

Statistical Analysis of Gene-Environment Data
 Jimmy Thomas Efird, Ph.D
 Department of Epidemiology and Biostatistics, UCSF School of Medicine

ABSTRACT

The objective of this paper is to assess the effect of gene-environment interaction on the odds ratio of environmental exposure given disease in the population [OR(E|D)], when neither a gene nor environmental effect is present alone. A nomogram is provided for determining OR(E|D) for different values of gene frequency and gene-environment effect given disease [OR(GE|D)].

INTRODUCTION

Both genes and environment are important determinants of disease. However, above unity odds ratios (OR) for environmental exposure given disease may be observed when neither a gene nor environment effect are present alone. Below we illustrate the effect of gene-environment interaction on the exposure OR given the above scenerio.

METHODOLOGY

Assuming that genotype is independent of environmental exposure, e.g.,

$P(G|E)=P(G|\bar{E})=P(G)=g$, and applying the chain rule of probability, the OR of environmental exposure associated with disease in the population may be expressed as

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

$$= \frac{\left[\frac{P(D|GE)g + P(D|\bar{G}\bar{E})(1-g)}{P(\bar{D}|GE)g + P(\bar{D}|\bar{G}\bar{E})(1-g)} \right]}{\left[\frac{P(D|G\bar{E})g + P(D|\bar{G}E)(1-g)}{P(\bar{D}|G\bar{E})g + P(\bar{D}|\bar{G}E)(1-g)} \right]} \quad (2)$$

If neither a gene nor environment effect exists alone [e.g., $OR(G\bar{E}|D)=OR(\bar{G}E|D)=1$] and the risk of disease is sufficiently small over the period of observation (e.g., <0.05 for the exposed and unexposed joint gene-environment effect),¹ the above expression approximately reduces to

$$\left[\frac{P(GE|D)/P(\bar{G}\bar{E}|D)g}{P(G\bar{E}|D)/P(\bar{G}E|D)} \right] + 1 - g \quad (3)$$

$$= OR(GE|D)g + 1 - g, \quad (4)$$

where $OR(GE|D)$ denotes the OR of a gene-environment effect given disease.

Using Woolf's approximation² and treating g as fixed, the 95% confidence interval (CI) for expression (4) is given as

$$e^{\log[OR(GE|D)g + 1 - g] \pm 1.96g \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}}, \quad (5)$$

where

a =number of diseased individuals with positive genotype and environmental exposure

b =number of non-diseased individuals with positive genotype and environmental exposure

c =number of diseased individuals with negative genotype and environmental exposure

d =number of non-diseased individuals with negative genotype and environmental exposure.

Further, if

$$Z = \frac{\log[OR(GE|D)g + 1 - g]}{g \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}} \geq Z_{\alpha/2} \quad (6)$$

where $Z_{\alpha/2}$ denotes the critical region of the standard normal distribution, the null hypothesis that $OR(E|D)=1$ may be rejected at the α -level of significance. The relationship between the 95% CI for expression (4) and the hypothesis test described in

expression (6) is evident upon replacing the denominator of expression (6) with $-\frac{[\text{Log}(\text{LCI})-\text{Log}\{\text{OR}(\text{GEID})g+1-g\}]}{1.96}$, where LCI denotes the 95% lower confidence for OR(EID). Rearranging the above terms, it follows that

$$\text{OR}(\text{GEID}) = \frac{\text{OR}(\text{EID}) - 1 + g}{g}, \quad (7)$$

where the 95% CI for expression (7) is given as

$$\frac{\text{OR}(\text{EID}) - 1 + g}{g} e^{\pm \frac{[\text{log}\{\text{OR}(\text{EID}) - \text{log}(\text{LCI})\}]}{g}} \quad (8)$$

EXAMPLE

Suppose that the odds ratio for heterocyclic amine exposure given colon cancer is 3.4 (95% CI=1.3-9.2) and that the genotype for slow acetylation occurs in 75% of a known population. Assuming that neither heterocyclic amine exposure nor the presence of the slow acetylator polymorphism alone increases the estimated risk for cancer, we see that $\text{OR}(\text{GEID})=4.2$, 95% CI=1.2-15.1.

RESULTS

The effect of gene-environment interaction on the odds ratio for disease is illustrated in Figure 1. For a fixed genotype frequency, the OR of environmental exposure given disease increases proportionally to the OR of a gene-environment effect for disease in the population. Further, the overall magnitude of the effect increases as gene frequency increases.

DISCUSSION

Epidemiologist and genetists have long recognized that for some diseases, the presence of a certain genotype or environmental exposure alone may not lead to the development of disease.³ The classic example relates to the interaction between phenylketonuria genotype and dietary intake of phenylalanine in the case of mental retardation.³ 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 elevated. 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 risk factors are present.⁴ Under circumstances as just described, erroneous inference concerning the role of environmental or genetic factors in disease etiology

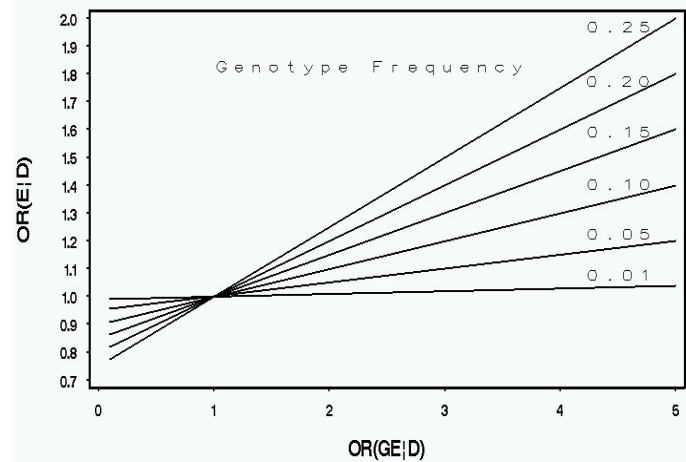
may occur when failing to account for gene-environment interaction effects.^{3,5,6}

By statistically modelling gene-environment interaction, we have shown that the odds ratio of environmental exposure given disease is a function of genotype frequency and the odds ratio of the joint gene-environmental effect given disease. These analyses provide a framework for evaluating the genetic difference between individuals in their reaction to pharmaceutical agents, response to medical treatment, and susceptibility to other environmental factors (e.g., physical, chemical, biologic).³

ACKNOWLEDGEMENTS

The author especially thanks Dr. Lorene M. Nelson (Division of Epidemiology, Stanford School of Medicine) for introducing him to the methodology of gene-environmental interaction analysis while a student at Stanford School of Medicine. Thanks also to Paige M. Bacci (Department of Epidemiology and Biostatistics, UCSF School of Medicine), Dr. Kristin Cobb (Division of Epidemiology, Stanford School of Medicine), and Dr. Dora Il'yasova (Department of Cancer Biology, Wake Forest University School of Medicine) for reviewing the manuscript and providing useful comments and suggestions.

Figure 1: Gene-environment Interaction



REFERENCES

1. 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.
2. Woolf B. On estimating the relation between blood group and disease. *Ann Hum Genet* 1955;**19**:251-253.
3. Khoury M, Adams M, Flanders W. An epidemiologic approach to ecogenetics. *Am J Hum Genet* 1988;**42**:89-95.
4. 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 edition. New York: McGraw-Hill, 1983:1629-1653.
5. Ottman R. An epidemiologic approach to gene-environment interaction. *Genet Epidemiol* 1990;**7**:177-185.
6. Khoury M, Stewart W, Beaty T. The effect of genetic susceptibility on causal inference in epidemiologic studies. *Am J Epidemiology* 1987;**126**:561-567.