European Journal of Epidemiology (2005) 20: 389–393
DOI 10.1007/s10654-005-2018-3

A method for indirectly estimating gene-environment effect modification and power given only genotype frequency and odds ratio of environmental exposure

Jimmy Thomas Efird
John A. Burns School of Medicine, University of Hawai'i at Mānoa, 1960 East-West Road, Room T201b, Honolulu, HI 96822-2319, USA

Accepted in revised form 1 February 2005

Abstract. Both genes and environment are important determinants of disease. In this paper we model gene-environment effect modification on the odds ratio scale OR(GE|D) and show how to indirectly estimate the effect and 95% confidence intervals (CI) for the simple case of no main genetic and environmental effects [i.e., $OR(G\bar{E}|D) = OR(\bar{G}E|D) = 1$]. A statistic is presented to test the null hypothesis OR(GE|D) = 1 and to calculate corresponding power, given the odds ratio for environmental exposure OR(E|D) and population genotype frequency (g). Direct extension of the above model provides a mathematical framework for estimating confidence bounds in more complex cases involving partial genetic and/or environmental effects.

Key words: Gene-environment effect estimation, Genotype frequency, Odds ratio

Abbreviations: CI = confidence interval; G6PD = glucose-6-phosphate dehydrogenase; LCI = 95% lower confidence interval; OR = odds ratio; PKU = phenylketonuria

Introduction

Epidemiologists long have recognized that for some diseases, the presence of a certain genotype or environmental exposure alone may not lead to the development of disease [1] but that disease can result from their combined effect. The classic example relates to the association between phenylketonuria (PKU) genotype and dietary intake of phenylalanine in the case of mental retardation [1]. Here, neither presence of the phenylketonuria genotype nor consumption of phenylalanine alone affects the risk of developing disease. Yet, when present in combination, risk is increased. Similarly, neither glucose-6-phosphate dehydrogenase (G6PD) deficiency nor fava bean consumption alone influences the development of severe hemolytic anemia. However, the disease may develop when both factors are present [2]. Erroneous interpretation concerning the role of environmental or genetic factors in disease etiology may occur when failing to account for gene-environment effect modification [1, 3, 4].

Below we derive a method to statistically test whether gene-environment effect modification on the odds ratio (OR) scale exists in the simple case when no main genetic and environmental effects are present. This method is based only on the OR of exposure and the population genotype (e.g., no individual genotype data are required from the case-control study). Furthermore, no knowledge of two-by-two cell-counts is necessary to compute a *z*-statistic or study power.

Methodology

The OR for environmental exposure associated with disease in the population is defined as

$$OR(E|D) = \frac{P(E|D)/P(\bar{E}|D)}{P(E|\bar{D})/P(\bar{E}|\bar{D})}. \quad (1)$$

Noting that the OR for environmental exposure and disease are equivalent [5] and assuming that disease is rare in both the exposed and unexposed populations [6], we see Equation (1) simplifies to the expression for relative risk (RR) of disease, i.e.,

$$= \frac{P(D|E)}{P(D|\bar{E})}. \quad (2)$$

Applying the chain rule of probability [i.e., $P(A|B) = P(A|CB)P(C|B) + P(A|\bar{C}B)P(\bar{C}|B)$] [1] and assuming that genotype (g) is independent of environmental exposure [i.e., $P(G|E) = P(G|\bar{E}) = P(G) = g$], Equation (2) may be rewritten as

$$= \frac{P(D|GE)g + P(D|\bar{G}E)(1-g)}{P(D|G\bar{E})g + P(D|\bar{G}\bar{E})(1-g)} \quad (3)$$

$$= \frac{[P(D|GE)g/P(D|\bar{G}\bar{E})] + [P(D|\bar{G}E)(1-g)/P(D|\bar{G}\bar{E})]}{[P(D|G\bar{E})g/P(D|\bar{G}\bar{E})] + (1-g)}. \quad (4)$$

Again invoking the rare disease assumption and interchanging disease and exposure ORs, Equation (4) becomes

$$= \frac{OR(GE|D)g + OR(\bar{G}E|D)(1-g)}{OR(G\bar{E}|D)g + (1-g)}, \quad (5)$$

where $OR(GE|D) = \{P(GE|D)/P(\bar{G}\bar{E}|D)\}/\{P(GE|\bar{D})/P(\bar{G}\bar{E}|\bar{D})\}$ and denotes the OR of a gene-environment effect given disease. Similarly, $OR(\bar{G}E|D)$ refers to the OR of an environmental effect in the absence of a genetic effect and $OR(G\bar{E}|D)$ the OR of a genetic effect in the absence of an environmental effect. The referent group in each case is the absence of a genetic and environmental effect (i.e., $\bar{G}\bar{E}$). Rearranging Equation (5), a general expression for the OR of a gene-environment effect is given as

$$OR(GE|D) = \{OR(E|D)[OR(G\bar{E}|D)g + (1-g)] - OR(\bar{G}E|D)(1-g)\} \quad (6)$$

Considering the simple case when $OR(G\bar{E}|D) = OR(\bar{G}E|D) = 1$, we see that

$$OR(GE|D) = [OR(E|D) - 1 + g]/g, \quad (7)$$

where $[|OR(E|D)| - 1] \geq g]$ per the above unity constraints on the joint conditional probabilities. A 95% normal theory confidence interval (CI) [7, 8] for Equation (7) is given as

$$\exp\left\{ \log\{[OR(E|D) - 1 + g]/g\} \pm 1.96 \cdot \sqrt{var\{\log\{[OR(E|D) - 1 + g]/g\}}\right\} \quad (8)$$

Treating (g) as fixed and using a first-order Taylor series expansion and the Cramer–Rao lower bounds, we see that [7, 9]

$$Var\left\{ \log\{[OR(E|D) - 1 + g]/g\}\right\} \quad (9)$$

$$\cong \left[\frac{d\log\{[OR(E|D) - 1 + g]/g\}}{d[OR(E|D)]}\right]^2 \cdot Var[OR(E|D)] \quad (10)$$

$$\cong \left\{OR(E|D)^2 \cdot \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)\right\} / \{OR(E|D) - 1 + g\}^2, \quad (11)$$

whereVar[OR(E|D)] was estimated using the method of Fleiss [10] and a is the number of diseased individuals with environmental exposure, b the number of non-diseased individuals with environmental exposure, c the number of diseased individuals without environmental exposure, and d is the number of non-diseased individuals without environmental exposure.

Substituting the results of Equation (11) into Equation (8), an estimate for the 95% CI for OR(GE|D) may be written as

$$\exp\left\{ \log\{[OR(E|D) - 1 + g]/g\}. \pm 1.96 \cdot \frac{OR(E|D)\sqrt{\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)}}{\{OR(E|D) - 1 + g\}}\right\}. \quad (12)$$

In situations where (g) cannot be considered fixed, the "Delta" method [11] may be easily extended to estimate the variance.

Next, we express the 95% CI for OR(GE|D) in terms of the 95% lower confidence limit (LCL) for OR(E|D). By definition [8], the 95% LCL for OR(E|D) is given as

$$LCL = OR(E|D) \cdot \exp\left\{-1.96 \cdot \sqrt{\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)}\right\}. \quad (13)$$

$$\Rightarrow \log(LCL) = \log[OR(E|D)] - 1.96\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \quad (14)$$

$$\Rightarrow \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} = \{\log[OR(E|D)] - \log(LCL)\}/1.96. \quad (15)$$

Substituting Equation (15) into Equation (12), the 95% CI for OR(GE|D) may be rewritten as

$$\exp\left\{ \log\{[OR(E|D) - 1 + g]/g\} \pm \frac{OR(E|D) \cdot \{\log[OR(E|D)] - \log(LCL)\}}{\{OR(E|D) - 1 + g\}}\right\}. \quad (16)$$

Further, if the pivotal statistic [7]

$$Z = \left| \frac{\{OR(E|D) - 1 + g\} \cdot \log\{[OR(E|D) - 1 + g]/g\}}{OR(E|D)\sqrt{\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)}} \right|, \quad (17)$$

where $Z_{\alpha/2}$ denotes the critical region of the standard normal distribution, then the null hypothesis that $OR(GE|D) = 1$ may be rejected at the α-level of significance. The power corresponding to this test statistic [5] is given as

$$= \left| \frac{(z_{\alpha/2}) \cdot \{OR(E|D) - 1 + g\} \cdot \log\{[OR(E|D) - 1 + g]/g\}}{OR(E|D) \cdot \{\log[OR(E|D)] - \log(LCL)\}} \right| \geq z_{\alpha/2}, \tag{18}$$

$$Z_{\text{power}} = Z - Z_{\alpha/2}. \tag{19}$$

Example

A study in a hypothetical population is being planned to "directly" estimate the OR for effect modification between arsenic exposure (E) and the genotype for slow acetylation (G), given that the study participants have been diagnosed with skin cancer (D). Suppose that a recent retrospective study of 250 cases and 250 controls in the same population as the proposed study observed a 1.7-fold OR(E|D) [95% CI: 1.1, 2.9] for skin cancer among participants exposed to arsenic compared with those not exposed. Although participants were not genotyped for slow acetylation in the latter study, other published reports have noted that the genotype frequency for slow acetylation (g) in this population is approximately 70%. Using the above OR, 95% CI, and (g) values in Equations (7) and (16) and assuming that neither arsenic exposure nor the presence of the slow acetylator polymorphism alone increase the OR for skin cancer, we "indirectly" estimate $OR(GE|D) = 2.0$ (95% CI: 1.2, 3.4). The null hypothesis $OR(GE|D) = 1$ is rejected at the $\alpha = 0.01$ level of significance. However, the new case-control study based upon the above parameters and sample size only would have power = 72.9% to detect $OR(GE|D) \leq 2.0$ at the $\alpha = 0.05$ level of significance, indicating that a larger sample size would be needed to be powered at the standard level of 80% (e.g., a false null finding in the new case control study could be due to a small sample size). A sample of approximately 345 cases and 345 controls would be required in the new study to achieve ≥80% power, given that all other parameters are held constant.

Sensitivity analysis

The sensitivity to the assumption of no main effects is shown in Table 1, given specific values of $OR(E|D)$, $OR(G\bar{E}|D), OR(\bar{G}E|D)$, and (g). Partial genetic and/or environmental effects can greatly influence $OR(GE|D)$. For example, when $OR(E|D) = 3.0$, $OR(G\bar{E}|D) = 2, OR(\bar{G}E|D) = 2.0$, and $(g) = 0.01$, $OR(GE|D) = 105$ versus 201 in the absence of genetic and environmental main effects. Further, above unity ORs for gene-environment effect modification may be present even though a null OR for environmental exposure is observed in a case-control study [e.g., $OR(E|D) = 1.0, OR(G\bar{E}|D) = 5.0$, $OR(\bar{G}E|D) = 2.0, (g) = 0.50, OR(GE|D = 4.0]$.

Discussion

Failing to account for gene-environment effect modification could conceal effects of genotype on risk of disease and lead to misleading interpretation of study results [1, 12, 13]. Inconsistent associations across studies between a disease and a suspected risk factor could be due to heterogeneity in the studied population with respect to unknown gene-environment effect modification [14]. The presence of gene-environment effect modification also may affect study power and indicate that a larger sample size is required when planning a new study.

Several limitations should be considered when indirectly estimating gene-environment effect modification. Bias of an unknown magnitude may result if the rare disease assumption is violated. However, bias of this type is common to all cumulative incidence case-control studies when disease risk is high between both the exposed and/or unexposed populations [6] When environmental exposure is influenced by genetically controlled behavioral determinants, as may be true in the case of alcohol consumption, it is conceivable that genotype may not be independent of environmental exposure, i.e., $P(G|E) \neq P(G|\bar{E})$. Only environmental factors that are well established as independent of underlying genetic factors should be considered when indirectly estimating gene-environment effect modification. Further, "indirect" estimation of gene-environment effect modification is only meaningful in the simple case presented if a true joint "biologic" effect exists between (G) and (E), and no main genetic and environmental effects are present. Nonetheless, this model emphases the broader importance of gene-environment effect modification in the interpretation of study results.

In contemporary research, with considerable advances in molecular genetic techniques, the method presented here provides a framework to evaluate the joint genetic effect for individual reaction to pharmaceutical agents, response to medical treatment, and susceptibility to other environmental factors (e.g., physical, chemical, biologic) [1].

Acknowledgements

The author thanks Dr Elizabeth A. Holly and Paige M. Bracci (Department of Epidemiology and Biostatistics, UCSF School of Medicine), and Drs Lorene M. Nelson and Kristin Cobb (Division of

Table 1. Odds ratio for gene-environment effect modification given specific values for OR(E|D), OR($G\bar{E}|D$), OR($\bar{G}E|D$) and (g). OR(GE|D)[a]

OR($G\bar{E}|D$)	1				2				3				5			
OR($\bar{G}E|D$)	1	2	3	5	1	2	3	5	1	2	3	5	1	2	3	5
OR(E|D) = 1																
(g)																
0.01	1.0	–[b]	–[b]	–[b]	2.0	–[b]	–[b]	–[b]	3.0	–[b]	–[b]	–[b]	5.0	–[b]	–[b]	–[b]
0.05	1.0	–[b]	–[b]	–[b]	2.0	–[b]	–[b]	–[b]	3.0	–[b]	–[b]	–[b]	5.0	–[b]	–[b]	–[b]
0.10	1.0	–[b]	–[b]	–[b]	2.0	–[b]	–[b]	–[b]	3.0	–[b]	–[b]	–[b]	5.0	–[b]	–[b]	–[b]
0.25	1.0	–[b]	–[b]	–[b]	2.0	–[b]	–[b]	–[b]	3.0	–[b]	–[b]	–[b]	5.0	–[b]	–[b]	–[b]
0.50	1.0	–[b]	–[b]	–[b]	2.0	1.0	–[b]	–[b]	3.0	2.0	1.0	–[b]	5.0	4.0	3.0	1.0
0.75	1.0	0.67	0.33	–[b]	2.0	1.7	1.3	0.67	3.0	2.7	2.3	1.7	5.0	4.7	4.3	3.7
0.90	1.0	0.89	0.78	0.56	2.0	1.9	1.8	1.6	3.0	2.9	2.8	2.6	5.0	4.9	4.8	4.6
0.95	1.0	0.95	0.89	0.79	2.0	1.9	1.9	1.8	3.0	2.9	2.9	2.8	5.0	4.9	4.9	4.8
OR(E|D) = 2																
(g)																
0.01	101	2.0	–[b]	–[b]	103	4.0	–[b]	–[b]	105	6.0	–[b]	–[b]	109	10	–[b]	–[b]
0.05	21	2.0	–[b]	–[b]	23	4.0	–[b]	–[b]	25	6.0	–[b]	–[b]	29	10	–[b]	–[b]
0.10	11	2.0	–[b]	–[b]	13	4.0	–[b]	–[b]	15	6.0	–[b]	–[b]	19	10	–[b]	–[b]
0.25	5.0	2.0	–[b]	–[b]	7.0	4.0	1.0	–[b]	9.0	6.0	3.0	–[b]	13	10	7.0	1.0
0.50	3.0	2.0	1.0	–[b]	5.0	4.0	3.0	1.0	7.0	6.0	5.0	3.0	11	10	9.0	7.0
0.75	2.3	2.0	1.7	1.0	4.3	4.0	3.7	3.0	6.3	6.0	5.7	5.0	10	10	9.7	9.0
0.90	2.1	2.0	1.9	1.7	4.1	4.0	3.9	3.7	6.1	6.0	5.9	5.7	10	10	9.9	9.7
0.95	2.1	2.0	1.9	1.8	4.1	4.0	3.9	3.8	6.1	6.0	5.9	5.8	10	10	9.9	9.8
OR(E|D) = 3																
(g)																
0.01	201	102	3.0	–[b]	204	105	6.0	–[b]	207	108	9.0	–[b]	213	114	15	–[b]
0.05	41	22	3.0	–[b]	44	25	6.0	–[b]	47	28	9.0	–[b]	53	34	15	–[b]
0.10	21	12	3.0	–[b]	24	15	6.0	–[b]	27	18	9.0	–[b]	33	24	15	–[b]
0.25	9.0	6.0	3.0	–[b]	12	9.0	6.0	–[b]	15	12	9.0	3.0	21	18	15	9.0
0.50	5.0	4.0	3.0	1.0	8.0	7.0	6.0	4.0	11	10	9.0	7.0	17	16	15	13
0.75	3.7	3.3	3.0	2.3	6.7	6.3	6.0	5.3	9.7	9.3	9.0	8.3	16	15	15	14
0.90	3.2	3.1	3.0	2.8	6.2	6.1	6.0	5.8	9.2	9.1	9.0	8.8	15	15	15	15
0.95	3.1	3.1	3.0	2.9	6.1	6.1	6.0	5.9	9.1	9.1	9.0	8.9	15	15	15	15
OR(E|D) = 5																
(g)																
0.01	401	302	203	5.0	406	307	208	10	411	312	213	15	421	322	223	25
0.05	81	62	43	5.0	86	67	48	10	91	72	53	15	101	82	63	25
0.10	41	32	23	5.0	46	37	28	10	51	42	33	15	61	52	43	25
0.25	17	14	11	5.0	22	19	16	10	27	24	21	15	37	34	31	25
0.50	9.0	8.0	7.0	5.0	14	13	12	10	19	18	17	15	29	28	27	25
0.75	6.3	6.0	5.7	5.0	11	11	11	10	16	16	16	15	26	26	26	25
0.90	5.4	5.3	5.2	5.0	10	10	10	10	15	15	15	15	25	25	25	25
0.95	5.2	5.2	5.1	5.0	10	10	10	10	15	15	15	15	25	25	25	25

[a] $OR(GE|D) = \left\{OR(E|D)[OR(G\bar{E}|D)g + (1-g)] - OR(\bar{G}E|D)(1-g)\right\}$.
[b] out of limits.

Epidemiology, Stanford School of Medicine) for reviewing the original manuscript and providing useful comments and suggestions. Thanks also to Alice S. Whittemore (Division of Epidemiology, Stanford School of Medicine) for introducing the author to the basic mathematical concepts underlying gene-environment analysis. This work is supported by NIH Grant Number G12RR003061 from the National Center for Research Resources.

References

1. Khoury M, Adams M, Flanders W. An epidemiologic approach to ecogenetics. Am J Hum Genet 1988; 42: 89–95.
2. Beutler E. Glucose-6-phosphate dehydrogenase deficiency. In: Stanbury J, Wyngaarden J, Fredrickson D, Goldstein J, Brown M (eds): The metabolic basis of inherited disease, 5th edn. New York: McGraw-Hill, 1983, p. 1629–1653.

3. Ottman R. An epidemiologic approach to gene-environment interaction. Genet Epidemiol 1990; 7: 177–185.
4. Khoury M, Stewart W, Beaty T. The effect of genetic susceptibility on causal inference in epidemiologic studies. Am J Epidemiol 1987; 126: 561–567.
5. Kelsey J, Whittemore A, Evans A, Thompson W. methods in observational epidemiology. New York: Oxford University Press, 1996.
6. Cornfield J. A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. J Natl Cancer Inst 1951; 11: 1269–1275.
7. Mood A, Graybill F, Boes D. Introduction to the theory of statistics. New York: McGraw-Hill, 1974.
8. Kahn H, Sempos C. Statistical methods in epidemiology. New York: Oxford University Press, 1989.
9. Armitage P. Statistical methods in medical research. New York: Wiley, 1971.
10. Fleiss J. Statistical methods for rates and Proportions. New York: John Wiley & Sons, 1981.
11. Cox C. Delta Method. In: Armitage P, Colton T (eds): Encyclopedia of biostatistics. Chichester, UK: Wiley, 1998, p. 1125–1127.
12. Hwang S, Beaty T, Liang K, Coresh J, Khoury M. Minimum sample size estimation to detect gene-environment interaction in case-control designs. Am J Epidemiol 1994; 140: 1029–1037.
13. 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.
14. Andrieu N, Goldstein A. Epidemiologic and genetic approaches in the study of gene-environment interaction: an overview of available methods. Epidemiol Rev 1998; 20: 137–147.

Address for correspondence: Dr Jimmy Thomas Efird, John A. Burus School of Medicine, 1960 East-West Road, Room T201b, University of Haiwai'i at Mānoa, Honolulu, HI 96822-2319, USA.
E-mail: Jimmy.efird@stanfordalummni.org