Research Article

Statistics in Medicine

Received 7 May 2009, Accepted 25 June 2010 Published online 25 August 2010 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4028

An easy-to-implement approach for analyzing case–control and case-only studies assuming gene–environment independence and Hardy–Weinberg equilibrium

Wen-Chung Lee,[a*†] Liang-Yi Wang[b] and K. F. Cheng[c]

The case–control study is a simple and an useful method to characterize the effect of a gene, the effect of an exposure, as well as the interaction between the two. The control-free case-only study is yet an even simpler design, if interest is centered on gene–environment interaction only. It requires the sometimes plausible assumption that the gene under study is independent of exposures among the non-diseased in the study populations. The Hardy–Weinberg equilibrium is also sometimes reasonable to assume. This paper presents an easy-to-implement approach for analyzing case–control and case-only studies under the above dual assumptions. The proposed approach, the 'conditional logistic regression with counterfactuals', offers the flexibility for complex modeling yet remains well within the reach to the practicing epidemiologists. When the dual assumptions are met, the conditional logistic regression with counterfactuals is unbiased and has the correct type I error rates. It also results in smaller variances and achieves higher powers as compared with using the conventional analysis (unconditional logistic regression). Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: genetic association studies; epidemiologic methods; gene–environment interaction

Introduction

The occurrences of most human diseases are the results of interplay between genes and environmental exposures [1, 2]. The case–control study is a simple and an useful method to characterize the effect of a gene, the effect of an exposure, as well as the interaction between the two, whereas the control-free case-only study is yet an even simpler design, if interest is centered on gene–environment interaction only [3, 4].

It is sometimes reasonable to assume that the gene under study is in Hardy–Weinberg equilibrium (HWE) and is independent of exposures (GE independence) among the non-diseased subjects in the study population: the population genetics theory dictates that a gene will achieve HWE within one generation of random mating in any given population [5]; and a subject's genes which are determined from birth often will not predict his/her subsequent environmental exposures. The previous studies showed that imposing GE independence and/or HWE can improve the statistical efficiency of a case–control study [6–9], whereas the GE independence is a necessary condition for the validity of a case-only study [10].

In this paper, we show how to analyze case–control and case-only studies assuming GE independence and HWE with readily available statistical packages. We demonstrate the method in a real data set. We also perform the Monte–Carlo simulation to investigate the statistical performances of the method.

[a] *Research Center for Genes, Environment and Human Health and Graduate Institute of Epidemiology, College of Public Health, National Taiwan University, Taiwan*

[b] *Institute of Public Health, College of Medicine, National Cheng-Kung University, Taiwan*

[c] *Graduate Institute of Biostatistics, China Medical University, Taiwan*

[*] *Correspondence to: Wen-Chung Lee, Rm. 536, No. 17, Xuzhou Rd., Taipei 100, Taiwan.*

[†] *E-mail: wenchung@ntu.edu.tw*

Backgrounds and notations

Assume that we are interested in characterizing the effect of a gene, the effect of an exposure, and possible gene–environment interactions, on the risk of a specific disease. Let $G=0$, 1, and 2, represent the number of variant alleles a subject carries. We define G_1 to be 1, if $G=1$, and 0, otherwise; and G_2 to be 1, if $G=2$, and 0, otherwise. (G_1 and G_2 are the corresponding dummy variables for the 3-level G.) Let E represent the exposure status of a subject: $E=1$ for an exposed subject, $E=0$ for an unexposed subject. (For now, we assume one single binary exposure variable. The situations of exposure in any measurement scale, binary, categorical or continuous, and of multiple exposures will be discussed later.) The disease status of a subject is denoted by D, with $D=1$ for diseased and $D=0$ otherwise.

We assume that the disease risk in the study population follows a logistic model:

$$\log\left(\frac{\Pr(D=1|G,E)}{\Pr(D=0|G,E)}\right)=\mu+\alpha_1G_1+\alpha_2G_2+\beta E+\gamma_1G_1E+\gamma_2G_2E. \tag{1}$$

For a rare disease, this is simply multiplicative in risk. The $\exp(\mu)$ is the background disease odds in the population. The $\exp(\alpha_1)=\psi_1$ is the odds ratio of disease for those carrying one variant allele versus those carrying none, and the $\exp(\alpha_2)=\psi_2$, the odds ratio for those carrying two versus those carrying none. The $\exp(\beta)$ is the odds ratio associated with the environmental exposure variable, and the $\exp(\gamma_1)$ and $\exp(\gamma_2)$ are the odds ratios associated with gene–environment interactions. We also assume GE independence and HWE among the non-diseased subjects in the study population, that is,

$$\begin{aligned}\Pr(G|E,D=0)&=\Pr(G|D=0)\\&=(1-p)^2\times 2^{G_1}\times[p/(1-p)]^G\\&=\Pr(G=0|D=0)\times\exp(G_1\log 2+\delta G),\end{aligned} \tag{2}$$

where p is allele frequency of the variant allele among the non-diseased subjects and $\delta=\log[p/(1-p)]$ is log allele frequency odds.

A case–control study (and a case-only study) recruits only a certain number of subjects in the study population but not all. Denote R as the recruitment status of a subject, $R=1$, if the subject is recruited in the study; $R=0$, if otherwise. (R is a random variable and plays a pivotal role in this paper as will be evident in the subsequent developments.) Appendix A shows that for a case–control study conducted in the above population, the likelihood function is

$$\Pr(D,G|E,R=1)=\frac{\exp(G_1\log 2+\delta G+\mu^*D+\alpha_1G_1D+\alpha_2G_2D+\beta ED+\gamma_1G_1ED+\gamma_2G_2ED)}{\sum_{d=0}^{1}\sum_{g=0}^{2}\exp(g_1\log 2+\delta g+\mu^*d+\alpha_1g_1d+\alpha_2g_2d+\beta Ed+\gamma_1g_1Ed+\gamma_2g_2Ed)}, \tag{3}$$

where $\mu^*=\mu+\log[\Pr(R=1|D=1)/\Pr(R=1|D=0)]$ is a nuisance parameter. For a case-only study composed of disease subjects only, the likelihood function is

$$\Pr(G|E,R=1)=\frac{\exp(G_1\log 2+\delta^*G+\eta G_2+\gamma_1G_1E+\gamma_2G_2E)}{\sum_{g=0}^{2}\exp(g_1\log 2+\delta^*g+\eta g_2+\gamma_1g_1E+\gamma_2g_2E)}, \tag{4}$$

where $\delta^*=\delta+\alpha_1$ (a nuisance parameter) and $\eta=\alpha_2-2\alpha_1$. We call (3) and (4) the 'conditional-on-exposure' likelihoods. It is of interest to note that the original logistic model in (1) is conditional on both G and E. But with the dual assumptions in (2) imposed, the likelihoods in (3) and (4) become conditional on E only.

The η parameter [$\exp(\eta)=\psi_2/\psi_1^2$] in (4) deserves special attention. It measures the departure from a multiplicative gene-dose model, with value of 0 implying a perfect multiplicative model ($\psi_2=\psi_1^2$) or that the gene has no effect on disease risk ($\psi_1=\psi_2=1$); value >0, a supra-multiplicative model (e.g. an autosomal recessive model); and value <0, a sub-multiplicative model (e.g. an autosomal dominant model).

Implementing the conditional-on-exposure analysis

We adopt a counterfactual approach for fitting the above conditional-on-exposure likelihoods. The counterfactual analysis has been used in many other genetic or epidemiologic settings [11–19]. Here, we found it applicable to the present context as well. To be specific, we create 5 counterfactual subjects (all of them with $R=0$) to each recruited subject ($R=1$) for the case–control data. The exposure variable (E) of these counterfactual subjects is deliberately set to be exactly the same as the recruited subject himself/herself, but the status of disease (D) and gene (G) is different (the five subjects represents the five different ways of making $[D, G]$-different counterfactuals). Treating the recruitment status (R) of a subject, factual or counterfactual, as a random variable, we perform a conditional logistic regression analysis (based on the 1:5 matched data) with the following regression equation: $G_1 \log 2+\delta G+\mu^* D+\alpha_1 G_1 D+\alpha_2 G_2 D+\beta ED+\gamma_1 G_1 ED+\gamma_2 G_2 ED$. Although this regression equation appears daunting involving interactions terms between two factors (‘gene×disease’, ‘environment×disease’) and even up to between three factors (‘gene×environment×disease’), it is actually a very simple linear function in terms of the regression coefficients, δ, μ^*, α_1, α_2, β, γ_1, γ_2, and can be fitted using the readily available statistical packages. (The $G_1 \log 2$ is to be declared as the ‘offset’ term.) One should note that such a conditional logistic regression analysis for 1:5 matched data has exactly the same likelihood as in (3), and therefore represents an easy way to implement the conditional-on-exposure analysis.

For the case-only data, we create 2 counterfactual $R=0$ subjects to each $R=1$ subject, with the counterfactuals having exactly the same E as the factual subject but different genes (there are two ways of making this). And the conditional logistic regression has the following linear form: $G_1 \log 2+\delta^* G+\eta G_2+\gamma_1 G_1 E+\gamma_2 G_2 E$. This regression equation requires up to two-factor interaction terms. Such a conditional logistic regression analysis for 1:2 matched data has exactly the same likelihood as in (4).

SAS codes for implementing the above ‘conditional logistic regression with counterfactuals’ are presented in Appendices B (for case–control study) and C (for case-only study).

An example

We use the data of a case–control study conducted by Sam *et al.* [20] as an example. The study examined the relationship between *CYP1A1* polymorphisms and smoking on the risks of upper aerodigestive tract (UADT) cancers in an Indian population. Table I shows the distribution of *CYP1A1* genotypes (*CYP1A1*$*$*2A* being the variant allele) and smoking status among the subjects in that study. We test the assumptions of GE independence and HWE in the control subjects in this dataset, the p values are 0.7563 (for GE independence) and 0.6789 (for HWE), respectively.

As the dual assumptions of GE independence and HWE are tenable, we perform the conditional logistic regression with counterfactuals for this data. The results are presented in Table II. For comparison, we also present the results using the standard (unconditional) logistic regression for this same data. It can be clearly seen that the proposed conditional logistic regression with counterfactuals results in smaller standard errors for all the corresponding regression coefficients as compared with the conventional unconditional logistic regression.

In both the approaches, the individuals with the *CYP1A1*$*2A/*2A$ genotype ($G=2$) (the α_2 coefficient) or those who smoke ($E=1$) (the β coefficient) have a statistically significant risk for UADT cancers. However, using the standard unconditional logistic regression, we found that the UADT cancer risk among individuals with the *CYP1A1*$*2A/*1A$ genotype ($G=1$) (the α_1 coefficient) is not significantly different ($p=0.0735$) from the risk of those with the *CYP1A1*$*1A/*1A$ genotype ($G=0$). But using the proposed conditional logistic regression with counterfactuals, the difference in risks reaches the statistical significance ($p=0.0026$). The interactions between *CYP1A1* gene and smoking (the γ_1 and γ_2 coefficients) do not reach statistical significance using either approach, though.

For a demonstration, we also perform a case-only analysis to the case subjects in Sam *et al.*’s study (results also presented in Table II). With the dual assumptions of HWE and GE independence imposed, the results for gene–environment interactions (the estimates and the standard errors of γ_1 and γ_2) using a case-only analysis are exactly the same as the results when using a case–control analysis. In addition, there is an exact equality between δ^* $(=-0.3698$, in case-only analysis) and $\delta+\alpha_1$ $(=-0.9923+0.6226=-0.3698$, in case–control analysis) and between η $(=-0.3234$, in case-only analysis) and $\alpha_2-2\alpha_1$ $(=0.9217-2\times 0.6226=-0.3234$, in case–control analysis). The δ^* is nuisance but η is informative. We see that η is not significant in this data ($p=0.3574$), implying that the relationship between *CYP1A1* genotypes and UADT cancers is conforming to a multiplicative gene-dose model or that the *CYP1A1* gene has no effect on UADT cancers risk.

Table I. Distribution of *CYP1A1* genotypes and smoking status among the subjects in Sam *et al.*’s study [20].

	Smokers		Non-smokers	
CYP1A1	Number of cases	Number of controls	Number of cases	Number of controls
$*2A/*2A$	44	5	19	9
$*2A/*1A$	123	29	76	62
$*1A/*1A$	91	45	55	70

Table II. Analysis results for the data in Table I.

					exp(Regression coefficient)	
Regression term	Regression coefficient	Estimate	Standard error	P value	Estimate	95 per cent confidence interval
Standard unconditional logistic regression (*cases and controls*)						
Intercept	μ^*	−0.2412	0.1802	0.1808	0.79	0.55–1.12
G_1	α_1	0.4448	0.2485	0.0735	1.56	0.96–2.54
G_2	α_2	0.9884	0.4430	0.0257	2.69	1.13–6.40
E	β	0.9454	0.2563	0.0002	2.57	1.56–4.25
G_1E	γ_1	0.2959	0.3709	0.4250	1.34	0.65–2.78
G_2E	γ_2	0.4822	0.6724	0.4733	1.62	0.43–6.05
Conditional logistic regression with counterfactuals (*cases and controls*)						
G	δ	−0.9923	0.1073	<0.0001	0.37	0.30–0.46
D	μ^*	−0.3108	0.1693	0.0663	0.73	0.53–1.02
G_1D	α_1	0.6226	0.2070	0.0026	1.86	1.24–2.80
G_2D	α_2	0.9217	0.3419	0.0070	2.51	1.29–4.91
ED	β	1.0828	0.2212	<0.0001	2.95	1.91–4.56
G_1ED	γ_1	−0.0221	0.2246	0.9217	0.98	0.63–1.52
G_2ED	γ_2	0.3362	0.3233	0.2984	1.40	0.74–2.64
Conditional logistic regression with counterfactuals (*cases only*)						
G	δ^*	−0.3698	0.1770	0.0367	0.69	0.49–0.98
G_2	η	−0.3234	0.3514	0.3574	0.72	0.36–1.44
G_1E	γ_1	−0.0221	0.2246	0.9217	0.98	0.63–1.52
G_2E	γ_2	0.3362	0.3233	0.2984	1.40	0.74–2.64

A simulation study

We perform a small-scale simulation study to examine the statistical properties of the conditional logistic regression with counterfactuals. For simplicity, we assume a binary exposure E ($E=0,1$) and a gene G ($G=0,1,2$). For the non-diseased subjects in the study population, we assume GE independence and HWE with the allele frequency and the exposure prevalence both being set at 0.5. We assume two scenarios for the genetic effects. In scenario I, a null gene is assumed. It has no bearing on the disease risk whatsoever (genetic main effects: $\alpha_1=\alpha_2=0.0000$; gene–environment interactions: $\gamma_1=\gamma_2=0.0000$). For scenario II, an autosomal recessive gene is assumed. Its genetic main effects are: $\alpha_1=0.0000$ (odds ratio=1.0 for those with $G=1$ as compared to those with $G=0$) and $\alpha_2=0.4055$ (odds ratio$=1.5$ for those with $G=2$ as compared to those with $G=0$), respectively. And its gene–environment interaction parameters are $\gamma_1=0.0000$ and $\gamma_2=0.4055$, respectively. As for the environmental main effect, we assume that the odds ratio for those with $E=1$ as compared to those with $E=0$ is 2.0 ($\beta=0.6931$) for both the scenarios.

In a case–control study conducted in the study population, we assume 500 cases (the diseased subjects) and various numbers of controls (the non-diseased subjects): 250, 500, 750, and 1000, respectively. The disease probabilities of subjects in the study population are assumed to follow the logistic model in (1), with parameters given in the preceding paragraph. The controls are assumed to be a random sample of the non-diseased subjects in the population.

The conditional logistic regression with counterfactuals is employed for the analysis of the simulated case–control data. (For a comparison, we also perform the standard unconditional logistic regression.) The simulation was done for a total of 10 000 times for each setting. The bias of a regression coefficient is calculated as the difference between the mean of the estimates and its true value. The variance of a regression coefficient is calculated as the sample variance of the estimates. We also calculate the average of the estimated variances. The type I error rates (for α_1, α_2, γ_1, and γ_2 in scenario I; α_1 and γ_1 in scenario II) and the powers (for β in scenario I; α_2, β and γ_2 in scenario II) are calculated as the total number of rejections of the null hypothesis that a specific regression coefficient is zero divided by 10 000 using a significance level of 0.05.

Table III presents the biases and variances of the regression coefficients in scenarios I and II. We see that the conditional logistic regression with counterfactuals produces regression coefficients that are approximately unbiased. We also see that the averages of the estimated variances closely match with the corresponding empirical variances of the estimates, indicating that the standard estimates of variances in the conditional logistic regression with counterfactuals are quite accurate.

Figure 1 (A: scenario I; B: scenario II) shows the (empirical) variances of the regression coefficients for the conditional logistic regression with counterfactuals (solid lines). It can be seen that as the number of control subjects increases, the

Table III. Biases and variances of the regression coefficients in the conditional logistic regression with counterfactuals.

	Scenario I: a null gene				Scenario II: an autosomal recessive gene			
	Number of control subjects				Number of control subjects			
Regression coefficient	250	500	750	1000	250	500	750	1000
Bias								
α_1	0.0096	0.0058	0.0050	0.0050	0.0066	0.0087	0.0100	0.0035
α_2	0.0037	−0.0028	−0.0010	−0.0018	0.0082	0.0048	0.0040	0.0016
β	0.0107	0.0060	0.0029	0.0055	0.0070	0.0106	0.0065	0.0075
γ_1	−0.0075	−0.0013	0.0001	−0.0026	−0.0018	−0.0063	−0.0035	−0.0023
γ_2	−0.0066	0.0032	0.0023	0.0023	−0.0034	0.0009	0.0022	−0.0005
Variance of the estimates								
α_1	0.0445	0.0411	0.0394	0.0389	0.0540	0.0500	0.0488	0.0478
α_2	0.0814	0.0659	0.0606	0.0576	0.0835	0.0672	0.0623	0.0589
β	0.0529	0.0453	0.0424	0.0405	0.0615	0.0545	0.0519	0.0493
γ_1	0.0541	0.0544	0.0545	0.0554	0.0688	0.0692	0.0683	0.0672
γ_2	0.0727	0.0746	0.0739	0.0738	0.0737	0.0740	0.0739	0.0722
Average of the estimated variances								
α_1	0.0448	0.0407	0.0393	0.0386	0.0541	0.0502	0.0488	0.0480
α_2	0.0813	0.0652	0.0598	0.0571	0.0832	0.0674	0.0620	0.0592
β	0.0528	0.0447	0.0420	0.0407	0.0621	0.0541	0.0515	0.0500
γ_1	0.0549	0.0548	0.0548	0.0548	0.0688	0.0689	0.0689	0.0687
γ_2	0.0734	0.0735	0.0734	0.0733	0.0731	0.0732	0.0733	0.0731

variances for the genetic (α_1 and α_2) and environmental (β) main effects decrease. On the other hand, the variances for the gene–environment interactions (γ_1 and γ_2) remain roughly constant, irrespective of the number of control subjects. In contrast, we see that the variances of the regression coefficients for the standard unconditional logistic regression (dotted lines) are much larger than those in the logistic regression with counterfactuals.

Table IV presents the type I error rates and powers of the conditional logistic regression with counterfactuals. We see that the type I error rates (underlined) are very close to the nominal significance level of 0.05. As for the powers, we see that as the number of control subjects increases, the powers for testing the genetic (α_2 for the autosomal recessive gene in scenario II) and the environmental (β) main effects increase. But the powers for testing the gene–environment interaction (γ_2 for the autosomal recessive gene in scenario II) remain roughly constant, irrespective of the number of control subjects. By comparison, we see that the powers of the regression coefficients for the standard unconditional logistic regression (in parenthesis) are lower as compared to those in the logistic regression with counterfactuals.

We also perform a case-only analysis based on the simulated 500 cases for the scenario II autosomal recessive gene. Again, the conditional logistic regression with counterfactuals produced approximately unbiased estimates for the γ_1 (bias$=-0.0047$), γ_2 (bias$=0.0014$) and η $(=\alpha_2-2\alpha_1)$ (bias$=-0.0171$) parameters. It is also notable that the variance of γ_1 $(=0.0699)$ and the variance of γ_2 $(=0.0759)$ are approximately equal to the corresponding values produced from a case–control analysis using the conditional logistic regression with counterfactuals in Figure 1B.

Discussion

In this paper, we employ the conditional logistic regression with counterfactuals for fitting the conditional-on-exposure likelihoods. An alternative is to use the multinomial logistic regression. The likelihood (3) can be treated as a 6-class multinomial logistic regression, and the likelihood (4), a 3-class multinomial logistic regression, both with class-specific intercepts and slopes. However, for the case–control study in (3), one needs to impose special constraints on the model parameters (detailed in Appendix D) and we know of no commercial statistical package that can impose such constraints. Therefore, we recommend using the counterfactual approach.

To accommodate an exposure variable, E, in any measurement scale, the disease risk model in (1) can be extended to

$$\log\left(\frac{\Pr(D=1|G,E)}{\Pr(D=0|G,E)}\right)=\mu+\alpha_1G_1+\alpha_2G_2+\boldsymbol{\beta}^{\mathbf{t}}\mathbf{f}(E)+\boldsymbol{\gamma}_1^{\mathbf{t}}\mathbf{f}_1(E)G_1+\boldsymbol{\gamma}_2^{\mathbf{t}}\mathbf{f}_2(E)G_2,$$

where $\mathbf{f}(\cdot)$, $\mathbf{f}_1(\cdot)$ and $\mathbf{f}_2(\cdot)$ are vector functions for the coding of E, and the $\boldsymbol{\beta}$, $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ are vectors of parameters. This is the most general formulation, as it allows the shape of the exposure risk to differ by genotype. The case–control

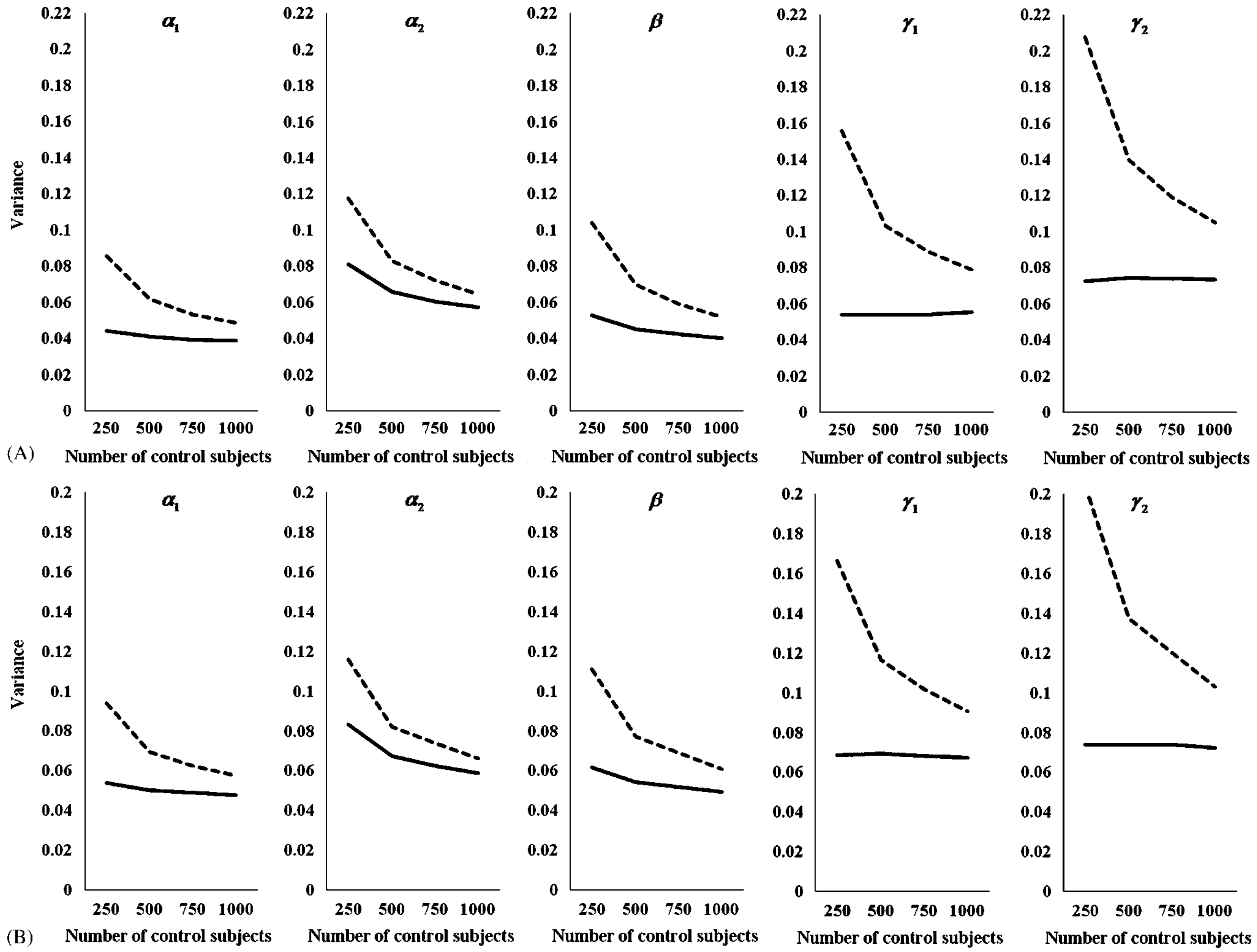

Figure 1. Variances of the regression coefficients for the conditional logistic regression with counterfactuals (solid lines) and the standard unconditional logistic regression (dotted lines): (A) scenario I: a null gene; (B) scenario II: an autosomal recessive gene.

Table IV. Type I error rates (underlined) and powers of the conditional logistic regression with counterfactuals. Shown in parenthesis are the powers of the standard unconditional logistic regression.

	Scenario I: a null gene				Scenario II: an autosomal recessive gene			
	Number of control subjects				Number of control subjects			
Regression coefficient	250	500	750	1000	250	500	750	1000
α_1	$\underline{0.0442}$	$\underline{0.0484}$	$\underline{0.0462}$	$\underline{0.0513}$	$\underline{0.0466}$	$\underline{0.0457}$	$\underline{0.0469}$	$\underline{0.0464}$
α_2	$\underline{0.0488}$	$\underline{0.0448}$	$\underline{0.0480}$	$\underline{0.0489}$	0.2970	0.3443	0.3712	0.3810
					(0.2212)	(0.2877)	(0.3240)	(0.3439)
β	0.8744	0.9191	0.9345	0.9464	0.8140	0.8702	0.8835	0.8992
	(0.5937)	(0.7591)	(0.8262)	(0.8713)	(0.5592)	(0.7188)	(0.7754)	(0.8211)
γ_1	$\underline{0.0461}$	$\underline{0.0496}$	$\underline{0.0481}$	$\underline{0.0511}$	$\underline{0.0491}$	$\underline{0.0517}$	$\underline{0.0474}$	$\underline{0.0465}$
γ_2	$\underline{0.0452}$	$\underline{0.0511}$	$\underline{0.0502}$	$\underline{0.0500}$	0.3263	0.3305	0.3333	0.3275
					(0.1399)	(0.1930)	(0.2291)	(0.2364)

likelihood in (3) becomes

$$\Pr(D,G|E,R=1)=\frac{\exp[G_1\log 2+\delta G+\mu^* D+\alpha_1 G_1 D+\alpha_2 G_2 D+\boldsymbol{\beta}^{\mathbf{t}}\mathbf{f}(E)D+\boldsymbol{\gamma}_1^{\mathbf{t}}\mathbf{f}_1(E)G_1 D+\boldsymbol{\gamma}_2^{\mathbf{t}}\mathbf{f}_2(E)G_2 D]}{\sum_{d=0}^{1}\sum_{g=0}^{2}\exp[g_1\log 2++\delta g+\mu^* d+\alpha_1 g_1 d+\alpha_2 g_2 d+\boldsymbol{\beta}^{\mathbf{t}}\mathbf{f}(E)d+\boldsymbol{\gamma}_1^{\mathbf{t}}\mathbf{f}_1(E)g_1 d+\boldsymbol{\gamma}_2^{\mathbf{t}}\mathbf{f}_2(E)g_2 d]}$$

and the case-only likelihood in (4) becomes

$$\Pr(G|E,R=1)=\frac{\exp[G_1\log 2+\delta^* G+\eta G_2+\boldsymbol{\gamma}_1^{\mathbf{t}}\mathbf{f}_1(E)G_1+\boldsymbol{\gamma}_2^{\mathbf{t}}\mathbf{f}_2(E)G_2]}{\sum_{g=0}^{2}\exp[g_1\log 2+\delta^* g+\eta g_2+\boldsymbol{\gamma}_1^{\mathbf{t}}\mathbf{f}_1(E)g_1+\boldsymbol{\gamma}_2^{\mathbf{t}}\mathbf{f}_2(E)g_2]}.$$

Note that mis-specifying $\mathbf{f}(\cdot)$, the main effect of E, will confound the case–control analysis [21]. However, the term cancels out from the case-only analysis. And an anonymous reviewer correctly pointed out that a hypothesis testing for gene–environment interaction in a case-only study cannot be biased by mis-specification of $\mathbf{f}(\cdot)$, though getting $\mathbf{f}_1(\cdot)$ and $\mathbf{f}_2(\cdot)$ wrong could affect power.

Both the case–control study and the case-only study are vulnerable to 'population stratification biases' [22–24]. The method proposed in this paper can be extended to the situation when the study population is not a homogeneous one but instead is composed of several population strata. The assumptions of GE independence and HWE hold within each stratum, but do not hold in the study population taken as a whole. Let $S_1,\ldots,S_q$ (indexed by i) be the stratum indicators for a subject, such that S_i is 1, if the subject belongs to the ith stratum, and 0, otherwise. For stratum adjustment, we need to condition on $S_1,\ldots,S_q$ (if these information are available) as well as on E. For case–control study, this becomes:

$$\Pr(D,G|S_1,\ldots,S_q,E,R=1)=\frac{\exp(G_1\log 2+\sum_i\delta_i S_i G+\sum_i\mu_i^* S_i D+\alpha_1 G_1 D+\alpha_2 G_2 D+\beta ED+\gamma_1 G_1 ED+\gamma_2 G_2 ED)}{\sum_{d=0}^{1}\sum_{g=0}^{2}\exp(g_1\log 2+\sum_i\delta_i S_i g+\sum_i\mu_i^* S_i d+\alpha_1 g_1 d+\alpha_2 g_2 d+\beta Ed+\gamma_1 g_1 Ed+\gamma_2 g_2 Ed)}.$$

The additional 'stratum×gene' interaction term allows the allele frequency odds to vary between strata, and the additional 'stratum×disease' interaction terms, the background disease odds to vary. And for the case-only study, it becomes:

$$\Pr(G|S_1,\ldots,S_q,E,R=1)=\frac{\exp(G_1\log 2+\sum_i\delta_i^* S_i G+\eta G_2+\gamma_1 G_1 E+\gamma_2 G_2 E)}{\sum_{g=0}^{2}\exp(g_1\log 2+\sum_i\delta_i^* S_i g+\eta g_2+\gamma_1 g_1 E+\gamma_2 g_2 E)}.$$

There is only one additional term: the 'stratum×gene' interaction term. The above 'conditional-on-stratum-and-exposure' analyses can also be implemented using the conditional logistic regression with counterfactuals. The exposure variables and the stratum indicators of these counterfactual subjects are set to be exactly the same as the recruited factual subject himself/herself.

In this paper, we impose dual assumptions of GE independence and HWE. With simple modifications, one can impose only the GE independence but not the HWE. The case–control likelihood in (3) is to be modified as

$$\Pr(D,G|E,R=1)=\frac{\exp(\varepsilon_1 G_1+\varepsilon_2 G_2+\mu^* D+\alpha_1 G_1 D+\alpha_2 G_2 D+\beta ED+\gamma_1 G_1 ED+\gamma_2 G_2 ED)}{\sum_{d=0}^{1}\sum_{g=0}^{2}\exp(\varepsilon_1 g_1+\varepsilon_2 g_2+\mu^* d+\alpha_1 g_1 d+\alpha_2 g_2 d+\beta Ed+\gamma_1 g_1 Ed+\gamma_2 g_2 Ed)},$$

where $\varepsilon_1=\log[\Pr(G=1|D=0)/\Pr(G=0|D=0)]$ and $\varepsilon_2=\log[\Pr(G=2|D=0)/\Pr(G=0|D=0)]$. Compared to (3) where both assumptions were made, this likelihood has one extra number of parameters. The case-only likelihood in (4) is to be modified as

$$\Pr(G|E,R=1)=\frac{\exp(\varepsilon_1^* G_1+\varepsilon_2^* G_2+\gamma_1 G_1 E+\gamma_2 G_2 E)}{\sum_{g=0}^{2}\exp(\varepsilon_1^* g_1+\varepsilon_2^* g_2+\gamma_1 g_1 E+\gamma_2 g_2 E)},$$

where $\varepsilon_1^*=\varepsilon_1+\alpha_1$ (ε_1 and α_1 are confounded) and $\varepsilon_2^*=\varepsilon_2+\alpha_2$ (ε_2 and α_2 are confounded). Note that without imposing the HWE assumption, we are no longer able to estimate the η parameter ($\eta=\alpha_2-2\alpha_1$) that measures the departure from a multiplicative gene-dose model.

For brevity, we only consider one single exposure in this paper. Extension to multiple exposures is straightforward. If every exposure is independent of the study gene, we can simply condition on all of them simultaneously (as exposure variables). If there is GE independence for some of the exposures but not for the others, we still condition on all of them. In the conditioning, those GE independent exposures are treated as the exposure variables, but the others are treated as 'stratification variables' (see above). That is, we allow the allele frequency odds to vary across the various strata defined by the cross-classification of these GE non-independent exposures.

The unconditional logistic regression has been the mainstay for analysis of case–control data for nearly half a century. On the other hand, it has also been known for long that its efficiency may be less than ideal, because it disregards completely the information contained in the distribution of covariates themselves (e.g. GE independence and HWE) when in fact it may be informative for the model parameters [25]. To exploit the covariates information of GE independence and/or HWE for the analysis of case–control data, the recent decade witnessed a plethora of methods which preach the conditional-on-exposure principle [6–9]. This paper is not a theoretical advance over previous works. Nevertheless, we do present herein an easy-to-implement counterfactual approach, which offers the flexibility for the complex modeling yet remains well within the reach to the practicing epidemiologists.

As compared to the case–control study, the case-only design for studying gene–environment interactions is a recent invention [3, 4, 10]. Attempts have been made to cast the case-only study in a regression framework [10, 26, 27]. However, those studies considered the GE independence assumption only and did not say what to do with the HWE assumption in a case-only study. Lee [28] on the other hand assumed HWE and proposed to use the case-only study as a gene hunting tool (by testing whether or not $\psi_2=\psi_1^2$). However, he did not consider the environmental exposures.

In this paper, we present a method to analyze case–control and case-only studies assuming GE independence and HWE with the readily available statistical packages. When both assumptions are met, the method (conditional logistic regression with counterfactuals) is unbiased and has the correct type I error rates. It also results in smaller variances and achieves higher powers as compared to using conventional analysis (unconditional logistic regression). However, when the assumptions are not met, conditional logistic regression with counterfactuals will be biased. In the analysis of Sam et al.'s data, we tested the assumptions of GE independence ($p=0.7563$) and HWE ($p=0.6789$) in the control subjects before applying the conditional logistic regression with counterfactuals. However, these *ad hoc* empirical checks may lack adequate power to support (or reject) the assumptions of GE independence and HWE. Recently, a number of researchers [29–32] have been working on methods that can enjoy the best of the two worlds, that is, to have smaller variances when the assumptions are met (efficiency), and are less biased when the assumptions are not (robustness). However, their methods are computationally much more involved and are beyond the scope of this paper.

Appendix A: Derivation of case–control and case-only likelihoods enforcing GE independence and HWE

From (2), we have,

$$\begin{aligned}\Pr(D=0, G|E) &= \Pr(D=0|E)\times\Pr(G|E, D=0)\\ &= \Pr(D=0|E)\times\Pr(G=0|D=0)\times\exp(G_1\log 2+\delta G).\end{aligned}$$

From (1), we have,

$$\Pr(D=1, G|E)=\Pr(D=0, G|E)\times\exp(\mu+\alpha_1 G_1+\alpha_2 G_2+\beta E+\gamma_1 G_1 E+\gamma_2 G_2 E).$$

And therefore,

$$\begin{aligned}\Pr(D=1, G|E) &= \Pr(D=0|E)\times\Pr(G=0|D=0)\\ &\quad\times\exp(G_1\log 2+\delta G+\mu+\alpha_1 G_1+\alpha_2 G_2+\beta E+\gamma_1 G_1 E+\gamma_2 G_2 E).\end{aligned}$$

Since both $\Pr(D=0, G|E)$ and $\Pr(D=1, G|E)$ have the same multiplier: $\Pr(D=0|E)\times\Pr(G=0|D=0)$, we see that

$$\Pr(D, G|E)=\frac{\exp(G_1\log 2+\delta G+\mu D+\alpha_1 G_1 D+\alpha_2 G_2 D+\beta ED+\gamma_1 G_1 ED+\gamma_2 G_2 ED)}{\sum_{d=0}^{1}\sum_{g=0}^{2}\exp(g_1\log 2+\delta g+\mu d+\alpha_1 g_1 d+\alpha_2 g_2 d+\beta Ed+\gamma_1 g_1 Ed+\gamma_2 g_2 Ed)}.$$

Because subject recruitment in a case–control study is dependent only on disease status but not on gene or exposure, we have $\Pr(R=1|D=1, G, E)=\Pr(R=1|D=1)$ and $\Pr(R=1|D=0, G, E)=\Pr(R=1|D=0)$. It is easy to further show that among the subjects recruited for study, the above conditional-on-exposure probability becomes

$$\Pr(D, G|E, R=1)=\frac{\exp(G_1\log 2+\delta G+\mu^* D+\alpha_1 G_1 D+\alpha_2 G_2 D+\beta ED+\gamma_1 G_1 ED+\gamma_2 G_2 ED)}{\sum_{d=0}^{1}\sum_{g=0}^{2}\exp(g_1\log 2+\delta g+\mu^* d+\alpha_1 g_1 d+\alpha_2 g_2 d+\beta Ed+\gamma_1 g_1 Ed+\gamma_2 g_2 Ed)},$$

where $\mu^*=\mu+\log[\Pr(R=1|D=1)/\Pr(R=1|D=0)]$. For a case-only study, we have

$$\begin{aligned}\Pr(G|E, R=1) &= \Pr(D=1, G|E, R=1)\\ &= \frac{\Pr(D=1, G|E)\times\Pr(R=1|D=1, G, E)}{\sum_{g=0}^{2}\Pr(D=1, g|E)\times\Pr(R=1|D=1, g, E)}\end{aligned}$$

$$= \frac{\dfrac{\Pr(D=1,G|E)}{\Pr(D=0|E)\times\Pr(G=0|D=0)}\times\Pr(R=1|D=1)}{\sum_{g=0}^{2}\left[\dfrac{\Pr(D=1,g|E)}{\Pr(D=0|E)\times\Pr(G=0|D=0)}\times\Pr(R=1|D=1)\right]}$$

$$= \frac{\exp(G_1\log 2+\delta G+\mu+\alpha_1 G_1+\alpha_2 G_2+\beta E+\gamma_1 G_1 E+\gamma_2 G_2 E)}{\sum_{g=0}^{2}\exp(g_1\log 2+\delta g+\mu+\alpha_1 g_1+\alpha_2 g_2+\beta E+\gamma_1 g_1 E+\gamma_2 g_2 E)}$$

$$= \frac{\exp(G_1\log 2+\delta^* G+\eta G_2+\gamma_1 G_1 E+\gamma_2 G_2 E)}{\sum_{g=0}^{2}\exp(g_1\log 2+\delta^* g+\eta g_2+\gamma_1 g_1 E+\gamma_2 g_2 E)},$$

where $\delta^*=\delta+\alpha_1$ and $\eta=\alpha_2-2\alpha_1$.

Appendix B: SAS codes for case–control study

```
/**     Prepare your SAS data set (data=casecontrol)     **/
/**     that contains variables: ID, G, E, and D         **/
/**     ID: subject identification                       **/
/**     G=0,1,2: the number of variant alleles           **/
/**     E: exposure status                               **/
/**     D: disease status                                **/


data counterfactual;
   set casecontrol;
   do temp_i=0 to 5;
      R=(temp_i=0);
      temp_g=mod((G+temp_i), 3);
      temp_d=mod((D+temp_i), 2);
      output;
      end;
run;


data counterfactual;
   set counterfactual;
   G=temp_g; D=temp_d;
   G1=(G=1); G2=(G=2);
   sur=1-R;
   G1D=G1*D; G2D=G2*D;
   ED=E*D;
   G1ED=G1*E*D; G2ED=G2*E*D;
   offset=G1*log(2);
   drop temp_g temp_d temp_i;
   run;


/* Conditional logistic regression with counterfactuals (cases and controls)*/
proc phreg data=counterfactual nosummary;
   model sur*R(0)=D G G1D G2D ED G1ED G2ED/offset=offset ;
   strata ID;
   run;
```

Appendix C: SAS codes for case-only study

```
/**     Prepare your SAS data set (data=caseonly)     **/
/**     that contains variables: ID, G, and E          **/
/**     ID: subject identification                     **/
/**     G=0,1,2: the number of variant alleles         **/
/**     E: exposure status                             **/


data counterfactual;
   set caseonly;
   do temp_i=0 to 2;
      R=(temp_i=0);
      temp_g=mod((G+temp_i), 3);
      output;
      end;
run;


data counterfactual;
   set counterfactual;
   G=temp_g;
   G1=(G=1); G2=(G=2);
   sur=1-R;
   G1E=G1*E; G2E=G2*E;
   offset= G1*log(2);
   drop temp_g temp_i;
   run;


/* Conditional logistic regression with counterfactuals (cases only)*/
proc phreg data=counterfactual nosummary;
   model sur*R(0)=G G2 G1E G2E/offset=offset;
   strata ID;
   run;
```

Appendix D: Multinomial logistic regression

For the case–control study, the conditional-on-exposure likelihood involves six classes defined by the cross-classification of disease status ($D=0$ or 1) and genotypes ($G=0$, 1, or 2). Define the class indicator as $k=3D+G$ ($k=0,1,\ldots,5$) and the multinomial likelihood as

$$\Pr(\text{class}=k|E)=\frac{\exp(G_1\log 2+a_k+b_kE)}{\sum_{j=0}^{5}\exp(G_1\log 2+a_j+b_jE)}.$$

Comparing this with (3) in the text, we found that for the two likelihoods to be exactly the same, we need to impose the following two sets of special constraints for this multinomial likelihood: (1) $b_1=b_2=0$ (for GE independence among the controls); and (2) $a_2=2a_1$ (for HWE among the controls), in addition to the usual constraints of $a_0=b_0=0$. Also, we found that the parameters in these two likelihoods are related through the following equations: $\delta=a_1$, $\mu^*=a_3$, $\alpha_1=a_4-a_1-a_3$, $\alpha_2=a_5-a_2-a_3$, $\beta=b_3$, $\gamma_1=b_4-b_3$, and $\gamma_2=b_5-b_3$. For the case-only study, define the class indicator as $l=G$ ($l=0,1,2$) and the multinomial likelihood as

$$\Pr(\text{class}=l|E)=\frac{\exp(G_1\log 2+a_l+b_lE)}{\sum_{j=0}^{2}\exp(G_1\log 2+a_j+b_jE)}.$$

Besides the usual $a_0=b_0=0$, no special constraint is needed this time. (The case-only study does not have a control group upon which to impose the two assumptions of GE independence and HWE.) This multinomial likelihood is exactly the same as (4) in the text, with the following relationships: $\delta^*=a_1$, $\eta=a_2-2a_1$, $\gamma_1=b_1$, and $\gamma_2=b_2$.

Acknowledgement

This paper is partly supported by grants from the National Science Council, Taiwan, Republic of China.

References

1. Hunter DJ. Gene–environment interactions in human diseases. *Nature Reviews Genetics* 2005; **6**:287–298.
2. Olden K. Commentary: From phenotype, to genotype, to gene–environment interaction and risk for complex diseases. *International Journal of Epidemiology* 2007; **36**:18–20.
3. Begg CB, Zhang ZF. Statistical analysis of molecular epidemiology studies employing case-series. *Cancer Epidemiology Biomarkers and Prevention* 1994; **3**:173–175.
4. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case–control studies. *Statistics in Medicine* 1994; **13**:153–162.
5. Sham P. *Statistics in Human Genetics*. Oxford University Press Inc.: New York, 1998; 39–43.
6. Umbach DM, Weinberg CR. Designing and analyzing case–control studies to exploit independence of genotype and exposure. *Statistics in Medicine* 1997; **16**:1731–1743.
7. Chatterjee M, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene–environment independence in case–control studies. *Biometrika* 2005; **92**:399–418.
8. Cheng KF, Lin WJ. Retrospective analysis of case–control studies when the population is in Hardy–Weinberg equilibrium. *Statistics in Medicine* 2005; **24**:3289–3310.
9. Chen YH, Kao JT. Multinomial logistic regression approach to haplotype association analysis in population-based case–control studies. *BMC Genetics* 2006; **7**:43.
10. Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. Limitations of the case-only design for identifying gene–environment interaction. *American Journal of Epidemiology* 2001; **154**:687–693.
11. Self SG, Longton G, Kopecky KJ, Liang KY. On estimating HLA/disease association with application to a study of aplastic anemia. *Biometrics* 1991; **47**:53–61.
12. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *American Journal of Epidemiology* 1991; **133**:144–153.
13. Farrington CP, Nash J, Miller E. Case series analysis of adverse reactions to vaccines: a comparative evaluation. *American Journal of Epidemiology* 1996; **143**:1165–1173.
14. Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene–environment interaction: case–control studies with no controls! *American Journal of Epidemiology* 1996; **144**:207–213.
15. Zaffanella LE, Savitz DA, Greenland S, Ebi KL. The residential case-specular method to study wire codes, magnetic fields, and disease. *Epidemiology* 1998; **9**:16–20.
16. Greenland S. A unified approach to the analysis of case-distribution (case-only) studies. *Statistics in Medicne* 1999; **18**:1–15.
17. Lee WC. Genetic association studies of adult-onset diseases using the case-spouse and case-offspring designs. *American Journal of Epidemiology* 2003; **158**:1023–1032.
18. Cordell HJ, Barratt BJ, Clayton DG. Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene–gene and gene–environment interactions, and parent-of-origin effects. *Genetic Epidemiology* 2004; **26**:167–185.
19. Lee WC, Chang CH. Assessing effects of disease genes and gene–environment interactions using the case-spouse design. *Journal of Epidemiology and Community Health* 2006; **60**:683–685.
20. Sam SS, Thomas V, Reddy SK, Surianarayanan G, Chandrasekaran A. CYP1A1 polymorphisms and the risk of upper aerodigestive tract cancers in an Indian population. *Head and Neck* 2008; **30**:1566–1574.
21. Greenland S. Basic problems in interaction assessment. *Environmental Health Perspectives* 1993; **101**(S4):59–66.
22. Lee WC, Wang LY. Simple formulas for gauging the potential impacts of population-stratification bias. *American Journal of Epidemiology* 2008; **167**:86–89.
23. Wang LY, Lee WC. Population stratification bias in case-only study for gene–environment interactions. *American Journal of Epidemiology* 2008; **168**:197–201.
24. Lee WC, Wang LY. Reducing population stratification bias: stratum matching is better than exposure. *Journal of Clinical Epidemiology* 2009; **62**:62–66.
25. Breslow NE, Day NE. *Statistical Methods in Cancer Research*, *the Analysis of Case–Control Studies*, vol. 1. International Agency for Research on Cancer: Lyon, 1980; 202–205.
26. Gatto NM, Campbell UB, Rundle AG, Ahsan H. Further development of the case-only design for assessing gene–environment interaction: evaluation of and adjustment for bias. *International Journal of Epidemiology* 2004; **33**:1014–1024.
27. Cheng KF. A maximum likelihood method for studying gene–environment interactions under conditional independence of genotype and exposure. *Statistics in Medicine* 2006; **25**:3093–3109.
28. Lee WC. Searching for disease-susceptibility loci by Hardy–Weinberg disequilibrium in a gene bank of affected individuals. *American Journal of Epidemiology* 2003; **158**:397–400.
29. 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**:685–694.
30. Mukherjee B, Ahn J, Gruber SB, Rennert G, Moreno V, Chatterjee N. Tests for gene–environment interaction from case–control data: a novel study of type I error, power and designs. *Genetic Epidemiology* 2008; **32**:615–626.
31. Chen YH, Chatterjee N, Carroll RJ. Shrinkage estimators for robust and efficient inference in haplotype-based case–control studies. *Journal of the American Statistical Association* 2009; **104**:220–233.
32. Li D, Conti DV. Detecting gene–environment interactions using a combined case-only and case–control approach. *American Journal of Epidemiology* 2009; **169**:497–504.