American Journal of Epidemiology
Vol. 153, No. 10
Printed in U.S.A.

PRACTICE OF EPIDEMIOLOGY

Commentary: Facing the Challenge of Gene-Environment Interaction: The Two-by-Four Table and Beyond

Lorenzo D. Botto[1] and Muin J. Khoury[2]

As a result of the Human Genome Project, epidemiologists can study thousands of genes and their interaction with the environment. The challenge is how to best present and analyze such studies of multiple genetic and environmental factors. The authors suggest emphasizing the fundamental core of gene-environment interaction—the separate assessment of the effects of individual and joint risk factors. In the simple analysis of one genotype and an exposure (both dichotomous), such study can be summarized in a two-by-four table. The advantages of such a table for data presentation and analysis are many: The table displays the data efficiently and highlights sample size issues; it allows for evaluation of the independent and joint roles of genotype and exposure on disease risk; and it emphasizes effect estimation over model testing. Researchers can easily estimate relative risks and attributable fractions and test different models of interaction. The two-by-four table is a useful tool for presenting, analyzing, and synthesizing data on gene-environment interaction. To highlight the role of gene-environment interaction in disease causation, the authors propose that the two-by-four table is the fundamental unit of epidemiologic analysis. *Am J Epidemiol* 2001;153:1016–20.

environment; epidemiologic methods; genes; Human Genome Project

Almost all human diseases result from gene-environment interaction. Proving, documenting, and quantifying this statement is a long-sought goal of the scientific community and one that, if achieved, could provide fundamental insights into the causes, courses, and prevention of many conditions. Knowing what genes to assess has been a major challenge—a challenge that the explosive growth of genetic technology is rapidly overcoming. As a result of the Human Genome Project, the sequences of thousands of genes are already available, and the complete catalogue of human genes is within reach (1, 2). Thus, although the development of environmental biomarkers has been less spectacular, scientists already can study diseases in relation to multiple genes and their variants. For example, the risk for venous thrombosis is being studied in relation to variants of the factor V, prothrombin, and 5,10-methylenetetrahydrofolate reductase genes, as well as to blood homocysteine levels and oral contraceptive use (3–6). Similarly, the risk for spina bifida is being studied in relation to variants of folate-related genes (e.g., 5,10-methylenetetrahydrofolate reductase, cystathione-beta-synthase, methionine synthase, and methionine synthase reductase) and blood levels of selected vitamins (folate, B_{12}) (7–10). We can safely predict that such studies of multiple genetic factors will increase in the near future.

From an epidemiologic perspective, however, this bounty of risk factor information presents a major challenge: how to best present and analyze studies of multiple genetic and environmental factors. Epidemiologists have long grappled with this issue, usually in relation to the concept of interaction (see Greenland and Rothman (11) for a summary), but the literature reveals no consistent approach.

In this commentary, we suggest emphasizing the fundamental core of gene-environment interaction—the separate assessment of the effects of individual and joint risk factors. Such an approach has many practical advantages that we illustrate with some simple study scenarios.

THE TWO-BY-FOUR TABLE: THE FUNDAMENTAL UNIT OF ANALYSIS OF GENE- ENVIRONMENT INTERACTION

The simplest case of gene-environment interaction is that of two dichotomous factors (e.g., presence or absence of a genotype and presence or absence of an environmental risk factor). In the case of a single biallelic gene, it can be argued that the genetic exposure has inherently three, rather than two, levels (zero, one, or two alleles). In some such cases, it may be worthwhile to show the full data. In many cases, however, it is useful to think more generally of presence or

Received for publication May 19, 2000, and accepted for publication August 23, 2000.

[1] Birth Defects and Genetic Diseases Branch, National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention, Atlanta, GA.

[2] Office of Genetics and Disease Prevention, National Center for Environmental Health, Centers for Disease Control and Prevention, Atlanta, GA.

Reprint requests to Dr. Lorenzo D. Botto, Birth Defects and Pediatric Genetics Branch, National Center for Environmental Health, Mailstop F-45, Centers for Disease Control and Prevention, 4770 Buford Highway NE, Atlanta, GA 30341.

TABLE 1. Layout for a case-control study assessing the effect of a genotype and an environmental factor

G*	E*	Cases	Controls	Odds ratio		Contrast	Main information
+	+	a	b	ah/bg	A	A vs. D	Joint genotype and environmental factor vs. none
+	−	c	d	ch/dg	B	B vs. D	Genotype alone vs. none
−	+	e	f	eh/fg	C	C vs. D	Environmental factor alone vs. none
−	−	g	h	1	D		Common reference

Other measures	Odds ratio	Main information
Case only odds ratio	ag/ce	Departure from multiplicative model of interaction
Control only odds ratio	bh/df	Independence of factors in the population
Multiplicative interaction	$A/(B \times C)$	Deviation from multiplicative model of interaction
Additive interaction	$A - (B + C - 1)$	Deviation from additive model of interaction
Stratified 1a	ad/bc	Association with environmental factor among people *with* the genotype
Stratified 1b	eh/fg	Association with environmental factor among people *without* the genotype
Stratified 2a	af/be	Association with genotype among people *exposed* to environmental factor
Stratified 2b	ch/dg	Association with genotype among people *not exposed* to environmental factor

* G, genotype; E, environmental factor.

absence of a genotype (or a set of genotypes). Such a genotype (or set of genotypes) could include, in principle, combinations of alleles at multiple loci. For example, in a dominant system in whom carriers of one or two copies of a specific allele may have the same genetic susceptibility, the genotype could be dichotomized as with or without the high-risk allele.

For illustration, we present scenarios in a case-control setting, in which we assume the ideal conditions of an unbiased, unconfounded, incident-case study. We will further assume that the odds ratios in the study are valid estimations of relative risks.

A case-control study of a genotype and an exposure (both dichotomous) can be presented in a two-by-four table (table 1). The same reference group is used to compute three odds ratios, those for each factor alone and those for the combination of genotype and exposure. Such odds ratios are the basic, direct measures of association.

Such presentation has several advantages (table 2). The role of each factor is independently assessed both in terms of odds ratios and of attributable fraction. The odds ratios can be combined to assess departures from specified models of interactions (e.g., multiplicative or additive). Table 2 also provides the distribution of the exposures among controls and helps evaluate the independence of the distribution of the genetic and environmental factors in the underlying population. Finally, a case-only odds ratio can be easily derived and used as a comparison with findings from case-only studies in the literature.

The two-by-four table approach to presenting gene-environment interactions is appealing for several reasons. First, it is efficient: It summarizes, without loss of detail, seven two-by-two tables and generates a comprehensive set of effect estimates that none of the latter, individually, can match. Second, the table highlights potential sample size issues: Cell sizes are directly presented, and confidence intervals show their effect on statistical power. Third, the two-by-four table approach emphasizes effect estimation over model testing: The relative risk estimates associated with the joint and individual exposures are the primary elements of an interaction, whereas departures from specific models of interactions are derived parameters and are explicitly labeled as such. Finally, because most, if not all, human diseases result from the joint effect of genes and the environment, it can be argued that the two-by-four table—and not the two-by-two table—is the fundamental unit of epidemiologic analysis.

TABLE 2. Advantages of the two-by-four table in the study of gene-environment interactions

Advantages of the 2×4 table
1. The primary data are displayed clearly and completely.
2. The primary measures of association—relative risk estimates for each factor alone and for the joint exposure—are readily generated. Because they use the same reference group, these estimates can be compared.
3. Attributable fractions can be computed separately for each exposure alone and for the joint exposure.
4. Relative risk estimates can be used to assess the relation between the joint exposure and the individual exposures. For example, the departure from additive or multiplicative models of interactions can be readily derived from the table.
5. Risk estimates stratified by either exposure can also be calculated if needed.
6. For case-control studies, the case-only and control-only odds ratios can be computed easily. For adequately chosen control groups, the control-only odds ratio estimates exposure dependencies in the underlying population.

TABLE 3. Analysis of oral contraceptive use, presence of factor V Leiden allele, and risk for venous thromboembolism*

Factor V†	OC†	Cases	Controls		OR†	95% CI†	AF-Exp† (%)	AF-Pop† (%)	Exposure frequency in controls (%)
+	+	25	2	ORge†	34.7	7.83, 310.0	97.1	15.7	1.2
+	–	10	4	ORg†	6.9	1.83, 31.80	85.6	5.5	2.4
–	+	84	63	ORe†	3.7	1.28, 6.32	73.0	39.6	37.3
–	–	36	100	Reference	Reference				59.2
Total		155	169						

	Expected ORge	Departure from expected
Additive	3.7 + 6.9 – 1 = 9.6	34.7 – 9.6 = 25.07
Multiplicative	3.7 × 6.9 = 25.7	34.7/25.7 = 1.4

* Modified from Vandenbroucke et al. Lancet 1994;344:1453–7. The departure of the observed from the expected effect of the joint exposure depends on the definition of no interaction, as shown below for simple additive and multiplicative definitions, where G = genotype and E = environmental factor.

† Factor V: +, presence of factor V Leiden allele (heterozygotes and homozygotes), –, absence of factor V Leiden allele; OC, oral contraceptive: +, current use of oral contraceptives; –, no current use of oral contraceptives; OR, odds ratio; CI, confidence interval; AF-Exp (%), attributable fraction (percent) among exposed cases; AF-Pop (%), attributable fraction (percent) among all cases in the population; ORge, odds ratio for disease among oral contraceptive users with the susceptibility genotype; ORg, odds ratio for disease among nonusers with the susceptibility genotype; ORe, odds ratio for disease among oral contraceptive users without the susceptibility genotype.

THE TWO-BY-FOUR TABLE: A SIMPLE APPLICATION

To illustrate the value of the two-by-four table approach with two dichotomous factors, we used a case-control study of venous thromboembolism in relation to factor V Leiden and oral contraceptive use (12). When so rearranged (table 3), the data show clearly the odds ratios for factor V Leiden and oral contraceptive use alone (6.9 and 3.7, respectively) and for their combination (34.7). In addition to these odds ratios, as their relevance to possible causation, table 3 shows information of public health interest, such as the relatively high frequency of these exposures (e.g., 2.4 percent for factor V Leiden and 1.2 percent for the joint exposure) and their attributable fractions (5.5 and 15.7 percent, respectively).

In contrast, classic stratified analysis, in which the association between the oral contraceptive use and venous thrombosis is assessed separately among those with and those without the factor V Leiden polymorphism, does not easily provide such primary information and emphasizes departure from a specified (multiplicative) model of interaction (table 4). When it is of interest, interaction models can be tested by using the data from the two-by-four table, but such testing is not restricted to multiplicative models, and the model itself is labeled (table 3). Finally, the relation of the risk factors within the case and the control groups can be assessed separately (table 4). In the first case, the relation can be summarized as a case-only odds ratio, as in a case-only study. The case-only design is an efficient and valid approach to screening for gene-environment interaction, under the assumption of independence of exposure and genotype in the population (13, 14), and the role of such studies in the epidemiologic approach to complex diseases has been reviewed (15, 16). In

TABLE 4. Comparing the stratified and case-only approaches with the 2 × 4 approach*

		Factor V present		Factor V absent		Ratio of odds ratios
		No. of cases	No. of controls	No. of cases	No. of controls	
Analysis stratified on factor 5						
Oral contraceptive use	+	25	2	84	63	
	–	10	4	36	100	
OR† (95% CI†)		5.0 (0.8, 31.8)		3.7 (2.2, 6.1)		1.4
Case-only and control-only odds ratios						
Case-only odds ratio		(25 × 36)/(10 × 84) = 1.1				
Control-only odds ratio		(2 × 100)/(4 × 63) = 0.8				1.4

* The data are taken from table 3. Note that ratios of odds ratios are identical to departure from the multiplicative model (table 3).

† OR, odds ratio; CI, confidence interval.

the second case, the association of risk factors among controls provides important information on potential dependencies of the risk factors in the underlying population. Taken together, these two odds ratios can be used to compare results of case-only studies in the literature and to verify the validity of their assumptions.

BEYOND THE TWO-BY-FOUR TABLE

The two-by-four table, although simple, may adequately summarize some, but probably not all, epidemiologic relations. First, there may be more than two relevant factors. Because the number of exposure combinations grows quickly (2^n for n dichotomous factors), the corresponding table rapidly becomes unwieldy. However, in some cases, it may be possible to revert to simpler tables by selecting appropriate contrasts of genotypes (or sets of genotypes) and environmental exposures. Second, the relation between exposure and outcome can be other than dichotomous, i.e., more generally graded or continuous (dose-response), as with smoking and lung cancer or obesity and hypertension. Dichotomizing exposure and disease has worked remarkably well for epidemiologists; how often dichotomizing genotypes will provide a reasonable summary of their effect is much less clear, particularly for common gene variants and common, complex diseases. In the general case of n exposures, each with its dose-response curve, the response surface is best described as an n-dimensional manifold. Third, as more factors are involved, the complexity of the interactions may be such that they cannot be adequately described by simple multiplicative or additive models.

These limitations also highlight two issues that will increasingly tax epidemiologists as they try to unravel the web of gene-environment interactions. First, new or improved epidemiologic methods may be needed to deal with such complex situations. For example, researchers have suggested utilizing regression models and neural networks, traditionally used in modeling the probability of clinical outcomes (17, 18), to study of gene-environment interactions (19–21). So far, these approaches have limitations: The output of regression models, for example, is model dependent; neural networks, although, in general, less dependent on prior model specification (19–21), may be limited in their ability to explicitly estimate dependencies among risk factors (19, 20).

The second issue relates to sample size. As the number of factors under study increases so do the strata that have to be defined within the study. At typical sample sizes, increasing the number of factors quickly reduces per-stratum size (table 2, control group), reducing statistical power. Thus, negative findings should be interpreted carefully. Scientists should consider conducting collaborative studies to increase sample size and power while striving to decrease or control for extraneous genetic heterogeneity.

Finally, the two-by-four table approach highlights and may in part resolve the issue of terminology. Historically, the assessment of multiple factors has been closely associated with the concept of interaction, which in itself has been controversial (11). In part, the debate was generated by the use of the term "interaction" for different concepts; for example, biologic, public health, and statistical concepts of interaction have been distinguished. The two-by-four table approach illustrates a possibility to avoid these terms, as true interactions will have a biologic basis, should be expressed also in public health terms, and can be subject to statistical evaluation. In their place, epidemiologists can simultaneously present and clearly label the odds ratios, attributable fractions, and excess case load, as well as statistical tests of explicitly defined interaction models.

In conclusion, researchers are challenged to apply epidemiologic methods to increasingly complex data on gene-environment interaction. Carefully conducted collaborative studies may provide adequate sample size. A clear presentation and analysis of the core elements of these interactions (the data distribution and the primary measures of association) may increase the information that can be extracted from the data. One should not be misled into believing that all gene-environment interaction can be immediately reduced to a simple eightfold table. Nevertheless, we propose that the two-by-four table and its immediate extensions are fundamental tools to documenting and studying gene-environment interaction.

REFERENCES

1. Hamosh A, Scott AF, Amberger J, et al. Online Mendelian inheritance in man (OMIM). Hum Mutat 2000;15:57–61.
2. Collins FS, Patrinos A, Jordan E, et al. New goals for the U.S. Human Genome Project: 1998–2003. Science 1998;282:682–9.
3. Gerhardt A, Scharf RE, Beckmann MW, et al. Prothrombin and factor V mutations in women with a history of thrombosis during pregnancy and the puerperium. N Engl J Med 2000;342: 374–80.
4. Akar N, Akar E, Akcay R, et al. Effect of methylenetetrahydrofolate reductase 677 C-T, 1298 A-C, and 1317 T-C on factor V 1691 mutation in Turkish deep vein thrombosis patients. Thromb Res 2000;97:163–7.
5. Martinelli I, Taioli E, Bucciarelli P, et al. Interaction between the G20210A mutation of the prothrombin gene and oral contraceptive use in deep vein thrombosis. Arterioscler Thromb Vasc Biol 1999;19:700–3.
6. Cattaneo M, Chantarangkul V, Taioli E, et al. The G20210A mutation of the prothrombin gene in patients with previous first episodes of deep-vein thrombosis: prevalence and association with factor V G1691A, methylenetetrahydrofolate reductase C677T and plasma prothrombin levels. Thromb Res 1999; 93:1–8.
7. Botto LD, Yang Q. 5,10-Methylenetetrahydrofolate reductase gene variants and congenital anomalies: a HuGE review. Am J Epidemiol 2000;151:862–77.
8. Christensen B, Arbour L, Tran P, et al. Genetic polymorphisms in methylenetetrahydrofolate reductase and methionine synthase, folate levels in red blood cells, and risk of neural tube defects. Am J Med Genet 1999;84:151–7.
9. Shaw GM, Rozen R, Finnell RH, et al. Maternal vitamin use, genetic variation of infant methylenetetrahydrofolate reductase, and risk for spina bifida. Am J Epidemiol 1998;148:30–7.
10. Wilson A, Platt R, Wu Q, et al. A common variant in methionine synthase reductase combined with low cobalamin (vitamin B12) increases risk for spina bifida. Mol Genet Metab 1999; 67:317–23.
11. Greenland S, Rothman KJ. Concepts of interaction. In: Greenland S, Rothman KJ, eds. Modern epidemiology. Philadelphia, PA: Lippincott, Williams & Wilkins, 1998:329–42.

12. Vandenbroucke JP, Koster T, Briet E, et al. Increased risk of venous thrombosis in oral-contraceptive users who are carriers of factor V Leiden mutation. Lancet 1994;344:1453–7.
13. Begg CB, Zhang ZF. Statistical analysis of molecular epidemiology studies employing case-series. Cancer Epidemiol Biomarkers Prev 1994;3:173–5.
14. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med 1994; 13:153–62.
15. Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls. Am J Epidemiol 1996; 144:207–13.
16. Yang Q, Khoury MJ. Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research. Epidemiol Rev 1997;19:33–43.
17. Ioannidis JPA, McQueen PG, Goedert JJ, et al. Use of neural networks to model complex immunogenetic associations of disease: human leukocyte antigen impact on the progression of human immunodeficiency virus infection. Am J Epidemiol 1998;147:464–71.
18. Marchevsky AM, Patel S, Wiley KJ, et al. Artificial neural networks and logistic regression as tools for prediction of survival in patients with stages I and II non-small cell lung cancer. Mod Pathol 1998;11:618–25.
19. Duh M-S, Walker AM, Ayanian JZ. Epidemiologic interpretation of artificial neural networks. Am J Epidemiol 1998;147: 1112–22.
20. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol 1996;49:1225–31.
21. Warner B, Misra M. Understanding neural networks as statistical tools. Am Stat 1996;50:284–93.