Short Communication

Data-Dredging Gene-Dose Analyses in Association Studies: Biases and their Corrections

Wen-Chung Lee and Hsiao-Yuan Huang

Graduate Institute of Epidemiology, College of Public Health, National Taiwan University, Taipei, Taiwan

Abstract

To examine the joint effect of multiple loci on disease risk, many case-control association studies used ''gene-dose analyses.'' However, some researchers defined high-risk genotypes (or alleles) as those that have higher genotypic (allelic) frequencies in the case group compared with the control group in the study. This will lead to the total number of the ''high-risk'' genotypes (alleles) tending to be higher for the cases than for the controls as well, even if none of the studied loci were related to the disease. Monte-Carlo simulations done in this study showed that such a ''data-dredging'' gene-dose analysis could produce grossly biased results. A permutation correction method was proposed which could correct the biases very effectively. (Cancer Epidemiol Biomarkers Prev 2005;14(12):3004–6)

Introduction

To examine the joint effect of multiple loci on disease risk, many case-control association studies used ''gene-dose analyses'' (we found 40 articles with gene-dose analysis from January 1998 to May 2005 in six esteemed cancer journals: *Cancer Epidemiology Biomarkers and Prevention, Cancer Research, Carcinogenesis, Clinical Cancer Research, International Journal of Cancer*, and the *Journal of the National Cancer Institute*). In these studies, the researchers first defined the high-risk genotype (or allele) at each locus. They then calculated the total number of the high-risk genotypes (alleles) a subject has for the cases and for the controls in their study. They categorized the total number of the high-risk genotypes (alleles) into a number of categories, say, ''0 to 2'', ''3 to 4'', and ''5+'', and they calculated the odds ratios (OR) and the associated confidence intervals (CI) for these categories. Finally, they did a trend test to see if the OR increases as the total number of the high-risk genotypes (alleles) increases. ''Dose-response relationship'' is one of the causal criteria put forward by Hill (1). A gene-dose effect is corroborative of a cause-and-effect relation between the genes and the disease.

However, some researchers defined the high-risk genotypes (alleles) as those that have higher genotypic (allelic) frequencies in the case group than in the control group in the study, or equivalently, as those with OR > 1 in the study [among the aforementioned 40 articles, researchers from 10 of these articles defined their high-risk genotypes (alleles) in this way; see Table 1 for a summary]. They then did the usual gene-dose analysis as if those defined are the bona fide high-risk genotypes (alleles)—the ''data-dredging'' gene-dose analysis. The definition guarantees the frequency of the ''high-risk'' genotypes (alleles) being higher in the case group than in the control group at each locus. Intuitively, this will lead to the total number of the high-risk genotypes (alleles) tending to be higher for the cases than for the controls as well, even if, in fact, none of the studied loci were related to the disease. Consequently, an excess of false-positive gene-dose effects may result.

Methods and Results

A simulation study is presented in Table 2. The loci are assumed to be unlinked or in linkage equilibrium with one another. We see that although none of the simulated loci is related to the disease, the OR increases as the total number of the high-risk genotypes increases, displaying a false sense of gene-dose effects. The coverage probabilities of the CIs are too low and the type I error rates of the trend test are inflated ($\alpha = 0.05$). As the number of the studied loci increases, the problem becomes more acute; with 10 loci, the OR can be as high as $\sim$2.00, the coverage probability, as low as $\sim$0.38, and the type I error rate, as high as $\sim$0.64 (the data-dredging articles in Table 1 with $\geq$5 loci and borderline level of significance are thus particularly vulnerable). The overestimation of the OR is less severe when sample sizes reached 1,000 as compared with studies with 500 samples. However, the undercoverage of the CIs and the inflation of the type I error rates are irrespective of sample size. Similar biases can be found when the studied loci are in linkage disequilibrium (less severe) and when the analysis is based on alleles (more severe; results not shown).

To correct the biases, we let each and every study subject retain his/her genetic data. Whereas we randomly permute the disease status (cases or controls) of all the study subjects (keeping the total number of cases and the total number of controls constant). A new round of data-dredging gene-dose analyses is done for this permuted data—to determine a new set of the ''high-risk'' genotypes (alleles), and then to calculate a new set of the ORs and a new χ^2 statistic of the trend test. The procedure is to be repeated 1,000 times (from our limited simulation experiences, this number of permutations is sufficient).

To get a permutation-corrected OR for a particular category of the total number of high-risk genotypes (alleles), we divide the OR of that category in the original data by the geometric

Received 8/8/05; revised 9/29/05; accepted 10/10/05.

Grant support: Supported in part by the National Science Council, Republic of China.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Requests for reprints: Wen-Chung Lee, Graduate Institute of Epidemiology, College of Public Health, National Taiwan University, No. 1, Jen-Ai Road, 1st Section, Taipei, Taiwan. Fax: 886-2-2351-1955. E-mail: wenchung@ha.mc.ntu.edu.tw

doi:10.1158/1055-9965.EPI-05-0605

Table 1. A sample of data-dredging gene-dose analysis in the literature

Citations*	Genes studied	Number of loci	Unit of analysis	Findings
Chiu et al., CEBP 2002; 11:646–53	collagen-related genes	6	genotype	increased risk of oral submucous fibrosis with increasing number of high-risk genotypes ($P < 0.05$)
Cho et al., CEBP 2003; 12:1100–4	DNA repair genes	2	genotype	OR (for nasopharyngeal carcinoma) = 3.0 for subjects with two high-risk genotypes
Fu et al., Cancer Res 2003; 63:2440–6	DNA repair genes	5	genotype	a 1.46-fold increase in risk for breast cancer with one additional high-risk genotype
Gu et al., Clin Cancer Res 2005;11:1408–15	DNA repair genes	8	allele	increased risk of bladder cancer recurrence with increasing number of high-risk alleles ($P < 0.001$)
Han et al., Cancer Res 2004;64:3009–13	DNA repair genes	5	allele	decreased risk of nonmelanoma skin cancer with increasing number of low-risk alleles ($P < 0.05$)
Hu et al., Int J Cancer 2005;115:478–83	DNA repair genes	2	genotype	OR (for lung cancer) = 2.41 for subjects with two high-risk genotypes
Koh et al., Carcinogenesis 2005;26:459–64	angiotensin-related genes	3	genotype	decreased risk of breast cancer with increasing number of low-risk genotypes ($P = 0.05$)
Li et al., Carcinogenesis 2005;26:1596–602	cell-cycle control genes	2	allele	increased risk of squamous cell carcinoma with increasing number of high-risk alleles ($P = 0.001$)
Sturgis et al., Carcinogenesis 1999;20:2125–9	DNA repair genes	2	genotype	OR (for squamous cell carcinoma) = 2.02 for subjects with two high-risk genotypes
Zhang et al., CEBP 2005;14:1188–93	folate metabolism genes	2	allele	increased risk of squamous cell carcinoma with increasing number of high-risk alleles ($P = 0.002$)

*CEBP, Cancer Epidemiology Biomarkers and Prevention; Cancer Res, Cancer Research; Clin Cancer Res, Clinical Cancer Research; Int J Cancer, International Journal of Cancer.

mean of the ORs of the same category in the 1,000 permuted data (the original log OR minus the arithmetic mean of the permutation log ORs). To get a permutation-corrected upper (lower) limit of the 95% CI for a particular category, we divide the OR of that category in the original data by the 2.5th (97.5) percentile of the ORs of the same category in the 1,000 permuted data (the 95% CI constructed in this way will cover 1.00, if the original OR does not differ significantly from 1.00, using permutation tests at $\alpha = 0.05$; it will not cover 1.00, if the OR differs significantly from 1.00). To get a permutation-corrected P value for the trend test, we calculate the proportion in the 1,000 permuted data that have the χ^2 statistic of the trend test larger than the same statistic in the original data (alternatively, we declare the trend test to be significant at $\alpha = 0.05$, if the original χ^2 statistic is larger than the 95th percentile of the permutation χ^2 statistics).

With regard to the data in Table 2, it can now be seen in Table 3 that the permutation-corrected ORs are very close to their true values of 1.00, the permutation-corrected CIs have coverage probabilities of ~0.95, and the permutation-corrected trend tests have type I error rates of ~0.05. The permutation correction achieves similarly satisfactory results when the studied loci are in linkage disequilibrium and when the analysis is based on alleles (data not shown).

Discussion

In a study in which we have a priori knowledge about the high-risk genotypes (alleles) for some of the loci but not for the rest, we may resort to performing a gene-dose analysis with a mixture of the ''predetermined'' high-risks and the

Table 2. Data-dredging gene-dose analyses with different numbers of loci and different sample sizes

Number of the high-risk genotypes	Sample size (500)		Sample size (1,000)	
	OR*	Coverage probability of the CI	OR*	Coverage probability of the CI
Three loci				
Number of high-risk genotypes				
0 to 1	1.0000	—	1.0000	—
2	1.1999	0.8906	1.1361	0.8883
3	1.3924	0.8678	1.2641	0.8627
Type I error rate of the trend test	0.1605		0.1596	
Five loci				
Number of high-risk genotypes				
0 to 1	1.0000	—	1.0000	—
2 to 3	1.2820	0.8577	1.1943	0.8509
4 to 5	1.6485	0.7280	1.4211	0.7300
Type I error rate of the trend test	0.3037		0.2980	
Ten loci				
Number of the high-risk genotypes				
0 to 3	1.0000	—	1.0000	—
4 to 6	1.4152	0.7399	1.2764	0.7373
7 to 10	2.0005	0.3892	1.6334	0.3808
Type I error rate of the trend test	0.6348		0.6436	

NOTE: The results are based on simulated case-control data (equal number of cases and controls). The loci are unlinked or in linkage equilibrium with one another. The genotypes in each loci have been classified into the ''1- genotype'' and the ''0- genotype.'' The frequencies for the 1- genotype are: 0.4, 0.5, and 0.6 (three loci); 0.3, 0.4, 0.5, 0.6, 0.7 (five loci); 0.4 (×3), 0.5 (×4), 0.6 (×3) (10 loci), and are the same in the case group and the control group. Ten thousand simulations are done for each scenario. The α level is set at 0.05.
*Geometric mean of 10,000 simulations.

Table 3. Data-dredging gene-dose analyses, permutation-corrected, with different number of loci and different sample sizes

Number of high-risk genotypes	Sample size (500)		Sample size (1,000)	
	OR*	Coverage probability of the CI	OR*	Coverage probability of the CI
Three loci				
Number of high-risk genotypes				
0 to 1	1.0000	—	1.0000	—
2	1.0015	0.9487	1.0000	0.9454
3	1.0019	0.9528	1.0032	0.9519
Type I error rate of the trend test	0.0510		0.0531	
Five loci				
Number of high-risk genotypes				
0 to 1	1.0000	—	1.0000	—
2 to 3	0.9991	0.9459	1.0025	0.9506
4 to 5	1.0014	0.9478	1.0014	0.9532
Type I error rate of the trend test	0.0494		0.0456	
Ten loci				
Number of high-risk genotypes				
0 to 3	1.0000	—	1.0000	—
4 to 6	0.9996	0.9493	0.9997	0.9475
7 to 10	0.9982	0.9496	1.0020	0.9491
Type I error rate of the trend test	0.0488		0.0498	

NOTE: The results are based on 10,000 simulations, with 1,000 permutations for each round of the simulation. Other settings are the same as those in Table 2.
*Geometric mean of 10,000 simulations.

''data-dredged'' high-risks. The correction method can deal with this; by copying the predetermining and the data-dredging (of the original data) to each and every round of the permutations. The method also works when other covariates beyond the studied loci (i.e., sex, age, smoking, etc.) need to be considered; by performing a logistic regression to derive the ''covariates-adjusted'' ORs (2) in each and every round of the permutations. Finally, the method can also be used for a matched case-control study; the ORs are to be calculated using the matched-data methods (2) and the permutations (of the disease status) are to be restricted to subjects in the same matching set.

In conclusion, the data-dredging gene-dose analysis as commonly employed in the literature will produce grossly biased results. The biases, however, could very effectively be corrected by the permutation correction method as described in this report.

References

1. Hill AB. The environment and health: association or causation? Proc R Soc Med 1965;58:295–300.
2. Breslow NE, Day NE. Statistical methods in cancer research. Vol. I, The analysis of case-control studies. Lyon (France): IARC Scientific Publication, No. 32; 1980.

Data-Dredging Gene-Dose Analyses in Association Studies: Biases and their Corrections

Wen-Chung Lee and Hsiao-Yuan Huang

Cancer Epidemiol Biomarkers Prev 2005;14:3004-3006.