American Journal of Epidemiology
Published by the Johns Hopkins Bloomberg School of Public Health 2008.

Vol. 169, No. 2
DOI: 10.1093/aje/kwn352
Advance Access publication November 20, 2008

Invited Commentary

Invited Commentary: Efficient Testing of Gene-Environment Interaction

Nilanjan Chatterjee and Sholom Wacholder

Initially submitted May 30, 2008; accepted for publication August 26, 2008.

Gene-environment-wide interaction studies of disease occurrence in human populations may be able to exploit the same agnostic approach to interrogating the human genome used by genome-wide association studies. The authors discuss 2 methods for taking advantage of possible independence between a single nucleotide polymorphism they call G (a genetic factor) and an environmental factor they call E while maintaining nominal type I error in studying G-E interaction when information on many genes is available. The first method is a simple 2-step procedure for testing the null hypothesis of no multiplicative interaction against the alternative hypothesis of a multiplicative interaction between an E and at least one of the markers genotyped in a genome-wide association study. The added power for the method derives from a clever work-around of a multiple testing procedure. The second is an empirical-Bayes–style shrinkage estimation framework for G-E interaction and the associated tests that can gain efficiency and power when the G-E independence assumption is met for most G's in the underlying population and yet, unlike the case-only method, is resistant to increased type I error when the underlying assumption of independence is violated. The development of new approaches to testing for interaction is an example of methodological progress leading to practical advantages.

association; environment; genes; genetic markers; genetics; genome

Abbreviations: E, environmental factor; G, genetic factor.

In this issue of the *Journal*, Murcray et al. (1) present a new approach to evaluating multiplicative gene-environment interaction in the context of a genome-wide association study, where there are M single nucleotide polymorphisms and a single E, the environmental factor under consideration. They propose a simple 2-step procedure for the null hypothesis of no multiplicative interaction against the alternative hypothesis of a multiplicative interaction. In step 1, they propose an α-level test for association in the 2×2 table of a single nucleotide polymorphism we call G (a genetic factor) crossed with E among cases and controls combined. If the P value for the test is above some $\alpha_1 < \alpha$, then the null hypothesis is accepted. Otherwise, in step 2, the P value from the standard test of multiplicative interaction between G and E in the $2 \times 2 \times 2$ table of disease status $\times$ G $\times$ E is compared with α/m, where m is the number of tests not accepted in step 1: if the P value is above α/m, then the hypothesis of no interaction is accepted; otherwise, the hypothesis of no interaction is rejected. When the standard assumptions hold, the independence of the 2 test statistics guarantees that only α/m among the m hypotheses are rejected, giving the desired property that only proportion α of all hypotheses is rejected, regardless of α_1.

Either greater p_{GE}, the fraction of G's associated with E, or increasing α_1 leads to an increase in m, the number of G's that reach step 2, and reduced power of the new method. Why does the power advantage diminish as m increases? As noted by the authors (1), the added power for the method derives from a clever work-around of a multiple testing procedure. The power of a standard analysis of a case-control study is calculated at an α level of α/M; The power of the Murcray et al. procedure is calculated at an α level of α/m in step 2 and therefore increases with decreasing m. Although this procedure guarantees that the family-wise error rate is at or below α, as does Bonferroni correction under the usual assumptions, it is not as conservative as Bonferroni adjustment at the level of each individual hypothesis. With Bonferroni adjustment, each of M hypotheses is of statistical

Correspondence to Dr. Nilanjan Chatterjee, 6120 Executive Boulevard, EPS 8020, Rockville, MD 20852 (e-mail: chattern@mail.nih.gov).

size α/M. Instead, the Murcray et al. procedure allows hypotheses with extreme G-E association but no interaction to be of a size above α/M, whereas hypotheses when the G-E association is near 1 are of a size below α/M.

Murcray et al. (1) use the G-E independence assumption to construct an efficient screening test for interaction at step 1. Their simulation studies demonstrate that if the independence assumption is valid for a large fraction of G-E combinations (say, $p_{GE} \leq 5\%$) under study, then the proposed 2-step method can have a substantial power advantage over the standard 1-step case-control test for interaction that completely ignores the natural G-E independence assumption. Thus, the method has increased power, yet retains the conservatism of the genome-wide significance level, which is needed to keep the chance of a false-positive finding low when the prior probabilities of each hypothesis are very low, as they will be with an agnostic approach (2).

The power advantage of the 2-step procedure over the standard 1-step method diminishes as M, the total number of markers, increases, everything else being equal. In particular, a 2-step procedure in a genome-wide association study with 500,000 single nucleotide polymorphisms and the standard α level of 10^{-7} for genome-wide significance would require, on average, an α level for the second step of 2×10^{-6} if $\alpha_1 = 0.05$ at the first step. Thus, the power of the 2-step procedure, which is bounded above by the power of the test used at the second step, would be only slightly higher than that of a 1-step method. In contrast, if one starts with a much smaller number of single nucleotide polymorphisms, say $M = 5{,}000$, the power gain attributable to the reduction in the number of tests due to the screening procedure at the first step using an α level of 0.05 indeed could be substantial, as demonstrated by the authors (1).

Recently, Mukherjee and Chatterjee (3) proposed a novel approach to ''1-step'' inference of gene-environment interaction by using an empirical Bayes–type shrinkage estimation framework. Their estimator is a weighted average of the case-only and case-control estimators of the logarithm of the interaction. The weights are based on the variance of the robust case-control estimate and the difference between the 2 observed estimates, which reflect the dependence between G and E among controls; note that in the $2 \times 2 \times 2$ table, the ratio of the interaction estimates is simply the G-by-E odds ratio in the controls (4). When the estimates from the standard and case-only estimates are similar, the empirical Bayes estimator puts more weight on the efficient case-only estimate, which is not robust to departure from G-E independence (5). As the difference between estimates increases, the estimator gives increasingly more weight to the case-control estimate, which is robust to departure from G-E independence. The weight for the standard case-control estimate also increases as its precision relative to the case-only estimate increases. Such empirical Bayes–type estimators, and the associated tests, can gain efficiency and power when the G-E independence assumption is met for most G's in the underlying population and yet, unlike the case-only method (4), are resistant to increased type I error when the underlying assumption of independence is violated (6).

The method proposed by Murcray et al. (1) also exploits the G-E independence assumption, but only through a first-step ''screening'' procedure that reduces the number of tests to be conducted at the second step. The empirical Bayes procedure, in contrast, gains efficiency by directly exploiting the likely independence assumption for the actual test for interaction. It will be interesting to compare the performance of the 2-step and empirical Bayes procedures under different scenarios of the distribution of G-E association; although they both exploit the independence assumption, they gain efficiency in very different ways. The performance comparison can now be informed by Davey Smith et al.'s (7) recent empirical study of the associations between pairs of 23 genetic variants and 96 nongenetic characteristics. Although no greater association than expected by chance was found in their study, further empirical studies of G-E association, particularly between the variants and environmental exposures important for G-E interaction, will be helpful in evaluating methods whose performance depends on G-E independence, such as Murcray et al.'s (1) and others' (3, 4, 6, 8, 9).

The development of new approaches to test for interaction is an example of methodological progress leading to practical advantages. The accompanying commentary by Khoury and Wacholder (10) shows how many more examples we need in the field of gene-environment interaction.

ACKNOWLEDGMENTS

Authors affiliation: Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institute of Health, Department of Health and Human Services, Bethesda, Maryland (Nilanjan Chatterjee, Sholom Wacholder).

This research was supported by a Gene-Environment Initiative (GEI) from the National Heart Lung and Blood Institute (R01 HL091172-01) and the Intramural Research Program of the National Institutes of Health, National Cancer Institute, Division of Cancer Epidemiology and Genetics.

The findings in this paper reflect the viewpoints of the authors and do not necessarily reflect the views of the Department of Health and Human Services.

Conflict of interest: none declared.

REFERENCES

1. Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. *Am J Epidemiol*. 2009;169(2):219–226.
2. Wacholder S, Chanock S, Garcia-Closas M, et al. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst*. 2004; 96(6):434–442.
3. Mukherjee B, Chatterjee N. Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*. 2008;64(3):685–694.
4. 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(2):153–162.

5. Albert PS, Ratnasinghe D, Tangrea J, et al. Limitations of the case-only design for identifying gene-environment interactions. *Am J Epidemiol*. 2001;154(8):687–693.
6. Mukherjee B, Ahn J, Gruber SB, et al. Tests for gene-environment interaction from case-control data: a novel study of type I error, power, and designs. *Genet Epidemiol*. 2008;32(7):615–626.
7. Davey Smith G, Lawlor DA, Harbord R, et al. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology [electronic article]. *PLoS Med*. 2007;4(12):e352.
8. Modan B, Hartge P, Hirsh-Yechezkel G, et al. Parity, oral contraceptives, and the risk of ovarian cancer among carriers and noncarriers of a *BRCA1* or *BRCA2* mutation. *N Engl J Med*. 2001;345(4):235–240.
9. Spinka C, Carroll RJ, Chatterjee N. Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genet Epidemiol*. 2005;29(2):108–127.
10. Khoury MJ, Wacholder S. Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies—challenges and opportunities. *Am J Epidemiol*. 2009;169(2):227–230.