STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:36–46
Published online 18 July 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2980

Sample size and optimal design for logistic regression with binary interaction

Eugene Demidenko*, †

Dartmouth Medical School, Hanover, NH 03755, U.S.A.

SUMMARY

There is no consensus on what test to use as the basis for sample size determination and power analysis. Some authors advocate the Wald test and some the likelihood-ratio test. We argue that the Wald test should be used because the Z-score is commonly applied for regression coefficient significance testing and therefore the same statistic should be used in the power function. We correct a widespread mistake on sample size determination when the variance of the maximum likelihood estimate (MLE) is estimated at null value. In our previous paper, we developed a correct sample size formula for logistic regression with single exposure (*Statist. Med.* 2007; **26**(18):3385–3397). In the present paper, closed-form formulas are derived for interaction studies with binary exposure and covariate in logistic regression. The formula for the optimal control–case ratio is derived such that it maximizes the power function given other parameters. Our sample size and power calculations with interaction can be carried out online at www.dartmouth.edu/~eugened.

KEY WORDS: binary data; gene–gene interaction; gene–environment interaction; information matrix; likelihood-ratio test; optimal allocation problem; Wald test

1. INTRODUCTION

Recent advances in genomics and declining cost of microarray analysis have increased the popularity of gene–environment and gene–gene interaction epidemiological studies. 'How many controls and cases must there be to achieve the desired power and what should be their proportion' are essential questions at the very early stage of study design.

A vast literature on sample size determination for logistic regression with interaction can be roughly divided into four groups with respect to the statistical test used: test on proportions, assuming that all variables are binary, and general statistical tests, such as likelihood-ratio, score, and Wald tests.

*Correspondence to: Eugene Demidenko, Dartmouth Medical School, Hanover, NH 03755, U.S.A.
†E-mail: eugened@dartmouth.edu

Received 25 October 2006
Accepted 18 May 2007

When all variables are binary, the logistic regression with interaction can be expressed as a $2 \times 2 \times 2$ contingency table. Then the test on interaction reduces to the Z-test on proportions—sometimes this test is called the Woolf test [1, 2]. Smith and Day [3] were, perhaps, the first authors to apply this idea to sample size determination with interaction. A limitation of their method is that they assumed an equal number of cases and controls. Hwang *et al*. [4] eliminated that assumption, but assumed that the exposure and covariate are independent. The same approach under the assumption of an equal number of cases and controls was later used by Yang *et al*. [5].

Whittemore [6] was the first to use the Wald test for sample size determination with logistic regression. To avoid the exact computation of the Fisher information matrix, she suggested an approximation when the response rate is small. Hsieh *et al*. [7] extended that approach and Shieh [8] compared this approximation with the likelihood-ratio test by Monte Carlo simulations. Foppa and Spiegelman [9] used the Wald test to determine the sample size in logistic regression with interaction when the gene (exposure) variable is categorical (the executable program can be downloaded at http://www.hsph.harvard.edu/faculty/spiegelman/ge_trend_v2.html).

Lubin and Gail [10] used the score test in the framework of logistic regression. Self and Mauritsen [11] and Shieh [12] applied the likelihood-ratio test for power analysis and sample size determination in a generalized linear model. Although general, these approaches need to be adapted to interaction studies because they require the specification of the joint distribution with the interaction variable being the product of the exposure and covariate. Gauderman [13, 14] developed an algorithm to derive the sample size for logistic regression with interaction using the likelihood-ratio test (one can freely download Windows-based software, QUANTO, based on this approach at http://hydra.usc.edu/gxe).

We, however, advocate the Wald test as the basis for power analysis and sample size determination. Our motivation is as follows. By definition, power is the probability of rejecting the null hypothesis evaluated at the alternative. Since the method for testing the significance of the regression coefficient, η, uses the Z-statistic,

$$Z = \frac{\widehat{\eta}_{\mathrm{ML}}}{\mathrm{SE}(\widehat{\eta}_{\mathrm{ML}})} \tag{1}$$

the power is $\Pr(|Z| > Z_{1-\alpha/2})$, where $Z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution. But this is the Wald test! Thus, the same test should be used for sample size determination. Although the Wald and likelihood-ratio tests are equivalent in the neighborhood of the null, these tests are different globally in large samples [15]. Consequently, the power and the respective sample size derived by the two tests will differ.

In the literature mentioned above, as well as in many other articles that use the Wald test, the sample size was calculated using the following formula for the total number of observations:

$$n = \frac{(Z_{1-\alpha/2}\sqrt{V_0} + Z_P\sqrt{V})^2}{\eta^2} \tag{2}$$

where V_0 is the variance of the maximum likelihood estimate (MLE) evaluated at the null hypothesis, $H_0 : \eta = 0$, and V is the variance evaluated at the MLE, $\widehat{\eta}_{\mathrm{ML}}$. We shall refer to this formula as to the *null-variance* formula. We guess that formula (2) emerged in connection with testing of the proportion, $H_0 : p = p_0$, where $V_0 = p_0(1-p_0)/n$ is the variance of the proportion in n

DOI: 10.1002/sim

Bernoulli trials under the null. But in the existing software, the variance and the test statistic, Z, are never evaluated at the null but at the MLE, as in (1). As another argument, we draw a parallel to the coefficient significance testing, $H_0 : \beta_i = 0$, in linear model using the t-test by computing $\widehat{\beta}_i / \sqrt{s^2 (\mathbf{X}'\mathbf{X})^{-1}_{ii}}$. Following (2), to estimate s^2, we should compute the residual sum of squares at $\beta_i = 0$ but we never do this. Remarkably, nobody *derives* the null-variance formula although it travels from one paper to another.

Instead, we suggest

$$n = \frac{(Z_{1-\alpha/2} + Z_P)^2}{\eta^2} V \tag{3}$$

for our sample size calculation. Formulas (2) and (3) usually produce close results, especially when the alternative odds ratio (OR) is close to 1. However, they may produce differences up to 30 per cent otherwise, as was shown in the previous paper [15].

The sample size formula (3) is general and can be applied in many settings and models. For example, formula (3) is widely used when comparing means of two groups with normal distribution [16], but for some reasons the null-variance formula (2) is used for logistic regression.

In our previous paper, we had developed a correct sample size formula for logistic regression with single exposure [15]. The goal of the present paper is to derive the Wald-based, closed-form power and sample size formulas for logistic regression with binary exposure and covariate and their interaction with no limitation on the design specification, such as an equal number of cases and controls or independence of the environment and gene factors. Obviously, alternatively, one could employ a categorical or continuous covariate design if the information on their distribution is available; however, the closed-form solution would not be available.

There is principally no difference between cohort and case–control studies in terms of parameter estimation and testing, as was shown by Prentice and Pyke [17]. Thus, we do not distinguish two designs. Since the intercept term in a case–control study determines the ratio of cases and normal subjects in the normal group (zero exposure and zero covariate), we seek an optimal study with minimum total sample size that yields the predefined power.

2. POWER AND SAMPLE SIZE FOR INTERACTION

In this section, we derive the Wald power and sample size for the interaction coefficient, η, in logistic regression defined by

$$\Pr(y = 1 | x, z) = \frac{\mathrm{e}^{\alpha_0 + \beta x + \gamma z + \eta(xz)}}{1 + \mathrm{e}^{\alpha_0 + \beta x + \gamma z + \eta(xz)}} \tag{4}$$

Variables x and z, and therefore the interaction term, xz, are binary. To be specific, we may assume that y codes the disease status, x is the exposure indicator, and z is a genotype at a disease-susceptibility locus with a single allele. Note that variables x and z may be interpreted in various ways that obviously will not affect the required sample size. For example, x may represent other genetic binary information, say a genotype at a secondary disease-susceptibility locus with a different allele. Then xz reflects the gene–gene interaction. To shorten the interpretation, we simply say that $z = 0$ corresponds to a *good* gene and $z = 1$ to a *bad* gene. For example, if exposure is a

DOI: 10.1002/sim

smoking status and y is cancer occurrence, the interaction xz reflects an elevated risk of cancer for a smoker with a bad gene.

The null hypothesis is $H_0 : \eta = 0$ with the alternative $H_A : \eta \neq 0$. The same model was used by Gauderman [13, 14], although he applied the likelihood-ratio test and we apply the Wald test. The power function of the Wald test, as the probability of rejecting the null when the alternative is true, can be well approximated as

$$\text{Power} = \Phi\left(-Z_{1-\alpha/2} + \frac{\eta\sqrt{n}}{\sqrt{V}}\right) \tag{5}$$

where $Z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution function, Φ; α is the test size (typically, $\alpha = 0.05$); n is the sample size; and V is the asymptotic variance of $\sqrt{n}\widehat{\eta}_{\text{ML}}$. One can use (5) to find the minimum detectable difference, η, or the sample size given power P (typically, $P = 80$ per cent or $P = 90$ per cent). For example, the sample size required to detect the interaction log OR, η with power P and significance level α is given by (3).

The major step in obtaining the power is computation of the variance V as a function of the regression coefficients. To shorten the notation, we use capital letters to denote the exponents of the regression coefficients. For example, $A = \mathrm{e}^{\alpha_0}$, $B = \mathrm{e}^{\beta}$ is the OR of the individual effect of the exposure, $G = \mathrm{e}^{\gamma}$ is the OR of the gene, and $K = \mathrm{e}^{\eta}$ is the interaction OR. To define probabilities for x and z, we use the notations $p_x = \Pr(x = 1)$ and $p_z = \Pr(z = 1)$. The relationship between x and z can be expressed as a conditional probability specified by logistic regression, namely,

$$P(x = 1|z) = \frac{\mathrm{e}^{c+\delta z}}{1 + \mathrm{e}^{c+\delta z}}$$

Parameter δ is defined *via* the OR $D = \mathrm{e}^{\delta}$, and a positive parameter $C = \mathrm{e}^{c}$ is found from the quadratic equation

$$1 - p_x = \frac{p_z}{1 + CD} + \frac{1 - p_z}{1 + C}$$

It is elementary to show that this equation has a unique positive solution

$$C = \frac{q + \sqrt{q^2 + 4p_x(1 - p_x)D}}{2(1 - p_x)D}$$

where $q = p_x(1 + D) + p_z(1 - D) - 1$. In practice, it is often assumed that gene and environment factors are independent, implying $D = 1$ [13]. Then, the above equation gives $C = p_x/(1 - p_x)$. However, in some instances, such as when z is a confounder, x and z may correlate.

The 4×4 information matrix for model (4) is derived in the Appendix. The variance of $\sqrt{n}\widehat{\eta}_{\text{ML}}$, as the (4,4)th element of the inverse matrix, is given by

$$V = \frac{1}{L} + \frac{1}{R} + \frac{1}{F} + \frac{1}{J} \tag{6}$$

DOI: 10.1002/sim

where quantities L, R, F, and J are defined as follows:

$$L = \frac{A(1-p_z)}{(1+A)^2(1+C)}, \quad R = \frac{ABCDGKp_z}{(1+ABGK)^2(1+CD)}$$

$$F = \frac{ABC(1-p_z)}{(1+AB)^2(1+C)}, \quad J = \frac{AGp_z}{(1+AG)^2(1+CD)} \tag{7}$$

Given V, the power and sample size are computed by formulas (5) and (3). Since we give the complete inverse information matrix, the sample size may be determined not only for the interaction term but also for any other coefficient in regression (4).

To compute the required sample size, we need to specify four groups of parameters (nine parameters in total):

1. Three parameters in formula (3): the significance level α; the power P; and the alternative $\log \mathrm{OR}$, $\eta = \ln K$.
2. Three parameters that specify the joint distribution of the exposure and the gene: the proportion of subjects in the general population with exposure, $p_x = \Pr(x=1)$; the proportion of subjects in the general population with the bad gene, $p_z = \Pr(z=1)$; and the OR between exposure and gene D.
3. Two ORs as individual effects of the exposure, $B = \mathrm{e}^{\beta}$, and the gene, $G = \mathrm{e}^{\gamma}$, with ORs computed based on coefficients from model (4).
4. The proportion of diseased subjects in the general population with no exposure and a good gene (baseline prevalence), $p_y = \Pr(y=1|x=0, z=0) = A/(1+A)$. In a case–control study, this corresponds to the proportion of cases among subjects with zero exposure and covariate.

Note that the above specifications of p_x and p_z assume marginal prevalence probabilities. Alternatively, we could specify conditional probabilities, such as probability of exposure in controls, as in [9]. For example, A may be determined using the ratio of *total* number of controls to subjects from the equation

$$\frac{p_{00}}{1+A} + \frac{p_{01}}{1+AG} + \frac{p_{10}}{1+AB} + \frac{p_{11}}{1+ABGK} = \frac{n_0}{n} \tag{8}$$

where four probabilities, $\{p_{ij}, i=0, 1; j=0, 1\}$, are defined in the Appendix. *Vice versa*, if A is determined, as in optimal design (Section 3), the proportion of controls to n is equal to the left-hand side of equation (8). Note that A is the expected ratio of cases to controls in the *normal* group (*no exposure*, *good gene*), while $(n-n_0)/n_0$ is the ratio of cases to controls in the entire sample. For example, an equal number of controls and cases in the entire sample does not imply an equal design in the normal group ($A=1$). Formula (8) should be used for the back and forth calculation.

Another comment is regarding setting up the value for the alternative OR of interaction, K. Following the line of common reasoning, we compare the effect of two groups, $x=0$, $z=0$ and $x=1$, $z=1$, which on the logit scale is $\Delta = \beta + \gamma + \eta$ with the interaction effect $\eta = \Delta - (\beta + \gamma)$. Thus, the pure interaction or *synergistic* effect on the OR scale is $K/(BG)$. Consequently, if we want to detect a synergistic OR, K_S, we set $K = K_S BG$.

DOI: 10.1002/sim

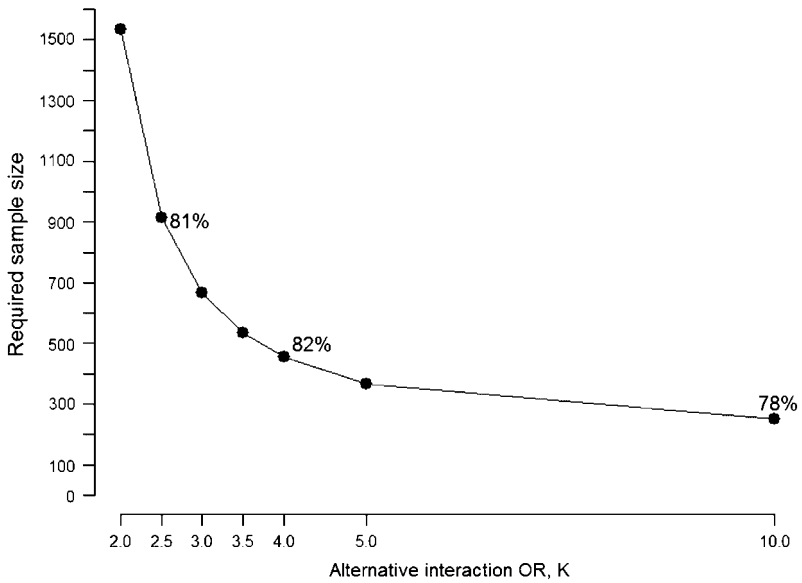

Figure 1. The required sample size for a gene–gene interaction study as a function of the alternative OR computed using the Wald test with the nominal power 80 per cent. There is good agreement between theoretical power and empirical power from simulations.

2.1. Example: study of G×G interaction for asthma

We use an example of gene–gene interaction considered by Gauderman [13]. As suggested by Gilliland *et al*. [18], two gene candidates are associated with the risk of asthma in children, myeloperoxidase (gene GSTM1) and tumor necrosis factor-α (gene GSTT1). Assuming that a case–control study is designed with an equal number of cases and controls, we wish to know the sample size to achieve power $P=0.8$ under significance level $\alpha=0.05$. Following Gauderman, we assume that the prevalence of the two genes is $p_x=0.4$ and $p_z=0.25$. Since a 50/50 design is suggested, $p_y=0.5$, so that $A=1$ (equal number of cases and controls in the *normal* group). A pure interaction model is assumed in which neither gene increases the risk by itself, so $B=G=1$. Although Gauderman did not specify it, we assume that the gene effect is independent, so that $D=1$.

The results of the sample size determination based on the Wald test for seven alternative interaction ORs are shown in Figure 1. To assess the difference between the theoretical and empirical power, we use Monte Carlo simulations based on 1000 experiments. The numbers at three ORs show the per cent of experiments when $|Z|>1.96$ for the interaction coefficient. As seen from this figure, the empirical power is close to the nominal for the three very diverse values of ORs considered.

3. OPTIMAL DESIGN OF THE CASE–CONTROL STUDY

As mentioned in the Introduction, the standard maximum likelihood theory for unmatched case–control studies holds and therefore the variance of the MLE and the sample size formulas derived

DOI: 10.1002/sim

above remain valid. The only difference is that $p_y = A/(1+A)$ is not interpretated as the disease prevalence rate but simply as the proportion of cases in the sample. Since the proportion of cases and controls in a case–control study can be chosen as part of the study design, we may find an optimal p_y that minimizes the sample size, n, given the power probability, P. This problem is known as the 'optimal allocation problem' and several authors have studied it. For example, Brittain and Schlesselman [19] suggest an optimal p_y for the test on proportion that minimizes the variance or maximizes the power. For optimal design/proportion of cases, we seek the p_y that minimizes the required sample size to achieve a given power P. Thus, following formula (3), the problem of optimal design of case–control interaction studies reduces to minimization of V as a function of A. As shown in the Appendix, the optimal ratio of cases to controls is

$$A_{\text{opt}} = \sqrt{\frac{(1+BC)\omega + (1+BCDK)\phi}{B(C+B)\omega + BG^2K(BK+CD)\phi}} \tag{9}$$

where

$$\omega = (1+C)DGKp_z, \quad \phi = (1+CD)(1-p_z)$$

It is elementary to check that when individual effects of exposure and gene are zero ($B = G = 1$) and exposure and gene are independent ($D = 1$), we have

$$A_{\text{opt}} = \sqrt{\frac{1+(K-1)(p_x+p_z-p_xp_z)}{K[K-(K-1)(p_x+p_z-p_xp_z)]}} \tag{10}$$

If $K = 1$ we have $A_{\text{opt}} = 1$. This means that the 50/50 design is optimal only when the alternative OR is 1. It should be noted that A_{opt} gives the optimal ratio of cases to controls in the group with zero exposure and good gene ($x = z = 0$). To obtain the optimal number of controls, n_0 in the total sample formula (8) should be used.

In Figure 2, we show the optimal number of cases to controls in the *normal* group ($x = z = 0$), computed by formula (10), and in the total sample, using formula (8), with several probabilities of x assuming that probability of z is 0.5. As in the figure, for $K > 1$ there should be less cases and more controls in the normal group, but when $K < 1$ the reverse is true. When p_x is close to 1 this proportion is almost 1, but it dramatically changes with alternative OR for small probability, p_x. The relationship between the optimal proportion of cases and controls in the total sample depends on the prevalence of the exposure. If $p_x < 0.5$ it resembles the previous case, but for large p_x it reverses. As evident from this graph, the ratio of cases to controls in the *normal* and entire group may be quite different, especially under extreme probabilities of exposure. This fact should be remembered when specifying the value of A for computing the power and n.

This approach can be extended to the design of cost-effective epidemiological studies that minimizes the cost function $p_0n_0 + p_1n_1$, where p_0 and p_1 are the costs of control and case.

3.1. Example continued

Recall, in our asthma example, the ratio of number of controls to cases in the *normal* group was 1. Assuming that in the alternative OR $K = 10$, we compute the variance of the interaction term per observation by formula (6), $V = 169.9$, so that the Wald power would be 80 per cent if the total sample size were $n = 252$. We can minimize V, and consequently the study, by using

DOI: 10.1002/sim

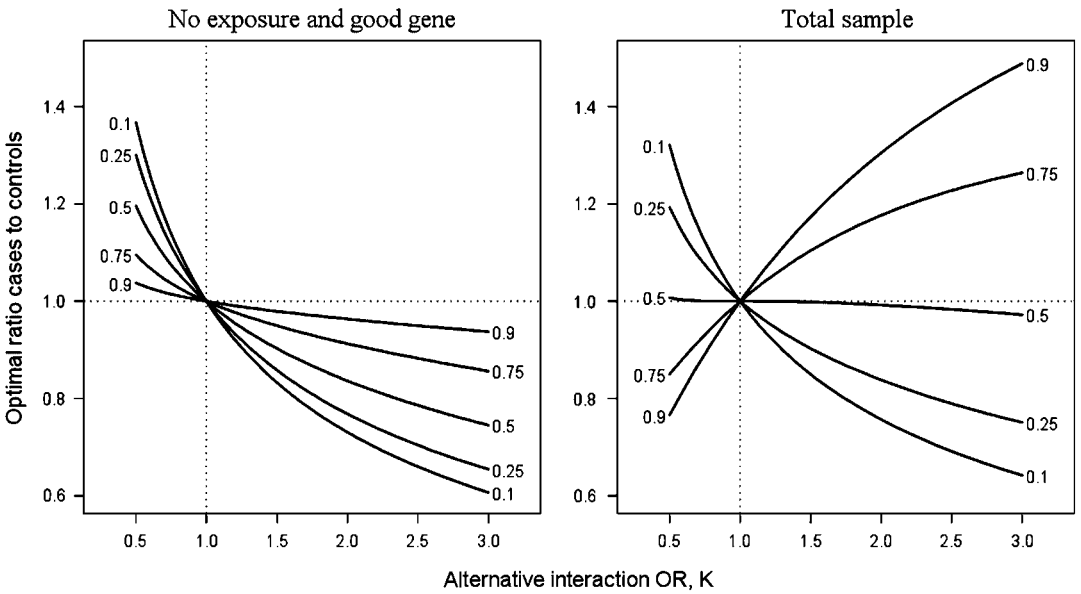

Figure 2. Optimal proportion of cases to controls in two groups for logistic regression with interaction for different values of p_x in the range from 0.1 to 0.9, and $p_z = \frac{1}{2}$.

the optimal case–control ratio. Since $B = G = D = 1$, we can apply formula (10), which yields $A_{\text{opt}} = 0.343$ with the minimum variance $V = 121.5$. As follows from formula (3), the reduction in the total sample size is $121.5/169.9 = 0.72$, about 30 per cent. Thus, the total sample size $n = 252 \times 0.72 = 180$ with the optimal ratio of cases to controls in the normal group gives the same power as $n = 252$, assuming that this ratio is 1. Using formula (8), we compute the proportion of controls in the entire sample (n_0/n): if $A = 1$ we have $n_0/n = 0.46$ and under the optimal design, when $A = 0.343$ we have $n_0/n = 0.69$. Thus, under the optimal design, there should be 124 controls and 56 cases in the entire sample of 180 participants.

4. SUMMARY AND DISCUSSION

Power function is the probability of rejecting the null hypothesis at the alternative. Therefore, power should be computed using the same statistic used to test the null. Since the Z-test is commonly used for coefficient significance testing, the same test should be used for power analysis and sample size determination. However, when the likelihood-ratio test is planned for use, the same test statistic should be used as the basis for power computation.

The Wald-based sample size and power calculations for logistic regression with binary interaction can be carried out online at www.dartmouth.edu/~eugened.

It is a mystery why the null variance formula is used in all papers on Wald-based sample size determination. One explanation is that it was initially borrowed from the test on proportion, $H_0 : p = p_0$. But in coefficient significance testing, such as in model (4), $H_0 : \eta = 0$, the variance is evaluated at the MLE, not at $\eta = 0$. Although the two formulas are close for small log ORs, they

DOI: 10.1002/sim

may lead to a considerable difference otherwise. Sometimes the difference in the tests is down played, suggesting that they are equivalent in large sample. Although this statement is true at the null, it is not true at the alternative. Consequently, different tests yield different sample sizes even asymptotically.

Since specification of the parameter values is never rigorous, it is a good practice to compute a sample size under different scenarios. Closed-form formulas presented in this work become very helpful to carry out the respective sensitivity analysis.

The power computation and sample size formula are applicable to case–control studies as well. For a case–control study, we find an optimal proportion that maximizes power and, respectively, minimizes the total sample size. We need to be careful when specifying the proportion of cases in the total sample or normal group. When the alternative OR is close to 1, the 50/50 design is close to optimal. But when OR becomes extreme, as may happen in gene–gene or gene–environment studies, we may substantially reduce the total number of subjects and yet have the same power. In our example, the optimal design reduces the number of subjects by 30 per cent, with the same power.

APPENDIX A

A.1. Information matrix

Letting $\mathbf{x}=(1, x, z, xz)'$, the 4×4 information matrix has the form

$$\mathbf{I}=E_{(x,z)}\left(\frac{e^{\alpha_0+\beta x+\gamma z+\eta(xz)}}{[1+e^{\alpha_0+\beta x+\gamma z+\eta(xz)}]^2}\mathbf{x}\mathbf{x}'\right)$$

$$=E_{(x,z)}\left(\frac{e^{\alpha_0+\beta x+\gamma z+\eta(xz)}}{[1+e^{\alpha_0+\beta x+\gamma z+\eta(xz)}]^2}\begin{bmatrix}1 & x & z & xz\\ x & x & xz & xz\\ z & xz & z & xz\\ xz & xz & xz & xz\end{bmatrix}\right)$$

where subindex (x, z) means that the expectation is taken over the joint distribution of x and z. The right-hand side of this expression follows from the identities $x^2=x$ and $z^2=z$. Since x and z take value 0 or 1, we express the information matrix as

$$\frac{e^{\alpha_0}}{[1+e^{\alpha_0}]^2}\mathbf{M}_1\Pr(x=0, z=0)+\frac{e^{\alpha_0+\beta}}{[1+e^{\alpha_0+\beta}]^2}\mathbf{M}_2\Pr(x=1, z=0)$$

$$+\frac{e^{\alpha_0+\gamma}}{[1+e^{\alpha_0+\gamma z}]^2}\mathbf{M}_3\Pr(x=0, z=1)+\frac{e^{\alpha_0+\beta+z+\eta}}{[1+e^{\alpha_0+\beta+\gamma+\eta}]^2}\mathbf{M}_4\Pr(x=1, z=1)$$

where

$$\mathbf{M}_1=\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix},\quad \mathbf{M}_2=\begin{bmatrix}1 & 1 & 0 & 0\\ 1 & 1 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$$

DOI: 10.1002/sim

$$\mathbf{M}_3 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{M}_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

To compute the expectation, we need the following joint probabilities:

$$p_{00} = \Pr(x=0, z=0) = \frac{1-p_z}{1+C}, \quad p_{10} = \Pr(x=1, z=0) = \frac{C(1-p_z)}{1+C}$$

$$p_{01} = \Pr(x=0, z=1) = \frac{p_z}{1+CD}, \quad p_{11} = \Pr(x=1, z=1) = \frac{CDp_z}{1+CD}$$

In notation (7), we write

$$\mathbf{I} = \begin{bmatrix} L+F+J+R & F+R & J+R & R \\ F+R & F+R & R & R \\ J+R & R & J+R & R \\ R & R & R & R \end{bmatrix}$$

After some algebra, we obtain the inverse matrix

$$\mathbf{I}^{-1} = \begin{bmatrix} \frac{1}{L} & -\frac{1}{L} & -\frac{1}{L} & \frac{1}{L} \\ -\frac{1}{L} & \frac{1}{L}+\frac{1}{F} & \frac{1}{L} & -\frac{1}{L}-\frac{1}{F} \\ -\frac{1}{L} & \frac{1}{L} & \frac{1}{L}+\frac{1}{J} & -\frac{1}{L}-\frac{1}{J} \\ \frac{1}{L} & -\frac{1}{L}-\frac{1}{F} & -\frac{1}{L}-\frac{1}{J} & \frac{1}{F}+\frac{1}{J}+\frac{1}{L}+\frac{1}{R} \end{bmatrix}$$

which is the square root of n times the covariance matrix of the MLE of $(\alpha_0, \beta, \gamma, \eta)$. The last element of this matrix is the asymptotic variance of the interaction coefficient estimate, $\sqrt{n}\widehat{\eta}_{\mathrm{ML}}$.

A.2. Optimal case–control ratio

We wish to find A that gives the minimum of V. Substituting formulas (7) into (6), we deduce that V is proportional to

$$\frac{[(1+A)^2BC + (1+AB)^2]\omega + [(1+ABGK)^2 + (1+AG)^2BCDK]\phi}{ABCDGKp_z(1-p_z)}$$

We notice that minimization of V can be reduced to minimization of $(a_2A^2 + a_1A + a_0)A^{-1}$, where a_0, a_1, and a_2 are coefficients depending on other specification parameters. But the minimum of the latter function is attained at $A = \sqrt{a_0/a_2}$, which gives formula (9).

DOI: 10.1002/sim

ACKNOWLEDGEMENTS

I am thankful to Sergey Demidenko, who helped me with programming and webpage design for sample size and power calculations. The author is also grateful for the reviewer's comments that helped to improve the paper.

REFERENCES

1. Rosner B. *Fundamentals of Biostatistics* (6th edn). Duxbury: Pacific Grove, 2005.
2. Gardiner J, Pathak D, Indurkhya A. Power calculations for detecting interaction in stratified 2×2 tables. *Statistics and Probability Letters* 1999; **41**(3):267–275.
3. Smith PG, Day NE. The design of case–control studies: the influence of confounding and interaction effects. *International Journal of Epidemiology* 1984; **13**(3):356–365.
4. Hwang S-J, Beaty TH, Liang K-L, Coresh J, Khoury MJ. Minimum sample size estimation to detect gene–environment interaction in case–control designs. *American Journal of Epidemiology* 1994; **140**(11):1029–1037.
5. Yang Q, Khoury MJ, Friedman JM, Flanders WD. On the use of population attributable fraction to determine sample size for case–control studies of gene–environment interaction. *Epidemiology* 2003; **14**(2):161–167.
6. Whittemore AS. Sample size for logistic regression with small response probability. *Journal of the American Statistical Association* 1981; **76**(373):27–32.
7. Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine* 1998; **17**(14):1623–1634.
8. Shieh G. On power and sample size calculations for likelihood-ratio tests in generalized linear models. *Biometrics* 2000; **56**(4):1192–1196.
9. Foppa I, Spiegelman D. Power, sample size calculations for case–control studies of gene–environment interactions with polytomous exposure variable. *American Journal of Epidemiology* 1997; **146**(7):596–604.
10. Lubin J, Gail M. On power and sample size calculations for studying features of the relative odds of disease. *American Journal of Epidemiology* 1990; **131**(3):552–566.
11. Self SG, Mauritsen RH. Power/sample size calculations for generalized linear models. *Biometrics* 1988; **44**(1): 79–86.
12. Shieh G. A comparison of two approaches for power and sample size calculations in logistic regression models. *Communications in Statistics—Statistical Simulations* 2000; **29**(3):763–791.
13. Gauderman WJ. Sample size requirements for association studies of gene–gene interaction. *American Journal of Epidemiology* 2002; **155**(5):478–484.
14. Gauderman WJ. Sample size requirements for matched case–control studies of gene–environment interaction. *Statistics in Medicine* 2002; **21**(1):35–50.
15. Demidenko E. Sample size determination for logistic regression revisited. *Statistics in Medicine* 2007; **26**(18): 3385–3397.
16. Machin D, Campbell MJ. *Design of Studies for Medical Research*. Wiley: New York, 2005.
17. Prentice RL, Pyke R. Logistic disease incidence models and case–control studies. *Biometrika* 1979; **66**(3):403–411.
18. Gilliland F, McConnell R, Peters J, Gong J. A theoretical basis for investigating ambient air pollution and children's respiratory health. *Environmental Health Prospective* 1999; **107**(3):403–407.
19. Brittain E, Schlesselman JJ. Optimal allocation for the comparison of proportions. *Biometrics* 1982; **38**(4): 1003–1009.

DOI: 10.1002/sim