American Journal of Epidemiology

Vol. 182, No. 1
DOI: 10.1093/aje/kwv030
Advance Access publication:
May 29, 2015

Practice of Epidemiology

Testing for Sufficient-Cause Gene-Environment Interactions Under the Assumptions of Independence and Hardy-Weinberg Equilibrium

Wen-Chung Lee*

* Correspondence to Dr. Wen-Chung Lee, Taiwan Research Center for Genes, Environment and Human Health and Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, No. 17 Xuzhou Road, Room 536, Taipei 100, Taiwan (e-mail: wenchung@ntu.edu.tw).

Initially submitted April 16, 2014; accepted for publication January 26, 2015.

To detect gene-environment interactions, a logistic regression model is typically fitted to a set of case-control data, and the focus is on testing of the cross-product terms (gene × environment) in the model. A significant result is indicative of a gene-environment interaction under a multiplicative model for disease odds. Based on the sufficient-cause model for rates, in this paper we put forward a general approach to testing for sufficient-cause gene-environment interactions in case-control studies. The proposed tests can be tailored to detect a particular type of sufficient-cause gene-environment interaction with greater sensitivity. These tests include testing for autosomal dominant, autosomal recessive, and gene-dosage interactions. The tests can also detect trend interactions (e.g., a larger gene-environment interaction with a higher level of environmental exposure) and threshold interactions (e.g., gene-environment interaction occurs only when environmental exposure reaches a certain threshold level). Two assumptions are necessary for the validity of the tests: 1) the rare-disease assumption and 2) the no-redundancy assumption. Another 2 assumptions are optional but, if imposed correctly, can boost the statistical powers of the tests: 3) the gene-environment independence assumption and 4) the Hardy-Weinberg equilibrium assumption. SAS code (SAS Institute, Inc., Cary, North Carolina) for implementing the methods is provided.

case-control studies; epidemiologic methods; gene-environment interaction; Hardy-Weinberg equilibrium; sufficient-component-cause model

Abbreviations: AD, autosomal dominant; AN, all or none; AR, autosomal recessive; CN, controls; CS, cases; *CYP1A1*, cytochrome P-450, family 1, subfamily A, polypeptide 1 gene; GD, gene-dosage; GEI, gene-environment independence; HWE, Hardy-Weinberg equilibrium; IC, interaction contrast; L, leaky.

The occurrence of most human diseases is the result of interplay between genetic and environmental factors (1, 2). To detect gene-environment interactions, epidemiologists often adopt a case-control study design, recruiting diseased subjects (cases) and nondiseased subjects (controls) and comparing their genotypes and environmental exposures. A logistic regression model is typically fitted to the data, and the focus is on hypothesis testing of the cross-product terms (gene × environment) in the model (3). A significant result is indicative of a gene-environment interaction under a multiplicative model for disease odds. Oftentimes, it is reasonable to assume that the genes under examination are in Hardy-Weinberg equilibrium (HWE) and are also independent of any environmental exposure among the nondiseased subjects in the study population (4–9). A modified logistic regression methodology developed by Lee et al. (9) can exploit these 2 assumptions and achieve higher statistical powers for the *multiplicative* interaction tests.

Based on the sufficient-cause model for *risks*, VanderWeele and Robins (10, 11) constructed statistical tests for causal mechanistic interactions between binary variables. Basically, these are *additive* interaction tests, examining whether the observed disease risks deviate too much from additivity. VanderWeele (12) then went on to expand the methodologies for including categorical or ordinal variables with 3 or more levels. This later development is of particular relevance to the

current study of detecting gene-environment interactions, because genes are often coded as ternary variables with levels indicating 0, 1, or 2 variant alleles; the environmental exposures under study can also have multiple levels.

Based on the sufficient-cause model for *rates*, Lee (13, 14) constructed statistical tests for causal mechanistic interactions between binary variables and determined that a rate-model–based test has a less stringent threshold for detecting causal mechanistic interactions than a corresponding risk-model–based test. In this paper, I build upon previous work (13, 14) to propose a general hypothesis-testing framework for sufficient-cause gene-environment interactions. The proposed tests can be tailored to detect a particular type of sufficient-cause gene-environment interaction with greater sensitivity. These include testing for autosomal dominant, autosomal recessive, or gene-dosage interactions. The tests can also detect trend interactions (e.g., a larger gene-environment interaction with a higher level of environmental exposure) and threshold interactions (e.g., gene-environment interaction that occurs only when the environmental exposure reaches a certain threshold level). Two assumptions are necessary for the validity of these tests: 1) the rare-disease assumption and 2) the no-redundancy assumption (13–15). Another 2 assumptions are optional but, if imposed correctly, can boost the statistical powers of the tests: 3) the gene-environment independence (GEI) assumption and 4) the HWE assumption.

METHODS

Sufficient-cause model for rates

Let G ($G = 0$, 1, or 2) represent the number of variant alleles a subject carries and E ($E = 0$ or 1) represent a binary environmental exposure. Let $\text{Rate}_{g,e}$ ($\text{Odds}_{g,e}$) denote the disease rate (odds) for subjects with $G = g$ and $E = e$ in the study population. The rare-disease assumption is invoked, so disease odds (for 1 unit of follow-up time) and disease rates are equivalent (3).

The above variables (G and E) together define a total of 12 classes of sufficient causes (i.e., $(3 + 1) \times (2 + 1) = 12$), including 1 "all-unknown" class (U_1), 3 gene-only classes (U_2, U_3, U_4), 2 environment-only classes (U_5, U_6), and 6 gene-environment interaction classes ($U_7 \sim U_{12}$) (Figure 1). Note that here we do not impose the assumption of monotonicity (16–20) on the genetic effect, the environmental effect, or the gene-environment interaction effect, so this represents the most general sufficient-cause model for a ternary G and a binary E.

The sufficient-cause model is partly deterministic and partly stochastic. The presence of risk factor(s) alone is not sufficient for the disease. Only when all of the unknown components (complement causes) also appear can the sufficient cause become complete and the disease occur. Let $\text{Rate}_{U_1} \sim \text{Rate}_{U_{12}}$ denote the "completion rates" of the aforementioned 12 classes of sufficient causes, respectively. The completion rate for a particular class is the instantaneous arrival rate of the unknown complement causes in that class (20). Here the no-redundancy assumption is applied. This assumption posits that within a sufficiently short time interval, there can only be, at most, 1 arrival event of the unknown complement causes for each and every subject in the population (13–15).

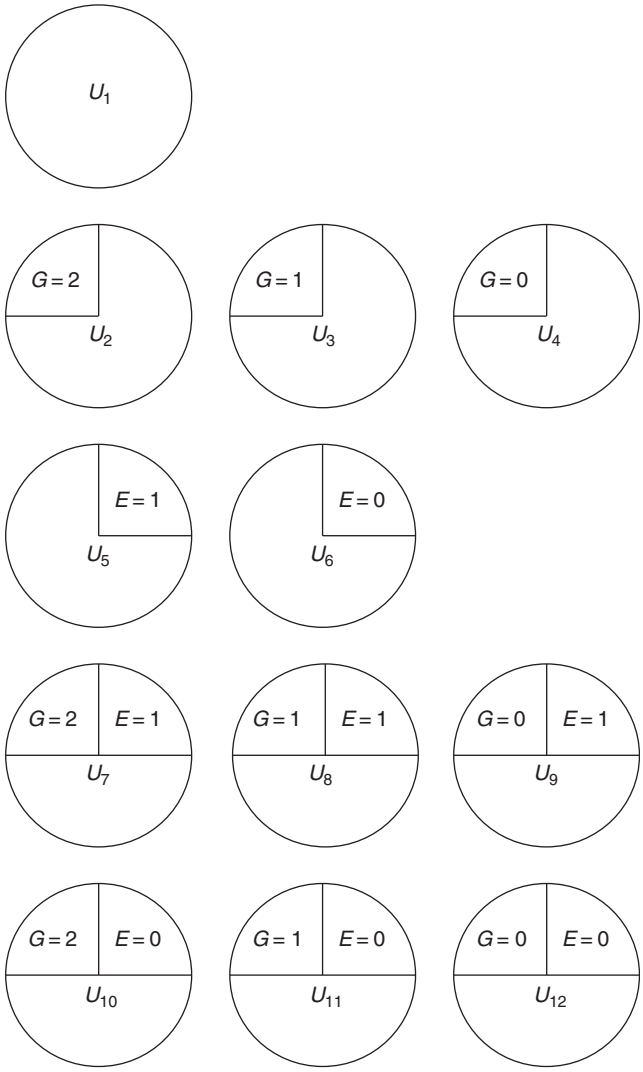

Figure 1. The 12 total classes of sufficient component causes for a ternary gene ($G = 0$, 1, or 2) and a binary environmental exposure ($E = 0$ or 1): the all-unknown class (U_1), the gene-only classes (U_2, U_3, U_4), the environment-only classes (U_5, U_6), and the gene-environment interaction classes ($U_7 \sim U_{12}$).

Under the no-redundancy assumption, the disease rate (and also the disease odds under the rare-disease assumption) for subjects with $G = g$ and $E = e$ is a simple arithmetic sum of the completion rates of the 4 "completable classes" (13, 14), that is,

$$\text{Odds}_{2,1} = \text{Rate}_{2,1} = \text{Rate}_{U_1} + \text{Rate}_{U_2} + \text{Rate}_{U_5} + \text{Rate}_{U_7},$$

$$\text{Odds}_{1,1} = \text{Rate}_{1,1} = \text{Rate}_{U_1} + \text{Rate}_{U_3} + \text{Rate}_{U_5} + \text{Rate}_{U_8},$$

$$\text{Odds}_{0,1} = \text{Rate}_{0,1} = \text{Rate}_{U_1} + \text{Rate}_{U_4} + \text{Rate}_{U_5} + \text{Rate}_{U_9},$$

$$\text{Odds}_{2,0} = \text{Rate}_{2,0} = \text{Rate}_{U_1} + \text{Rate}_{U_2} + \text{Rate}_{U_6} + \text{Rate}_{U_{10}},$$

$$\text{Odds}_{1,0} = \text{Rate}_{1,0} = \text{Rate}_{U_1} + \text{Rate}_{U_3} + \text{Rate}_{U_6} + \text{Rate}_{U_{11}},$$

and

$$\text{Odds}_{0,0} = \text{Rate}_{0,0} = \text{Rate}_{U_1} + \text{Rate}_{U_4} + \text{Rate}_{U_6} + \text{Rate}_{U_{12}},$$

respectively. In the case of subjects with $G = 2$ and $E = 1$, for example, they can develop the disease because of the completion of the U_1, U_2, U_5, or U_7 class; these 4 classes are their completable classes.

Testing for sufficient-cause gene-environment interactions

Now, consider the following interaction contrasts (ICs) which quantify departure from additivity:

$$\begin{aligned}\text{IC}_1 &= \text{Odds}_{1,1} - \text{Odds}_{0,1} - \text{Odds}_{1,0} + \text{Odds}_{0,0}\\ &= \text{Rate}_{U_8} - \text{Rate}_{U_9} - \text{Rate}_{U_{11}} + \text{Rate}_{U_{12}},\\ \text{IC}_2 &= \text{Odds}_{2,1} - \text{Odds}_{0,1} - \text{Odds}_{2,0} + \text{Odds}_{0,0}\\ &= \text{Rate}_{U_7} - \text{Rate}_{U_9} - \text{Rate}_{U_{10}} + \text{Rate}_{U_{12}},\end{aligned}$$

and

$$\begin{aligned}\text{IC}_3 &= w_1 \times \text{IC}_1 + w_2 \times \text{IC}_2\\ &= [w_2 \times \text{Odds}_{2,1} + w_1 \times \text{Odds}_{1,1} - (w_2 + w_1) \times \text{Odds}_{0,1}]\\ &\quad - [w_2 \times \text{Odds}_{2,0} + w_1 \times \text{Odds}_{1,0} - (w_2 + w_1) \times \text{Odds}_{0,0}]\\ &= [w_2 \times \text{Rate}_{U_7} + w_1 \times \text{Rate}_{U_8} - (w_2 + w_1) \times \text{Rate}_{U_9}]\\ &\quad - [w_2 \times \text{Rate}_{U_{10}} + w_1 \times \text{Rate}_{U_{11}} - (w_2 + w_1) \times \text{Rate}_{U_{12}}],\end{aligned}$$

with arbitrary w_1 and w_2. These ICs are linear combinations of the disease odds, and under the rare-disease and no-redundancy assumptions, they are equal to linear combinations of the completion rates of the interaction classes ($U_7 \sim U_{12}$).

The coefficients in ICs can alternatively be expressed as the products of contrast coefficients for the gene (G) and the environment (E), respectively:

$$\text{IC} = \sum_{i=1}^{3}\sum_{j=1}^{2} c_i^G \times c_j^E \times \text{Odds}_{i-1,j-1},$$

where the genetic contrast coefficients (c_1^G, c_2^G, c_3^G) are, respectively, $(-1,\ 1,\ 0)$ for IC_1; $(-1,\ 0,\ 1)$ for IC_2; and $(-w_1 - w_2,\ w_1,\ w_2)$ for IC_3. The environmental contrast coefficients (c_1^E, c_2^E) are $(-1,\ 1)$ for all 3 ICs. As long as the sum-to-zero constraints,

$$\sum_{i=1}^{3} c_i^G = \sum_{j=1}^{2} c_j^E = 0,$$

are respected, an IC thus constructed will be equal to a linear combination of the completion rates of—and only of—the interaction classes. Clearly, a nonzero IC is mathematically incompatible with all interaction classes having a zero completion rate. A 2-sided test on an IC,

$$\begin{cases} \text{H}_0\text{: IC} = 0\\ \text{H}_1\text{: IC} \neq 0,\end{cases}$$

is therefore a test for sufficient-cause gene-environment interaction.

The same principle applies in case-control studies, since the case-control odds estimated in a case-control study are a constant multiple (the reciprocal of the control sampling fraction of the study) of the corresponding disease odds in the underlying population (3). A straightforward way to estimate the case-control odds is to divide the number of cases ($\text{CS}_{g,e}$) by the corresponding number of controls ($\text{CN}_{g,e}$) in a stratum defined by genotype ($G = g$) and environmental exposure level ($E = e$) in a case-control data set, that is, $\widehat{\text{Odds}}_{g,e} = \text{CS}_{g,e}/\text{CN}_{g,e}$.Web Appendix 1 (available at http://aje.oxfordjournals.org/) details the method for dealing with a polychotomous environmental exposure with a total of l levels and shows that as long as the sum-to-zero constraints for the contrast coefficients are respected, all of the resulting ICs are legitimate tests for sufficient-cause gene-environment interactions. Asymptotically, the test is a χ^2 test with 1 degree of freedom (df) for each constructed IC. One can also perform a single χ^2 test for several ICs simultaneously. The degrees of freedom equal the number of (linearly independent) ICs. For a ternary G and an l-leveled E, the total number of degrees of freedom amenable for testing is $2 \times (l - 1)$. A simultaneous test with the maximum number of degrees of freedom is referred to as the global test for sufficient-cause gene-environment interactions.

By judiciously designing the contrast coefficients, we can tailor an IC to detect a particular type of departure from additivity (sufficient-cause interactions) with greater sensitivity. For example, we can use $(c_1^G, c_2^G, c_3^G) = (-1,\ 1/2,\ 1/2)$ for an autosomal dominant interaction, $(-1/2,\ -1/2,\ 1)$ for an autosomal recessive interaction, and $(-1,\ 0,\ 1)$ for a gene-dosage interaction. For $l \geq 3$, there are also many possible choices for the environmental contrast coefficients, such as trend interaction (e.g., larger gene-environment interaction with a higher level of environmental exposure) or threshold interaction (e.g., gene-environment interaction occurs only when the environmental exposure reaches a certain threshold level). Specifying contrast coefficients as above, under the specific models they are designed to detect, will cause the ICs to deviate further from zero on average. However, a larger mean deviation is not enough; to be statistically relevant, we need the resulting IC to have a smaller variance as well. Web Appendix 2 shows a weighted version of the IC test, where the user-specified contrast coefficients are weighted by the inverse variances of the case-control odds.

Imposing the independence and HWE assumptions

Because one of the 2 factors considered in this paper is genetic (G) and the other is environmental (E), we will next explore how to exploit the assumptions of GEI and HWE. If GEI can be assumed for the nondiseased subjects in the study population, a more efficient estimation for the case-control odds is

$$\widehat{\text{Odds}}_{g,e}^{\text{GEI}} = \frac{\text{CS}_{g,e}}{\hat{u}_g \times \text{CN}_{+,e}},$$

where $\hat{u}_g = \text{CN}_{g,+}/\text{CN}_{+,+}$ is the estimate of the $G = g$ genotype frequency in the nondiseased subjects in the study

population and the "+" in the subscript indicates summation of the corresponding index.

If GEI and HWE are both assumed for the nondiseased subjects in the study population, the estimates for the case-control odds are

$$\widehat{\text{Odds}}_{0,e}^{\text{GEI \& HWE}} = \frac{\text{CS}_{0,e}}{(1-\hat{u})^2 \times \text{CN}_{+,e}},$$

$$\widehat{\text{Odds}}_{1,e}^{\text{GEI \& HWE}} = \frac{\text{CS}_{1,e}}{2 \times \hat{u} \times (1-\hat{u}) \times \text{CN}_{+,e}},$$

and

$$\widehat{\text{Odds}}_{2,e}^{\text{GEI \& HWE}} = \frac{\text{CS}_{2,e}}{\hat{u}^2 \times \text{CN}_{+,e}},$$

respectively, where $\hat{u} = (\text{CN}_{2,+} + 0.5 \times \text{CN}_{1,+})/\text{CN}_{+,+}$ is the estimated allele frequency among nondiseased subjects in the study population.

Web Appendix 3 presents the formulae of the asymptotic variances for all estimates. Web Appendix 4 presents the SAS code (SAS Institute, Inc., Cary, North Carolina) needed for all of the calculations. Users can specify genetic and environment contrast coefficients to suit the most likely model of sufficient-cause interactions for the disease in question. Calculation of the weighted contrast coefficients is fully automatic, requiring no further input from the user. Web Appendix 5 presents an annotated example of such output.

A SIMULATION STUDY

We conducted a small-scale simulation study to examine the statistical properties of the proposed tests. We simulated data for a ternary gene (allele frequency 0.4) and a binary environmental exposure (prevalence 0.3), assuming GEI and HWE.

For the null hypothesis of no sufficient-cause gene-environment interaction (H_0), we set the completion rates for all of the interaction classes ($U_7 \sim U_{12}$) to zero (Table 1). We further constructed a number of alternative hypotheses (Table 1). For the genetic factor, these include autosomal dominant (AD), autosomal recessive (AR), and gene-dosage (GD) interactions, respectively. For the environmental factor, all alternative hypotheses assumed an all-or-none (AN) interaction pattern (interaction for the exposed subjects but not for the unexposed subjects, completion rates for $U_{10} \sim U_{12}$ being zero). For all hypotheses, the disease rates are on the order of 1 per 10,000 per year.

A case-control study with 1,000 cases and 1,000 controls was conducted in the study population. The proposed IC tests (both unweighted and weighted versions) were applied to the simulated data with various contrast coefficients, with or without the GEI and HWE assumptions. The α level was set at 0.05. A total of 1,000,000 simulations were performed for each scenario. At this high number of simulations, the error is no more than $\sqrt{(0.5 \times 0.5)/1{,}000{,}000} = 0.0005$.

Table 2 presents the type I error rates (under H_0 in Table 1) of the IC tests and the weighted IC tests using different (specified) genetic contrast coefficients. The corresponding weighted genetic contrast coefficients are shown in the same rows as the specified ones. As noted above, the weighted coefficients are automatically calculated from the data once the unweighted ones are specified/supplied. Here we show the averages from the total 1,000,000 simulations. Type I error rates of the unweighted and weighted IC tests are well controlled for all of the contrast coefficients we specified, with or without imposing the assumptions of GEI and HWE. If one or both assumptions fail, only the unweighted tests and the weighted IC tests without the failed assumption(s) can maintain the nominal α level (see Web Tables 1 and 2).

Table 3 presents the empirical powers of the IC tests and the weighted IC tests under the alternative hypothesis of an interaction between an autosomal dominant gene and an all-or-none environmental exposure ($H_{AD\times AN}$ in Table 1). For all of the genetic contrast coefficients specified, statistical power increases when more assumptions (GEI and/or HWE) are imposed.

Table 3 also shows that given the same assumption(s), the power is highest for the 1-df weighted IC test with correctly specified genetic contrast coefficients; that is, the test using (−0.78, 0.59, 0.19) that corresponds to the autosomal dominant coefficients (−1, 1/2, 1/2). IC tests using contrast coefficients closer to (−0.78, 0.59, 0.19) also have favorable powers—that is, 1) the unweighted IC test using (−1, 1/2, 1/2) and 2) the

Table 1. Completion Rates (per 100,000 Population per Year) for 12 Classes of Sufficient Component Causes Under Various Hypotheses in a Simulation Study of Gene-Environment Interaction

Hypothesis	Class of Sufficient Component Causes											
	All-Unknown Class (U_1)	Gene-Only Classes			Environment-Only Classes		Gene-Environment Interaction ($G \times E$) Classes					
		$G=2$ (U_2)	$G=1$ (U_3)	$G=0$ (U_4)	$E=1$ (U_5)	$E=0$ (U_6)	2×1 (U_7)	1×1 (U_8)	0×1 (U_9)	2×0 (U_{10})	1×0 (U_{11})	0×0 (U_{12})
H_0[a]	1	4	2	1	3	1	0	0	0	0	0	0
$H_{AD\times AN}$[b]	1	4	2	1	3	1	4	4	1	0	0	0
$H_{AR\times AN}$[c]	1	4	2	1	3	1	6	1	1	0	0	0
$H_{GD\times AN}$[d]	1	4	2	1	3	1	6	3	1	0	0	0

Abbreviations: AD, autosomal dominant; AN, all or none; AR, autosomal recessive; GD, gene-dosage.
[a] Null hypothesis.
[b] AD for gene, AN for environmental exposure.
[c] AR for gene, AN for environmental exposure.
[d] GD for gene, AN for environmental exposure.

Table 2. Genetic Contrast Coefficients and Type I Error Rates (Under H_0 in Table 1) of the Interaction Contrast Tests, for the Data Simulated Under the Assumptions of Gene-Environment Independence and Hardy-Weinberg Equilibrium

Type of Interaction More Sensitive to Detection	Genetic Contrast Coefficient[a]						Type I Error Rate					
	Specified			Weighted			Assuming Neither GEI nor HWE		Assuming GEI Only		Assuming Both GEI and HWE	
	$G=0$	$G=1$	$G=2$	$G=0$	$G=1$	$G=2$	IC Test	Weighted IC Test	IC Test	Weighted IC Test	IC Test	Weighted IC Test
Autosomal dominant	−1	1/2	1/2	−0.77	0.62	0.15	0.044	0.047	0.049	0.049	0.049	0.049
Autosomal recessive	−1/2	−1/2	1	−0.31	−0.49	0.80	0.045	0.044	0.049	0.049	0.050	0.050
Gene-dosage	−1	0	1	−0.81	0.32	0.49	0.044	0.044	0.049	0.049	0.050	0.049
Global	−1	1	0	−0.68	0.73	−0.05	0.044	0.044	0.049	0.049	0.050	0.050
	−1	0	1	−0.81	0.32	0.49						

Abbreviations: GEI, gene-environment independence; HWE, Hardy-Weinberg equilibrium; IC, interaction contrast.
[a] The environmental contrast coefficients are specified as (−1, 1) for ($E=0$, $E=1$) for all scenarios, with the corresponding weighted environmental contrast coefficients calculated as (−0.71, 0.71).

weighted IC test using (−0.81, 0.33, 0.48) that corresponds to the gene-dosage contrast coefficients (−1, 0, 1). The weighted and unweighted tests are the same for the 2-df global interaction test. This is because the pair of vectors spans the entire interaction subspace, regardless of the values used for the weights.

The 1-df weighted IC test with correctly specified genetic contrast coefficients has the highest power to detect the other 2 alternatives as well (Web Table 3 for $H_{AR\times AN}$, Web Table 4 for $H_{GD\times AN}$). The global test can detect all types of alternatives listed in Table 1 with reasonably high power (Table 3, Web Tables 3 and 4). We also tested other allele frequencies and exposure prevalences and achieved similar results (not shown). Based on the simulation results, we recommend using the 1-df weighted IC test when one has a priori knowledge about the interaction model and using the global test if one does not.

AN EXAMPLE

We used Sam et al.'s (21) case-control data on upper aerodigestive tract cancers as an example. In this study, Sam et al. examined the relationship between polymorphisms in the cytochrome P-450, family 1, subfamily A, polypeptide 1 gene (*CYP1A1*)—*CYP1A1*2A* and *CYP1A1*1A*, with the former being the variant allele—and smoking (yes/no) in the risk of contracting upper aerodigestive tract cancers (Table 4). None of the cells in Table 4 have counts of less than 5. The assumptions of GEI ($P=0.7563$) and HWE ($P=0.6789$) are both tenable among control subjects. Lee et al. (9) previously analyzed these data assuming a multiplicative model for disease odds. Under that model, they concluded that the (multiplicative) interactions between *CYP1A1* polymorphisms and smoking were nonsignificant at an α level of 0.05, with or without the assumptions of GEI and HWE.

We next applied the weighted IC test to these data for any sufficient-cause gene-environment interaction (Table 5). Without the assumptions of GEI and HWE, none of the genetic contrast coefficients we tried (autosomal dominant, autosomal recessive, gene-dosage, and global) resulted in statistically significant test results. However, with the GEI assumption invoked, the weighted IC tests became significant at $\alpha=0.05$ for all contrasts. When both the GEI assumption and the HWE assumption were invoked, all 3 sets of genetic

Table 3. Genetic Contrast Coefficients and Empirical Powers (Under $H_{AD\times AN}$ in Table 1) of the Interaction Contrast Tests, for the Data Simulated Under the Assumptions of Gene-Environment Independence and Hardy-Weinberg Equilibrium

Type of Interaction More Sensitive to Detection	Genetic Contrast Coefficient[a]						Empirical Power					
	Specified			Weighted			Assuming Neither GEI nor HWE		Assuming GEI Only		Assuming Both GEI and HWE	
	$G=0$	$G=1$	$G=2$	$G=0$	$G=1$	$G=2$	IC Test	Weighted IC Test	IC Test	Weighted IC Test	IC Test	Weighted IC Test
Autosomal dominant	−1	1/2	1/2	−0.78	0.59	0.19	0.52	0.63	0.84	0.90	0.85	0.91
Autosomal recessive	−1/2	−1/2	1	−0.39	−0.42	0.81	0.05	0.06	0.14	0.15	0.16	0.17
Gene-dosage	−1	0	1	−0.81	0.33	0.48	0.19	0.45	0.49	0.79	0.50	0.79
Global	−1	1	0	−0.70	0.71	−0.01	0.51	0.51	0.84	0.84	0.85	0.85
	−1	0	1	−0.81	0.32	0.48						

Abbreviations: GEI, gene-environment independence; HWE, Hardy-Weinberg equilibrium; IC, interaction contrast.
[a] The environmental contrast coefficients are specified as (−1, 1) for ($E=0$, $E=1$) for all scenarios, with the corresponding weighted environmental contrast coefficients calculated as (−0.71, 0.71).

Table 4. A Case-Control Data Set From the Study by Sam et al. (21) and the Estimated Case-Control Odds

	No. of Cases	No. of Controls	Estimated Case-Control Odds[a]		
			Assuming Neither GEI nor HWE	Assuming GEI Only	Assuming Both GEI and HWE
$E=0$					
$G=0$	55	70	0.79	0.75	0.73
$G=1$	76	62	1.23	1.30	1.37
$G=2$	19	9	2.11	2.12	1.84
$E=1$					
$G=0$	91	45	2.02	2.20	2.16
$G=1$	123	29	4.24	3.76	3.95
$G=2$	44	5	8.80	8.75	7.61

Abbreviations: GEI, gene-environment independence; HWE, Hardy-Weinberg equilibrium.

[a] The probability of a study subject's being a case divided by the probability of his/her being a control in a case-control study.

contrast coefficients designed to be sensitive to detecting autosomal dominant, autosomal recessive, and gene-dosage interactions, respectively, resulted in highly significant test results at $\alpha = 0.01$. If the multiple testing issue (a total of 12 tests performed simultaneously in this example) is taken into account, then 2 tests were significant at $\alpha = 0.05$ after Bonferroni correction (footnote "c" in Table 5), and all of the tests with the GEI assumption were significant at a false discovery rate (22) of 5%.

DISCUSSION

The sufficient-cause model is nonidentifiable, with the number of model parameters exceeding the total number of degrees of freedom in the data. For example, with a ternary gene and a binary environmental exposure, the model has a total of 12 completion rates (the model parameters) but cohort data can provide at most 6 gene- and environmental exposure–specific disease rates (the data degrees of freedom). An estimation of model parameters is not possible in a nonidentifiable model, but hypothesis testing is a different story. This paper demonstrates that the IC test can detect gene-environment interactions, even under the nonidentifiable sufficient-cause model. The significance of an IC test implies the presence of at least 1 interaction class involved in the given contrast; that is, some sufficient-cause gene-environment interaction is occurring. A nonsignificant test result, however, does not guarantee the opposite; a perfect cancellation of several interaction classes with nonzero completion rates also leads to IC = 0. An example is the "leaky" (L) environmental exposure where the interaction completion rates for the unexposed subjects (U_{10}, U_{11}, U_{12}) are exactly 1×10^{-5} shy of those for the exposed subjects (U_7, U_8, U_9) in Table 1. Here, we have no power whatsoever against $H_{AD\times L}$, $H_{AR\times L}$, and $H_{GD\times L}$ alternatives. This is an unavoidable limitation for any hypothesis testing of the parameters of a nonidentifiable model.

The rare-disease assumption is necessary for the proposed method, under which a sufficient-cause model for *rates* is equivalent to a sufficient-cause model for *odds*. Web Table 5 presents the biases of using odds to approximate rates. The biases are less than 0.05% for rates under 0.001 per year—the setting for studies of cancers, coronary heart diseases, etc. For more common diseases, such as hypertension or type 2 diabetes, the approximation breaks down, and the method proposed here would be inapplicable.

Another necessary assumption, the no-redundancy assumption, is more subtle and is not amenable to testing by itself. This is actually a much weaker assumption than the simple independent action assumption, which originated from toxicopharmacology (23) and found use in epidemiology in recent decades (17, 20, 24–31). In the language of the sufficient-cause model, the simple independent action assumption posits that the arrival events of the unknown complement causes in different classes of sufficient causes are independent of one another. The no-redundancy assumption can still hold, even if there is strong dependency in the arrival events (and the simple independent action assumption fails).

Table 5. Genetic Contrast Coefficients and *P* Values for the Data Shown in Table 4

Type of Interaction More Sensitive to Detection	Genetic Contrast Coefficient[a]						*P* Value From the Weighted Interaction Contrast Test		
	Specified			Weighted			Assuming Neither GEI nor HWE	Assuming GEI Only	Assuming Both GEI and HWE
	$G=0$	$G=1$	$G=2$	$G=0$	$G=1$	$G=2$			
Autosomal dominant	−1	1/2	1/2	−0.81	0.49	0.32	0.0696	0.0088[b]	0.0035[b,c]
Autosomal recessive	−1/2	−1/2	1	−0.77	0.16	0.62	0.1644	0.0209[b]	0.0062[b]
Gene-dosage	−1	0	1	−0.82	0.40	0.41	0.0991	0.0120[b]	0.0040[b,c]
Global	−1	1	0	−0.79	0.56	0.23	0.0858	0.0239[b]	0.0139[b]
	−1	0	1	−0.82	0.40	0.41			

Abbreviations: GEI, gene-environment independence; HWE, Hardy-Weinberg equilibrium.

[a] The environmental contrast coefficients are specified as (−1, 1) for ($E=0$, $E=1$) for all scenarios, with the corresponding weighted environmental contrast coefficients calculated as (−0.71, 0.71).

[b] Significant at a false discovery rate of 5%.

[c] Significant at $\alpha = 0.05$ after Bonferroni correction.

To break the no-redundancy assumption, the unknown complement causes for 2 different classes need to have a common constituent factor, and that common factor must be the last one to arrive among all of the constituents of these 2 classes, before all other classes are completed.

The remaining 2 assumptions, the GEI and HWE assumptions, are optional, and when their validity is in doubt, a researcher always has the liberty to put them to test using the data at hand. Han et al. (32) exploited the independence assumption to develop a constrained likelihood ratio test for gene-environment interactions under an additive risk model. However, their method cannot take into account the HWE assumption, and it is highly demanding computationally. By contrast, all the formulae presented here—with or without the GEI and HWE assumptions—are of the closed form, obviating the need for a computer-intensive iteration algorithm. Lee et al.'s (9) modified logistic regression model can exploit both assumptions and can be easily fitted using common statistical packages. However, it detects *multiplicative* interactions, not *sufficient-cause* gene-environment interactions, which are the focus of this paper.

While the proposed IC test can be tailored to detect a particular type of sufficient-cause gene-environment interaction with greater sensitivity, it is actually a nonspecific test. To pin down a specific interaction—for example, the U_7 interaction class—one can test for $\max(\text{Odds}_{2,1} - \text{Odds}_{0,1} - \text{Odds}_{2,0}, \text{Odds}_{2,1} - \text{Odds}_{1,1} - \text{Odds}_{2,0}) > 0$. (For a "$G = g$ and $E = e$" gene-environment interaction class, one tests for

$$\max_{\substack{g' \neq g \\ e' \neq e}} (\text{Odds}_{g,e} - \text{Odds}_{g',e} - \text{Odds}_{g,e'}) > 0.$$

VanderWeele referred to this as a "singular interaction" or an "epistatic interaction," since there are individuals for whom the outcome would occur if and only if $G = g$ and $E = e$ (33–35).) This is because $\text{Odds}_{2,1} - \text{Odds}_{0,1} - \text{Odds}_{2,0} = \text{Rate}_{U_7} - \text{Rate}_{U_1} - \text{Rate}_{U_4} - \text{Rate}_{U_6} - \text{Rate}_{U_{10}}$ and $\text{Odds}_{2,1} - \text{Odds}_{1,1} - \text{Odds}_{2,0} = \text{Rate}_{U_7} - \text{Rate}_{U_1} - \text{Rate}_{U_3} - \text{Rate}_{U_6} - \text{Rate}_{U_{10}}$, so if statistically either one is larger than zero, it must be the case that $\text{Rate}_{U_7} > 0$, and therefore the presence of the U_7 interaction class can be inferred. Further work is needed to develop empirical tests for these special forms of gene-environment interaction in case-control studies, with and without the assumptions of GEI and HWE.

Besides the genetic and environmental factors under consideration, there may be other factors that could confound gene-environment interactions. It may also be that the study population is not a homogeneous one but instead is composed of several population strata (36–38). The assumptions of GEI and HWE hold within each population stratum but do not hold in the population as a whole. To account for these, one can stratify the data according to confounders and population strata and then perform a separate IC test in each resulting stratum. With a proper multiple-testing correction for multiple strata, the presence of some sufficient-cause gene-environment interaction can be inferred if the result of any of these stratum-specific IC tests turns out to be significant. Further work is needed to develop stratified sufficient-cause interaction testing methods, both when the total number of strata is large (and the average stratum size is small, i.e., the sparse-data scenario) and when some of the stratifying variables interact with the specific gene and environmental exposure under study (sufficient-cause interactions between 3 or more variables).

ACKNOWLEDGMENTS

Author affiliation: Research Center for Genes, Environment and Human Health and Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan (Wen-Chung Lee).

This study was partly supported by grants from the Ministry of Science and Technology, Taiwan (grant NSC 102-2628-B-002-036-MY3) and National Taiwan University (grant NTU-CESRP-102R7622-8). No additional external funding was received.

The funders played no role in the study design, data collection, and analysis, the decision to publish, or the preparation of the manuscript.

Conflict of interest: none declared.

REFERENCES

1. Hunter DJ. Gene-environment interactions in human diseases. *Nat Rev Genet*. 2005;6(4):287–298.
2. Olden K. Commentary: from phenotype, to genotype, to gene-environment interaction and risk for complex diseases. *Int J Epidemiol*. 2007;36(1):18–20.
3. Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
4. Sham P. *Statistics in Human Genetics*. New York, NY: Oxford University Press; 1998.
5. Umbach DM, Weinberg CR. Designing and analysing case-control studies to exploit independence of genotype and exposure. *Stat Med*. 1997;16(15):1731–1743.
6. Chatterjee M, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*. 2005;92(2):399–418.
7. Cheng KF, Lin WJ. Retrospective analysis of case-control studies when the population is in Hardy-Weinberg equilibrium. *Stat Med*. 2005;24(21):3289–3310.
8. Chen YH, Kao JT. Multinomial logistic regression approach to haplotype association analysis in population-based case-control studies. *BMC Genet*. 2006;7:43.
9. Lee WC, Wang LY, Cheng KF. An easy-to-implement approach for analyzing case-control and case-only studies assuming gene-environment independence and Hardy-Weinberg equilibrium. *Stat Med*. 2010;29(24):2557–2567.
10. VanderWeele TJ, Robins JM. The identification of synergism in the sufficient-component-cause framework. *Epidemiology*. 2007;18(3):329–339.
11. VanderWeele TJ, Robins JM. Empirical and counterfactual conditions for sufficient cause interactions. *Biometrika*. 2008; 95(1):49–61.
12. VanderWeele TJ. Sufficient cause interactions for categorical and ordinal exposures with three levels. *Biometrika*. 2010; 97(3):647–659.
13. Lee WC. Testing synergisms in a no-redundancy sufficient-cause rate model. *Epidemiology*. 2013;24(1):174–175.

14. Lee WC. Assessing causal mechanistic interactions: a peril ratio index of synergy based on multiplicativity. *PLoS One*. 2013; 8(6):e67424.
15. Gatto NM, Campbell UB. Redundant causation from a sufficient cause perspective. *Epidemiol Perspect Innov*. 2010;7:5.
16. Greenland S, Brumback B. An overview of relations among causal modelling methods. *Int J Epidemiol*. 2002;31(5): 1030–1037.
17. Liao SF, Lee WC. Weighing the causal pies in case-control studies. *Ann Epidemiol*. 2010;20(7):568–573.
18. Suzuki E, Yamamoto E, Tsuda T. On the link between sufficient-cause model and potential-outcome model. *Epidemiology*. 2011;22(1):131–132.
19. Suzuki E, Yamamoto E, Tsuda T. On the relations between excess fraction, attributable fraction, and etiologic fraction. *Am J Epidemiol*. 2012;175(6):567–575.
20. Lee WC. Completion potentials of sufficient component causes. *Epidemiology*. 2012;23(3):446–453.
21. Sam SS, Thomas V, Reddy SK, et al. CYP1A1 polymorphisms and the risk of upper aerodigestive tract cancers in an Indian population. *Head Neck*. 2008;30(12):1566–1574.
22. Benjamini Y, Hochberg Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J R Stat Soc Series B*. 1995;57(1):289–300.
23. Bliss CI. The toxicity of poisons applied jointly. *Ann Appl Biol*. 1939;26(3):585–615.
24. Weinberg CR. Applicability of the simple independent action model to epidemiologic studies involving two factors and a dichotomous outcome. *Am J Epidemiol*. 1986;123(1): 162–173.
25. Allard R, Boivin JF. Measures of effect based on the sufficient causes model. 1. Risks and rates of disease associated with a single causative agent. *Epidemiology*. 1993;4(1):37–42.
26. Allard R, Boivin JF. Measures of effect based on the sufficient causes model. 2. Risks and rates of disease associated with a single preventive agent. *Epidemiology*. 1993;4(6): 517–523.
27. Weinberg CR. Can DAGs clarify effect modification? *Epidemiology*. 2007;18(5):569–572.
28. Weinberg CR. Less is more, except when less is less: studying joint effects. *Genomics*. 2009;93(1):10–12.
29. Wang TE, Lin CY, King CC, et al. Estimating pathogen-specific asymptomatic ratios. *Epidemiology*. 2010;21(5):726–728.
30. Madsen AM, Ottman R, Hodge SE. Causal models for investigating complex genetic disease: II. What causal models can tell us about penetrance for additive, heterogeneity, and multiplicative two-locus models. *Hum Hered*. 2011;72(1): 63–72.
31. Wang JF, Li XH, Christakos G, et al. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. *Int J Geogr Inf Sci*. 2010;24(1):107–127.
32. Han SS, Rosenberg PS, Garcia-Closas M, et al. Likelihood ratio test for detecting gene (G)-environment (E) interactions under an additive risk model exploiting G-E independence for case-control data. *Am J Epidemiol*. 2012;176(11):1060–1067.
33. VanderWeele TJ. Empirical tests for compositional epistasis. *Nat Rev Genet*. 2010;11(2):166.
34. VanderWeele TJ. Epistatic interactions. *Stat Appl Genet Mol Biol*. 2010;9:Article 1.
35. VanderWeele TJ. Sample size and power calculation for additive interactions. *Epidemiol Methods*. 2012;1(1):8.
36. Lee WC, Wang LY. Simple formulas for gauging the potential impacts of population stratification bias. *Am J Epidemiol*. 2008; 167(1):86–89.
37. Wang LY, Lee WC. Population stratification bias in the case-only study for gene-environment interactions. *Am J Epidemiol*. 2008;168(2):197–201.
38. Lee WC, Wang LY. Reducing population stratification bias: stratum matching is better than exposure. *J Clin Epidemiol*. 2009;62(1):62–66.