Power and Sample Size Calculations for Case-Control Studies of Gene-Environment Interactions with a Polytomous Exposure Variable

Ivo Foppa and Donna Spiegelman

Genetic polymorphisms may appear to the epidemiologist most commonly as different levels of susceptibility to exposure. Epidemiologic studies of heterogeneity in exposure susceptibility aim at estimating the parameter quantifying the gene-environment interaction. In this paper, the authors use a general approach to power and sample size calculations for case-control studies, which is applicable to settings where the exposure variable is polytomous and where the assumption of independence between the distribution of the genotype and the environmental factor may not be met. It was found through exploration of different scenarios that in the cases explored, power calculations were relatively insensitive to assumptions about the odds ratio for the exposure in the referent genotype category and to assumptions about the odds ratio for the genetic factor in the lowest exposure category, yet they were relatively sensitive to assumptions about gene frequency, particularly when gene frequency was low. In general, to detect a small to moderate gene-environment interaction effect, large sample sizes are needed. Because the examples studied represent only a small subset of possible scenarios that could occur in practice, the authors encourage the use of their user-friendly Fortran program for calculating power and sample size for gene-environment interactions with exposures grouped by quantities that are explicitly tailored to the study at hand. Am J Epidemiol 1997; 146:596-604.

case-control studies; epidemiologic methods; interaction; epidemiology, molecular; sample size; statistical power; statistics; study design

The recent development of easily available and powerful assays to identify different alleles has led to an increased interest in gene-environment interactions, manifesting as different levels of susceptibility to the health effects of exposures. When the interaction effects are the parameters of interest, power and sample size calculations based on the magnitude of the main effects will give misleading answers, most likely overestimating the power and underestimating the required sample size. This leads to higher than expected frequencies of incorrectly rejecting the alternative hypothesis of an interaction effect. Thus, adequate methods for power and sample size calculations are required to correctly interpret the results from case-control studies of gene-environment interactions.

In a recent paper, Hwang et al. (1) presented power and required sample size calculations for case-control studies of gene-environment interactions, with a binary exposure variable and a genetic polymorphism that permits grouping into a susceptible and a nonsusceptible genotype. They found that for most values of exposure and genotype frequency, a fourfold difference in the odds ratio for the exposure in individuals with the susceptibility genotype versus the odds ratio for the exposure in those without (6.0 vs. 1.5) can be detected with sample sizes smaller than 600 (200 cases, 400 controls) with at least 80 percent power (see (1) table 2, p. 1,032).

However, as exposure status is often represented by more than two exposure categories, a more general method to calculate power and sample size than the one used by Hwang et al. would be desirable. In addition, it is well known that collapsing an ordinal variable into two categories is statistically inefficient. The purpose of this study therefore was to calculate power and sample size for case-control studies of gene-environment interactions with polytomous exposure variables, or of any other interaction between a binary and a polytomous variable. Furthermore, our aim was to investigate the relative importance of each of the parameters that determine power and required sample size. Finally, we wish to provide a user-friendly Fortran program to perform these calculations.
(request via the Internet: ifoppa@hsph.harvard.edu or stdls@gauss.bwh.harvard.edu).

**METHODS**

**Notation and assumptions**

The data from a case-control study on gene-environment interaction, as examined in this paper, could be summarized by a $2 \times 2 \times Q$ table, where $Q$ represents the number of ordinal exposure levels. The three variables $G_i$, $D_i$, and $E_i$ determine the cell of the $2 \times 2 \times Q$ table into which a subject is classified, where $E_i$ is the exposure score assigned to subject $i$ ($i = 1, \ldots, N$). Typically, $E_i = q_i - 1$, where $q_i \in \{1, \ldots, Q\}$ corresponds to quartiles of exposure level in the controls (i.e., 1 is the lowest, $Q$ is the highest exposure quantile), and $N$ refers to the sample size.

$$G_i = \begin{cases} 1 & \text{if subject } i \text{ carries the gene that is assumed to confer susceptibility} \\ 0 & \text{otherwise} \end{cases}$$

$$D_i = \begin{cases} 1 & \text{if subject } i \text{ is a case} \\ 0 & \text{otherwise} \end{cases}$$

Commonly, independence between gene and exposure in the controls is assumed (1). We also assumed this to be true, i.e.,

$$\Pr(G = g, E = e | D = 0) = \Pr(G = g | D = 0)\Pr(E = e | D = 0),$$

where $g = 1, 2$ and $e = 0, \ldots, Q - 1$. In the Discussion, we generalize our results to situations in which genotype and exposure covary in the controls and examine the impact of violations of this assumption on the resulting power and sample size. We assume the log odds ratio of disease to be a linear function of the exposure, i.e., the odds ratio between adjacent exposure categories is assumed to be constant for both individuals without the susceptibility genotype ($OR_{E=G=0}$) and for individuals with the susceptibility genotype ($OR_{E=G=1}$) (see equation 1). Thus, the top-to-bottom quantile contrast of the exposure effect in nonsusceptible individuals is $(OR_{E=G=0})^{Q-1}$; and in susceptible individuals, it is $(OR_{E=G=1})^{Q-1}$. We denote the gene effect in the lowest exposure category by $OR_{G=E=0}$, and the gene-environment interaction effect is denoted by $\theta$, where

$$\theta = \frac{OR_{G=E=1}}{OR_{G=E=0}},$$

which obviously is not an odds ratio itself but rather a ratio of two odds ratios. The top-to-bottom quantile interaction effect is defined as

$$\theta^* = \frac{(OR_{E=G=1})^{Q-1}}{(OR_{E=G=0})^{Q-1}} = \left(\frac{OR_{E=G=1}}{OR_{E=G=0}}\right)^{Q-1} = \theta^{Q-1}.$$

These definitions imply the logistic model

$$\logit[\Pr(D = 1 | E, G)] = \beta_0 + \beta_E E + \beta_G G + \beta_{GE} EG,$$

(1)

where $\beta_E = \log(OR_{E=G=0})$, $\beta_G = \log(OR_{G=E=0})$, and $\beta_{GE} = \log(\theta)$.

**Power calculation**

For some $\beta_{GE,H_A} > 0$, the asymptotic power to discriminate $H_A: \beta_{GE} = \beta_{GE,H_A}$ from $H_0: \beta_G = 0$ is

$$\psi(N, \beta_{H_A}) \approx 1 - \Phi\left(\frac{z_{1-\alpha/2} \sigma_{G,H_A} - \beta_{GE,H_A}}{\sigma_{G,E,H_A}}\right),$$

(2)

(see (2), p. 217), where $\Phi(.)$ is the cumulative standard normal distribution function, $z_{1-\alpha/2}$ is the value of a standard normal variable, $Z$, such that $\Pr(Z \geq z) = 1 - \alpha/2$, $\alpha$ is the size of the test, and $\sigma_{G,E,H_A}$ and $\sigma_{G,H_A}$ are the standard errors of the estimates of $\beta_{GE,H_A}$ under $H_0$ and $H_A$, respectively. Given a particular value of $\beta_{GE,H_A}$, the standard errors and therefore the power are a function of the following seven parameters (see table 1): the magnitude of the type I error ($\alpha$), the sample size ($N$), the control-case ratio ($C$), the gene frequency ($p_G$), the exposure distribution in the controls (which we assume to be quantiles with $Q$ levels), the odds ratio for the genotype in the lowest exposure category, $OR_{G=E=0}$, and the odds ratio for the exposure among noncarriers of the susceptibility genotype, $OR_{E=G=0}$. In addition, of course, the power is a function of $\theta$. We derive the explicit closed-form relation between the power and the eight parameters (Appendix). Briefly, the Fisher information matrix for the log-likelihood of the data based on equation 1 is used to calculate $\sigma_{G,E,H_A}$ and $\sigma_{G,H_A}$, which are then used in formula 2 to obtain the power. The expected cell counts used to calculate the Fisher information are generated from the assumptions described in "Notation and assumptions" and the specified values of the seven parameters.

If at least one of the expected cell counts in the $2 \times 2 \times Q$ table was less than 5, we assumed that the asymptotic standard errors would be biased. For such cases, the resulting power calculations were omitted or
otherwise marked. Future research is required to obtain more information about when asymptotic methods no longer apply in this setting.

Sample size calculations
To obtain the minimum required sample size for a given power, we need to solve the nonlinear equation

$$1 - \Phi\left(\frac{\zeta_{a} \sigma_{GEH_0} - \beta_{GEH_1}}{\sigma_{GEH_1}}\right) = \psi_0$$

with respect to $N$, where $\psi_0$ is the power to be achieved, and $\sigma_{GEH_0}$, $\sigma_{GEH_1}$ are functions of $N$ and are determined by the seven parameters given in table 1 and $\theta$. As there is no closed-form solution to this problem, we used the Newton-Raphson algorithm (3) with numerical derivatives. Typically, convergence was obtained after five to six iterations, where convergence was defined as a difference in the obtained sample size between two successive iterations of less than $10^{-8}$.

RESULTS
For the scenarios described below, $\theta^*$ and one additional parameter, $OR_{G|E=0}$, $OR_{E|G=0}$, or $P_G$, were varied, and the remaining six parameters were held fixed at their default values as given in table 1. In each graph, different curves correspond to different magnitudes of $\theta^*$.

Relation between $OR_{G|E=0}$ and power
In figure 1, the power is shown as a function of $OR_{G|E=0}$. The odds ratio for the genotype in the lowest exposure category, $OR_{G|E=0}$, was varied between 1.0 and 5.0 (in steps of 0.1). Scenarios in which at least one expected cell count was smaller than five were deleted from the graph.

The curves appeared to be almost flat with a peak in the range between $OR_{G|E=0} = 1.0$ and $OR_{G|E=0} = 2.0$, depending on the magnitude of the interaction effect. For rarer genes (e.g., $P_G = 0.05$ or $P_G = 0.01$, graphs not shown), the peaks were shifted toward larger values of $OR_{G|E=0}$. For $P_G = 0.05$, some of the curves were steeper ($\theta = 2.0$ to $\theta = 4.0$, not shown) for low values of $OR_{G|E=0}$ (i.e., between $OR_{G|E=0} = 1.0$ and $OR_{G|E=0} = 2.0$). The flatness of the curves suggests that, at least in situations in which the other parameters are close to the default values used here, accurate specification of $OR_{G|E=0}$ is probably not crucial for getting a reasonably accurate estimate of the power.

Relation between $OR_{E|G=0}$ and power
As shown in figure 2, the power was examined as a function of the $OR_{E|G=0}$. The resulting picture resembled figure 2, and again over a large range of $OR_{E|G=0}$, the power changed only slightly. Therefore, for the setting examined, exact knowledge of $OR_{E|G=0}$ is not shown.

TABLE 1. Notation and default values for the parameters determining power and sample size

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Default value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I error size</td>
<td>$\alpha$</td>
<td>Two-sided, 0.05</td>
</tr>
<tr>
<td>Sample size</td>
<td>$N = m_0^* + m_1^*$</td>
<td>600</td>
</tr>
<tr>
<td>Control-case ratio</td>
<td>$C = m_0/m_1$</td>
<td>2.0</td>
</tr>
<tr>
<td>Prevalence of the gene in controls, i.e., $Pr(G = 1</td>
<td>D = 0)$</td>
<td>$P_G$</td>
</tr>
<tr>
<td>Distribution of exposure in the controls, i.e., $Pr(E = e</td>
<td>D = 0)$ for $q = 1, \ldots, Q$</td>
<td>$P_a$</td>
</tr>
<tr>
<td>Odds ratio for the exposure in individuals without the susceptibility genotype (top-to-bottom)</td>
<td>$OR_{E</td>
<td>G=0}$</td>
</tr>
<tr>
<td>Odds ratio for the genotype in the lowest exposure category</td>
<td>$OR_{G</td>
<td>E=0}$</td>
</tr>
</tbody>
</table>

* $m_0^*$, total number of controls; $m_1^*$, total number of cases.
† NA, because $Pr(E = e|D = 0)$ is 0.2 in all examples, no symbol was used.

probably not crucial for obtaining accurate estimates of the power, regardless of the number of quantiles chosen (not shown).

Relation between gene frequency \( p_G \) and power

In figure 3, the power of studies to detect various magnitudes of \( \theta^* \) is shown for different \( p_G \). The power to detect a small interaction effect \( (\theta^* = 2.0) \) was low for all gene frequencies. For stronger interaction effects, the power increased considerably as gene frequency increased to approximately 30 percent and then began to decrease again, although slowly. For gene frequencies less than 0.30 and for moderate gene-environment interactions, the power was sensitive to the value of \( p_G \). For stronger "background effects" of the genotype \( (e.g., \ OR_{G,E=0} = 3.0, \ graph \ not \ shown) \), the slope of the curves became even steeper but the range of highest sensitivity shifted toward smaller values of \( p_G \). Misspecification of the gene frequency can therefore result in over- or underestimation of the study power. For common gene frequencies \( (e.g., > 30 \ percent) \), the power was relatively insensitive to assumptions about the actual values of \( p_G \) used.

Relation between case-control ratio and power

In figure 4a, the power is shown graphed as a function of the control-case ratio, given a fixed sample size \( (n = 600) \). For all magnitudes of the interaction effect considered, the power decreased as the number of controls increased due to the reduced number of cases. Given the availability of a sufficient number of cases or a predetermined budgetary constraint, the optimal control-case ratio therefore is 1:1. However, if the number of cases is fixed \( (m_1 = 200) \), increasing the number of controls increases the power (figure 4b). The gain was greatest when the ratio was increased from 1 to 2 and only very little gain was evident for \( C \geq 4 \). Depending on whether the number of available cases or the overall cost is the limiting factor in study design, different control-case ratios may be most cost efficient.

Sample size calculations

In tables 2 and 3, sample sizes required to attain 80 and 90 percent power are presented for different \( p_G \), \( \theta^* \), \( \ OR_{E,G} = 0 \), and \( C \), respectively. Since sample size calculations are closely related to power calculations, the patterns found previously emerge again. For most scenarios we considered, the required sample sizes were prohibitively large. For most settings yielding required sample sizes less than 1,600, at least one of the expected cell sizes was smaller than 5. In these cases, asymptotic statistical methods such as those used in this paper may not be valid.
Among the interaction effect sizes examined, \( \theta^* \) of 3.0 or less uniformly required sample sizes larger than 1,000—often considerably larger. The sample size was less than 1,000 only if the \( \theta^* \) was 6.0 or greater and the gene was relatively common (\( p_G \) of 0.1 or greater to attain 80 percent power, and 0.25 or higher to attain 90 percent power).

**DISCUSSION**

We have presented power and sample size calculations for case-control studies of gene-environment interactions where exposure is represented by a polytomous, ordinal exposure variable and where there are a susceptible and a nonsusceptible genotype. This is a generalization of the setting considered by other authors (1, 4), who presented sample size and power calculations for case-control studies of gene-environment interactions with a binary exposure variable. Although the motivation for our work was to aid the design of case-control studies of gene-environment interactions, it should be noted that the results apply to a study of any ordinal polytomous exposure and binary exposure. Also, approaches for interactions of two dichotomous variables can be used with polytomous exposure variables after dichotomization; however, due to nonoptimal use of the data, this results in a loss of power. In what follows, we discuss the practical relevance of our results and then examine some potential shortcomings of our approach.

Of the parameters \( OR_{G=0} \), \( OR_{E=0} \). \( OR_{G=E=0} \), and \( p_G \), the gene frequency (\( p_G \)) most critically influences the power of a study to detect a given gene-environment interaction in settings similar to the ones examined here. If the prevalence of the genotype associated with exposure susceptibility is known to be common, i.e., approximately as common as the nonsusceptible genotype, then the lack of precise knowledge of \( p_G \) probably will not influence the power. This is the case for some of the genetic polymorphisms that are suspected to be involved in gene-environment interactions. For example, the GSTM1 nulled genotype, which is thought to be associated with an increased risk of several cancers, was shown to have a prevalence of 41.8 percent in a control population (5). In another study, a prevalence of 35 percent was found in a community sample of African-Americans whereas among Caucasians, the prevalence was 50 percent (6). Similarly, the slow acetylation phenotype, which is strongly associated with \( N \)-acetyltransferase genotype (7) and is related to increased susceptibility to bladder cancer (8, 9), has been found in 51-67 percent of different populations (9). However, if the gene frequency is relatively low, as for BRCA1, which has a prevalence estimated to be much less than 1 percent (10), precise knowledge of the actual value of \( p_G \) may be crucial to obtain an accurate estimate of the power. If, for example, the estimate of \( p_G \) used in the calculations is 10 percent but the true value is 5 percent (\( m_1 = 600, C = 2, OR_{E=0}^{BG} = 1.5, OR_{G=0}^{BG} = 1.5, \theta^* = 3.0 \), the power estimate would be 0.74 whereas the true power was 0.47 (see also figure 3). In the range of \( p_G \) between 30 percent and 40 percent, an error of similar magnitude will hardly affect the power or sample size calculated (for \( p_G = 0.3 \), the power is 0.95, whereas for \( p_G = 0.4 \), the power is 0.96). If a rough estimate of the distribution of the genotype in the population is available before a case-control study of a gene-environment interaction is conducted, the investigator may use this information to examine the potential impact of misspecification of \( p_G \) on the power and sample size estimates.

However, the power was relatively insensitive to assumptions about the baseline odds ratio for the gene.
notype, $OR_{E|G} = 0$, especially for small effects ($OR_{E|G} = 0 \leq 2.0$). Similarly, the background odds ratio for the exposure, $OR_{E|G} = 0$, only slightly affected the power to detect a given gene-interaction effect. Thus, inaccurate estimates of $OR_{E|G} = 0$ and $OR_{E|G} = 0$ are likely to yield reasonably accurate estimates of power and required sample size.

The case-control ratio strongly influences the study power if the number of cases is fixed. The increasing power from increasing the sample size from 400 to 600 by adding controls is dramatic. An additional increase in the case-control ratio is less effective in increasing the power and therefore may be increasingly less cost effective.

Our approach is easily generalizable to situations in which it is not reasonable to assume that gene and exposure are independent in the study population (see Appendix). In fact, the independence assumption may be wrong in many cases. In the example cited above (6), the distribution of the GSTM1 nulled genotype was found to vary among different ethnic groups (35 percent in African-American, 50 percent in Caucasian). Many relevant behavioral and environmental risk factors also vary by ethnic group. In the binary exposure case, we found that assuming independence of gene and exposure in the controls gives a higher power than when independence does not hold. For a moderate strength of dependence, where the odds of being exposed in individuals with the susceptibility genotype versus individuals without the susceptibility genotype, $OR_{E|G} = 0$, was between 0.4 and 2.0, and the values of the other parameters were as given in table 1, the calculated power was only slightly overestimated if the assumption of independence was made. For example, if we assumed independence, the estimated power to detect $\theta^*$ of 4.0 was 0.91; if $OR_{E|G} = 0$ was 0.43, the power was actually 0.90; and if $OR_{E|G} = 0$ was 2.33, the power was actually 0.89. If $\theta^*$ was only 3.0, the corresponding values for the power were 0.79 (independence), 0.78 ($OR_{E|G} = 0 = 0.43$), and 0.77 ($OR_{E|G} = 0 = 2.33$). Assuming lower gene frequencies, the error due to the incorrect independence assumption became slightly larger. Because, at least in the examples considered, the power was relatively insensitive to $OR_{E|G} = 0$, $OR_{E|G} = 0$, and $P_G$, it is not surprising that the power was not very

---

### Table 2. Minimum required sample sizes to achieve 80% and 90% power of detecting a particular top-bottom interaction effect ($\theta^*$) as functions of $p_G$ and $OR_{E|G} = 2$.

| $OR_{E|G} = 0$ | $\theta^*$ | Power = 0.8 | Power = 0.9 |
|---------------|-------------|-------------|-------------|
|               | 0.05 | 0.1 | 0.25 | 0.5 | 0.05 | 0.1 | 0.25 | 0.5 |
| 1.5           | 1.5  | 28,596 | 15,600 | 8,262 | 7,272 | 37,994 | 20,757 | 11,042 | 9,785 |
| 3.0           | 3.0  | 3,974 | 2,086 | 1,138 | 1,043 | 4,997 | 2,762 | 1,520 | 1,424 |
| 6.0           | 6.0  | 1,410 | 765  | 441  | 426  | 1,850 | 1,037 | 596  | 597  |
| 3.0           | 0.1  | 30,810 | 16,826 | 8,937 | 5,868 | 40,986 | 22,419 | 11,963 | 10,646 |
| 3.0           | 3.0  | 1,400 | 765  | 441  | 426  | 1,850 | 1,037 | 596  | 597  |
| 6.0           | 6.0  | 1,410 | 765  | 441  | 426  | 1,850 | 1,037 | 596  | 597  |

* All other parameters were held fixed at their default values.
† For values in italics, at least one of the expected cell counts was <5.

### Table 3. Minimum required sample sizes to achieve 80% and 90% power of detecting a particular top-bottom interaction

| $OR_{E|G} = 0$ | $\theta^*$ | Power = 0.8 | Power = 0.9 |
|---------------|-------------|-------------|-------------|
|               | 0.05 | 0.1 | 0.25 | 0.5 | 0.05 | 0.1 | 0.25 | 0.5 |
| 1.5           | 1.5  | 27,154 | 14,670 | 7,558 | 6,386 | 36,166 | 19,558 | 10,108 | 8,584 |
| 3.0           | 3.0  | 3,638 | 1,978 | 1,038 | 906  | 4,820 | 2,626 | 1,392 | 1,234 |
| 6.0           | 6.0  | 1,360 | 744  | 400  | 366  | 1,798 | 990  | 542  | 508  |
| 3.0           | 3.0  | 3,882 | 2,114 | 1,118 | 986  | 5,154 | 2,818 | 1,504 | 1,348 |
| 6.0           | 6.0  | 1,456 | 802  | 436  | 404  | 1,932 | 1,070 | 594  | 566  |
| 6.0           | 6.0  | 1,368 | 740  | 400  | 366  | 1,798 | 990  | 542  | 508  |

* All other parameters were held fixed at their default values.
† For values in italics, at least one of the expected cell counts was <5.

---

**Am J Epidemiol** Vol. 146, No. 7, 1997

---

Downloaded from http://aje.oxfordjournals.org/ at Ernst Mayr Library, the Museum Comp Zoology, Harvard University on April 23, 2013
sensitive to assumptions about the joint distribution of $E$ and $G$. In the polytomous case, the situation is more complicated because there are numerous possibilities for specifying the joint distribution of exposure and gene in the controls. However, sensitivity analysis may be useful in determining the impact of the different scenarios for the joint distribution of the gene and the environmental variables in the controls on study power, and thus in determining the approximate magnitude of error in power and sample size calculations, which could be due to incorrect assumptions about the joint distribution of the gene and the environmental variables in the controls.

We assumed a logistic model to be true. However, it may also be of interest to examine the effects of model misspecification on the power in the case of a polytomous exposure variable, as Greenland has done for the binary case (11). Furthermore, it would be desirable to more closely examine models more complex (12) than the ones currently routinely used in epidemiology and to investigate the power to detect more complex associations between the polytomous exposure variable and disease in a susceptible and in a nonsusceptible group. Another issue worthy of additional investigation is that for many scenarios, asymptotics may not be valid because of small expected cell sizes. Simulation studies may be useful to determine how accurate power and sample size estimates are in these cases. Furthermore, if rare genes are of particular interest, adaptations of the case control approach for family studies may offer more power than a conventional case-control design (13).

One of the most noteworthy results of our investigation is given in tables 2 and 3. For most combinations of parameter values examined, very large samples were needed to detect gene-environment interactions with reasonable power. This finding agrees with Smith and Day’s finding for case-control studies with binary exposure variables (4). However, Hwang et al. (1) concluded that—again for binary exposure—except for very high or very low exposure frequencies and for common susceptibility genotypes, “modest sample sizes will be adequate to detect gene-environment interaction” (p. 1036). The reason for this discordance lies in the fact that a $\theta^*$ of 4.0 for a binary exposure variable (as used in that paper) corresponds to a much higher $\theta^*$ if the exposure variable is polytomous. Thus, a “moderate” interaction effect in the binary case would be a “strong” effect in the polytomous case.

It is important to note that the examples studied in this paper represent only a small subset of possible scenarios that could occur. Rather than relying on the figures and tables that are most applicable to scenarios with default parameter values corresponding to table 1, we encourage readers to obtain our program and perform the calculations exactly as described the study at hand.

ACKNOWLEDGMENTS

This work was supported by grants NIH CA55075, CA65725 (D. S.), and 823B-040075 (I. F.) from the Swiss National Science Foundation.

The authors thank Mario Casella and Carrie Wager for their computer assistance and David Hunter and Edward Giovannucci for their helpful suggestions.

REFERENCES

APPENDIX

Derivation of the asymptotic standard error of $\beta_{OE}$ under $H_0$ and $H_A$

The asymptotic standard errors for $\hat{\beta}_{GE}$, $\sigma_{GE|H_0}$, and $\sigma_{GE|H_A}$ were used for power and sample size calculations. They were derived as the square root of the negative inverse of the expected information matrix of the log-likelihood of the data implied by equation 1.

$$L(\beta|X) = \prod_{i=1}^{N} \left( \frac{\exp(\beta_0 + \beta_E E_i + \beta_G G_i + \beta_{GE} G_i E_i)}{1 + \exp(\beta_0 + \beta_E E_i + \beta_G G_i + \beta_{GE} G_i E_i)} \right)^{D_i} \left( \frac{1}{1 + \exp(\beta_0 + \beta_E E_i + \beta_G G_i + \beta_{GE} G_i E_i)} \right)^{1-D_i},$$

(A.1)

where $X = (D_i, E_i, G_i), i = 1, \ldots, N,$ and $\beta_{GE}$ is zero under $H_0$ and has some specified value different from zero under $H_A$.

To denote all possible covariate patterns, we define the matrix $Z$, where

$$Z = \begin{bmatrix}
1 & 1 & 1 & \ldots & 1 & 1 & 1 & \ldots & 1 \\
0 & 0 & 0 & \ldots & 0 & 1 & 1 & \ldots & 1 \\
0 & 1 & 2 & \ldots & Q-1 & 0 & 1 & 2 & \ldots & Q-1 \\
0 & 0 & 0 & \ldots & 0 & 0 & 1 & 2 & \ldots & Q-1
\end{bmatrix}.$$

Let $z_j$, for $j = 1, \ldots, 2 \times Q$, be the $j$th column of $Z^T$, where the rows of $Z^T$ correspond to $1, G, E,$ and $G \times E$. Furthermore, let $z_{rs}$ be the $(r, s)$ element of $Z$ with $r = 1, \ldots, 2 \times Q$ and $s = 1, \ldots, 4$.

Under $H_0$, the $(u, v)$ element of the expected $4 \times 4$ information matrix for the likelihood (A.1) is

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

where $\beta_{H_0} = [\beta_0|H_0, \beta_E, \beta_G, 0]$. To calculate $E(m_{1,j})$, we make use of the relation between the assumed proportional hazards model in the underlying cohort, which gave rise to the case-control study, and the logistic regression model for the case-control study itself (14). Assuming the proportional hazards model for the disease incidence rate in the underlying cohort, the incidence rate for a fixed value of $E$ and $G$ under $H_0$, $I(E, G|H_0)$, is equal to $I_0 \times \exp(\beta_E E + \beta_G G)$, where $I_0$ is the baseline incidence rate, the expected value for the number of cases in the source population with covariate pattern $z_j$ is $E(M_{1,j|H_0})$,

$$E(M_{1,j|H_0}) = I_0 \times \exp(\beta_E z_{j2} + \beta_G z_{j3}) \times \Pr(E = z_{j2}, G = z_{j3}|D = 0),$$

and the expected value for the number of cases sampled at random from the source population, $E(m_{1,j|H_0})$, is

$$E(m_{1,j|H_0}) = f_c \times I_0 \times \exp(\beta_E z_{j2} + \beta_G z_{j3}) \times \Pr(E = z_{j2}, G = z_{j3}|D = 0),$$

where $f_c$ is the case sampling fraction. The probability of a certain covariate pattern $z_j$ among the cases in the study is

$$\Pr(E = z_{j2}, G = z_{j3}|D = 1, H_0) = \frac{f_c \times E(M_{1,j|H_0})}{\sum_{k=1}^{2\times Q} f_c \times E(M_{1,k|H_0})},$$

(A.2)

and

$$E(m_{0,j}|H_0) = \Pr(E = z_{j3}, G = z_{j3}|D = 1) \times m_0.$$ 

The expected number of controls with covariate pattern $z_j$, under $H_0$ and $H_A$, is

$$E(m_{0,j}) = \Pr(E = z_{j3}, G = z_{j3}|D = 0) \times m_0.$$ 

(A.3)

Under the assumption of independence of $E$ and $G$ in the controls $\Pr(E = z_{j3}|D = 0) \times \Pr(G = z_{j2}|D = 0)$ is substituted for $\Pr(E = z_{k2}, G = z_{k3}|D = 0)$ in the expressions A.2 and A.3.
Similarly, under $H_A$, the $(u, v)$ element of the expected information matrix is

$$E\left(\frac{\partial^2 \ell(\beta|Z)}{\partial \beta_u \partial \beta_v | H_A}\right) = -\sum_{j=1}^{2 \times Q} z_{ju} z_{jv} \times \left[ E(m_{0j}|H_A) + E(m_{1j}|H_A) \right] \times \frac{\exp(\beta_{H_A}^T z_j)}{[1 + \exp(\beta_{H_A}^T z_j)]^2},$$

with $\beta_{H_A} = [\beta_0|H_A, \beta_E, \beta_G, \beta_{GEM H_A}]$ and

$$E(m_{0j}) + E(m_{1j}|H_A) = \frac{\exp(\beta_{H_A}^T z_j) \times \Pr(E = z_\beta, G = z_\beta|D = 0)}{\sum_{k=1}^{2^Q} \exp(\beta_{H_A}^T z_k) \times \Pr(E = z_\beta, G = z_\beta|D = 0) \times m_1} \times m_1$$

$$+ \Pr(E = z_\beta|D = 0) \times \Pr(G = z_\beta|D = 0) \times m_0.$$

Asymptotic standard errors of $\hat{\beta}_{GE}, \sigma_{GE|H_A}$, and $\sigma_{GEM H_A}$ are the square root of the $(4, 4)$ element of the negative inverse of the expected information matrix under $H_0$ and $H_A$, respectively. Although the expression for these quantities is complex, it does have a closed form.