

$$a_e = \sum_{g=0}^{1} P(D=1|E=1,G=g)P(E=1,G=g) = h_0(t)e^{\beta_e}\left[P(E=1,G=0) + e^{\beta_g+\beta_{eg}}P(E=1,G=0)\right],$$

$$b_e = \sum_{g=0}^{1} P(D=1|E=0,G=g)P(E=0,G=g) = h_0(t)\left[P(E=0,G=0) + e^{\beta_g}P(E=0,G=1)\right],$$

$$c_e = \sum_{g=0}^{1} P(D=0|E=1,G=g)P(E=1,G=g) = f(t)\left[P(E=1,G=0) + P(E=1,G=1)\right],$$

$$d_e = \sum_{g=0}^{1} P(D=0|E=0,G=g)P(E=0,G=g) = f(t)\left[P(E=0,G=0) + P(E=0,G=1)\right].$$

Then the marginal odds ratio for E, $OR(E) = (a_e/c_e)/(b_e/d_e)$, can be expressed as

$$OR(E) = e^{\beta_e}\frac{P(E=0)\left[P(E=1,\,G=0) + e^{\,\beta\,+\beta_{eg})}\,P(E=1,\,G=1)\right]}{P(E=1)\left[P(E=0,\,G=0) + e^{\beta_g}\,P(E=0,\,G=1)\right]} \qquad \text{(A.1)}$$

To calculate the marginal odds ratio for G, OR(G), we need to sum over the individual cells indexed by the q levels of E. Using the same notation as in the above 2×2 table with E replaced

by G, yields the following four expected cell counts:

$$a_g = \sum_{e=0}^{q-1} P(D=1|E=e,G=1)P(E=e,G=1) = h_0(t)e^{\beta_g}\sum_{e=0}^{q-1} e^{\beta_e\ e\beta_{eg}\ e}P(E=e,G=1),$$
$$b_g = \sum_{e=0}^{q-1} P(D=1|E=e,G=0)P(E=e,G=0) = h_0(t)\sum_{e=0}^{q-1} P(E=e,G=0) = h(t)P(G=0),$$
$$c_g = \sum_{e=0}^{q-1} P(D=0|E=e,G=1)P(E=e,G=1) = f(t)\sum_{e=0}^{q-1} P(E=e,G=1) = f(t)P(G=1),$$
$$d_g = \sum_{e=0}^{q-1} P(D=0|E=e,G=0)P(E=e,G=0) = f(t)\sum_{e=0}^{q-1} P(E=e,G=0) = f(t)P(G=0).$$

Thus, the marginal odds ratio for G is $OR(G) = (a_g/c_g)/(b_g/d_g)$ and can be simplified to

$$OR(G) = e^{\beta_g}\frac{P(G=0)\left[\sum_{e=0}^{q-1} e^{\beta_e\ e+\beta_{eg}\ e}\ P(E=e,G=1)\right]}{P(G=1)\ \left[\sum_{e=0}^{q-1} e^{\ \beta_e\ e)}\ P(E=e,\ G=0)\right]}. \tag{A.2}$$

If we assume that E and G are independent, so that $P(E=e,G=g) = P(E=e)*p_g^g*(1-pg)^{1-g}$, where $p_g = P(G=1)$, then equations (A.1) and (A.2) can be further simplified to

$$\begin{aligned} OR(E) &= e^{\beta_e}\frac{(1-p_g)+e^{\beta\ +\beta_{eg}}p_g}{(1-p_g)+e^{\beta_g}p_g}, \\ OR(G) &= e^{\beta_g}\frac{\sum_{e=0}^{q-1} e^{\beta_e\ e+\beta_{eg}\ e}\ P(E=e)}{\sum_{e=0}^{q-1} e^{\beta_e\ e}\ P(E=e)}. \end{aligned} \tag{A.3}$$

From these two equations, it is easy to see that under the null hypothesis when $\beta_{eg} = 0$, the marginal and conditional odds ratios are equivalent (the equations reduce to $OR(E) = e^{\beta_e}$ and $OR(G) = e^{\beta_g}$).

For $OR(E)$, $OR(G)$, and β_{eg} fixed, the system of nonlinear equations (A.1) and (A.2), or their simplifications under independence of E and G, equation (A.3), are solved using the MINPACK (10) subroutine HYBRD1 which implements a modification of the Powell hybrid algorithm (11), for $OR(E|G=0) = e^{\beta_e}$ and $OR(G|E=0) = e^{\beta_g}$ to obtain the results of tables 1 and 2.

Table 1: Values of OR^A(E| G=0) and OR^A(G| E=0) for different val es of OR(G), OR(E), and ROR^{tb} when P(E=1|G=1)/P(E=1|G=0)= 1.0, with P(E=1)=0.05 and P(G=1)=0.25

ROR^{tb}	OR(G)	OR(E)							
		0.50		0.75		1.50		2.00	
		$OR^A(E\|G=0)$	$OR^A(G\|E=0)$	$OR^A(E\|G=0)$	$OR^A(G\|E=0)$	$OR^A(E\|G=0)$	$OR^A(G\|E=0)$	$OR^A(E\|G=0)$	$OR^A(G\|E=0)$
1.0	0.50	0.50	0.50	0.75	0.50	1.50	0.50	2.00	0.50
	0.75	0.50	0.75	0.75	0.75	1.50	0.75	2.00	0.75
	1.50	0.50	1.50	0.75	1.50	1.50	1.50	2.00	1.50
	2.00	0.50	2.00	0.75	2.00	1.50	2.00	2.00	2.00
2.0	0.50	0.44	0.49	0.66	0.48	1.32	0.47	1.76	0.46
	0.75	0.42	0.73	0.63	0.73	1.26	0.71	1.68	0.69
	1.50	0.38	1.47	0.57	1.46	1.14	1.42	1.52	1.40
	2.00	0.36	1.96	0.54	1.95	1.08	1.90	1.45	1.87
4.0	0.50	0.35	0.47	0.54	0.46	1.09	0.43	1.47	0.41
	0.75	0.32	0.71	0.48	0.70	0.98	0.65	1.32	0.63
	1.50	0.25	1.44	0.38	1.42	0.78	1.34	1.05	1.30
	2.00	0.23	1.93	0.35	1.90	0.71	1.81	0.95	1.75

Table 2: Values of OR^A(E| G=0) and OR^A(G| E=0) for different val es of OR(G), OR(E), and ROR^{tb} when P(E=1|G=1)/P(E=1|G=0)= 1.5, with P(E=1)=0.05 and P(G=1)=0.25

ROR^{tb}	OR(G)	OR(E)							
		0.50		0.75		1.50		2.00	
		$OR^A(E\|G=0)$	$OR^A(G\|E=0)$	$OR^A(E\|G=0)$	$OR^A(G\|E=0)$	$OR^A(E\|G=0)$	$OR^A(G\|E=0)$	$OR^A(E\|G=0)$	$OR^A(G\|E=0)$
1.0	0.50	0.53	0.51	0.79	0.50	1.59	0.49	2.12	0.49
	0.75	0.51	0.76	0.77	0.75	1.54	0.74	2.06	0.73
	1.50	0.48	1.52	0.72	1.51	1.44	1.48	1.92	1.47
	2.00	0.46	2.03	0.70	2.01	1.40	1.98	1.86	1.96
2.0	0.50	0.44	0.49	0.66	0.48	1.34	0.46	1.80	0.44
	0.75	0.40	0.74	0.61	0.73	1.23	0.69	1.65	0.67
	1.50	0.33	1.49	0.50	1.47	1.02	1.40	1.37	1.36
	2.00	0.31	1.99	0.46	1.96	0.94	1.88	1.26	1.83
4.0	0.50	0.33	0.47	0.51	0.46	1.05	0.41	1.43	0.39
	0.75	0.28	0.72	0.43	0.70	0.90	0.63	1.22	0.60
	1.50	0.21	1.46	0.32	1.43	0.66	1.33	0.89	1.27
	2.00	0.19	1.96	0.28	1.92	0.58	1.80	0.78	1.73