American Journal of Epidemiology
© The Author 2010. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

Vol. 173, No. 2
DOI: 10.1093/aje/kwq352
Advance Access publication:
November 17, 2010

Practice of Epidemiology

Genetic Association and Gene-Environment Interaction: A New Method for Overcoming the Lack of Exposure Information in Controls

Rémi Kazma*, **Marie-Claude Babron**, and **Emmanuelle Génin**

* Correspondence to Dr. Rémi Kazma, Institut National de la Santé et de la Recherche Médicale, U946, 27 rue Juliette Dodu, 75010 Paris, France (e-mail: remi.kazma@inserm.fr).

Initially submitted January 25, 2010; accepted for publication September 17, 2010.

The use of a reference control panel in genome-wide association studies is an interesting solution to the problem of how to reduce costs. In such designs, data on relevant environmental factors are usually collected only in cases, making it more difficult to deal with potential gene-environment interactions when testing for genetic association. However, under certain circumstances, neglecting an existing interaction with the environment may be detrimental in terms of statistical power to detect the genetic factor. In this paper, the authors propose a novel method based on a multinomial logistic regression model to overcome the lack of environmental exposure information in controls, by contrasting both exposed and unexposed cases with the control sample. For each case group, a genetic effect-size parameter is estimated, and the genetic association and the gene-environment interaction are tested jointly. The authors evaluate the performance of this method through asymptotic computations and simulations of cases and population controls under different models. In the presence of a gene-environment interaction, this approach outperforms other available methods that test for genetic association and gene-environment interaction either separately or jointly. Interestingly, it even has better power than the joint test requiring full knowledge of the environmental information in both cases and controls.

epidemiologic research design; genotype-environment interaction; genetic predisposition to disease; logistic models; multinomial model; partial information; population control

Abbreviations: LRT, likelihood ratio test; OR, odds ratio.

Most prevalent human diseases (cancer, cardiovascular disease, asthma, neuropsychiatric disorders, diabetes, etc.) are associated with multiple genetic and environmental factors acting jointly rather than independently (1). With the advent of human genome sequencing and the International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/), providing the distribution patterns of single nucleotide polymorphisms across the whole genome, a plethora of genome-wide association scans have been conducted, leading to the successful identification of over 100 association signals (2). In the last few years, huge samples of genotypes from various populations around the world have been collected. These genotype data, which are usually available to the scientific community, can serve as control data in association studies. Using such samples as reference control panels is a tempting solution for reducing costs and has already been implemented in large consortium studies, where the same individuals serve as controls against several sets of cases with different diseases (3–7). A drawback of this ''universal'' control design is that data on the relevant environmental factors for a specified disease are usually collected only in cases and not in controls, making it difficult to address the question of gene-environment ($G \times E$) interaction.

When dealing with $G \times E$ interaction in genetic studies of complex diseases, 2 complementary strategies can be outlined. The first strategy consists of simultaneously testing for genetic association and $G \times E$ interaction, arguing that neglecting an existing interaction with an environmental factor may hinder detection of the genetic factor (8). Kraft et al. (9) have shown that, under a wide variety of $G \times E$ interaction models, such a combined test outperforms a marginal association test in terms of statistical power. However, this strategy requires collecting exposure data on both cases and controls. The second strategy consists of testing for $G \times E$ interaction alone, arguing that after an initial genome-wide association scan in which genetic effects are considered alone, such an approach would be independent and would identify new genetic variants involved in pure $G \times E$ interaction (10). In that context, a logistic regression-based case-control design can model $G \times E$ interaction explicitly, but it requires full information on genetic and environmental data for both samples. When exposure information is available only for cases, a case-only design can be implemented by regressing the genotype (or allele) counts on the environmental variable (11, 12). If the gene and the environmental factor are independent at the population level, the case-only design can be more powerful in detecting $G \times E$ interaction than the case-control design. However, it can only estimate and test for the $G \times E$ interaction term (not the genetic and environmental main effects), and it is prone to false-positive findings or power loss should a gene-environment correlation exist in the underlying population. Regardless of the choice to test for $G \times E$ interaction or not, most methods in genetic epidemiology are based on the binomial form of the logistic regression model, where the outcome is binary and the dependent covariates are observed in all subjects. The absence of environmental information in controls makes such methods unfeasible in practice. Development of an alternative approach in such situations would provide new perspectives in the search for $G \times E$ interaction studies at the genome-wide level.

Here we present a novel method based on a multinomial logistic regression model that contrasts both exposed and unexposed case samples with a control sample for which there is no information on the environmental exposure. We evaluate the performance of this approach and compare it with the different genetic-association and $G \times E$-interaction tests available under various scenarios of $G \times E$ interaction.

MATERIALS AND METHODS

Disease penetrance model and notations

Let us consider a disease phenotype D, a genetic risk factor G, and an environmental risk factor E. The 3 variables are assumed to be dichotomous, 0 denoting absence and 1 presence of the disease or risk factor. The genetic factor is an autosomal biallelic genetic locus with a susceptibility allele S and population frequency q. Dominant genetic effects are modeled here as in the paper by Kraft et al. (9). Therefore, the variable G corresponds to having at least 1 copy of the S allele, and the frequency of the susceptibility genotypes becomes $f_G = q^2 + 2q(1 - q)$ under the assumption that genotypes are in Hardy-Weinberg proportions in the population. The independent effect of G is measured by the genotypic odds ratio (OR_G), which is equal to the ratio of the odds of G in cases and controls, all being unexposed:

$$\mathrm{OR}_G = \frac{P(G=1|D=1,E=0)/P(G=0|D=1,E=0)}{P(G=1|D=0,E=0)/P(G=0|D=0,E=0)}. \tag{1}$$

The independent effect of E is measured by the environmental odds ratio (OR_E), which is equal to the ratio of the odds of E in cases and controls, all being noncarriers of G:

$$\mathrm{OR}_E = \frac{P(E=1|D=1,G=0)/P(E=0|D=1,G=0)}{P(E=1|D=0,G=0)/P(E=0|D=0,G=0)}. \tag{2}$$

A possible gene-environment correlation is introduced by considering a coefficient θ_{GE}, as in the paper by Lindström et al. (13):

$$\theta_{GE} = \frac{P(E=1|G=1)/P(E=0|G=1)}{P(E=1|G=0)/P(E=0|G=0)}. \tag{3}$$

The base probability of $E = 1$ given $G = 0$ is denoted f_E. When $\theta_{GE} = 1$ (independence between G and E), $P(E = 1|G = 1) = P(E = 1|G = 0) = f_E$. The gene-environment correlation θ_{GE} measuring the degree of association between G and E in the population can thus vary between 0 and ∞.

Considering the joint distributions of G and E, the odds ratio associated with the presence of both G and E is measured by OR_{GE}:

$$\mathrm{OR}_{GE} = \frac{P(E=1,G=1|D=1)/P(E=0,G=0|D=1)}{P(E=1,G=1|D=0)/P(E=0,G=0|D=0)}. \tag{4}$$

In the absence of $G \times E$ interaction and under a multiplicative model on the odds scale (or an additive model on the log odds scale), OR_{GE} is expected to be equal to the product of OR_G and OR_E. Using equations 1, 2, and 4, the interaction term OR_I measuring the departure from this condition is expressed as

$$\mathrm{OR}_I = \frac{\mathrm{OR}_{GE}}{\mathrm{OR}_G \times \mathrm{OR}_E}. \tag{5}$$

Binomial logistic regression models

Different binomial logistic regression approaches are commonly used to test for association or $G \times E$ interaction. These approaches differ by the amount of exposure information they require and the hypotheses they test for (Table 1). Here, we first describe 2 methods that test for the G effect only, then 2 methods that test for $G \times E$ interaction only, and finally 1 method that tests simultaneously for G and $G \times E$ interaction.

Marginal genetic association test (referred to as marginal-G). When information on E is available in neither cases nor controls, we can only model the marginal effect of G as follows:

Table 1. Summary of Available Methods for Different Patterns of Exposure and Genotype Data Availability

Tested Effect	Data Availability in Cases		Data Availability in Controls		Logistic Model	Null Hypothesis	Degrees of Freedom	Notation in Figures and Tables
	Genotype	Exposure	Genotype	Exposure				
Genetic effect alone	Yes	No	Yes	No	Binomial	β_{GM}[a] $= 0$	1	Marginal-G
	Yes	Yes	Yes	Yes	Binomial	β_{GA}[b] $= 0$	1	Adjusted-G
Gene-environment (*G* × *E*) interaction alone	Yes	Yes	Yes	Yes	Binomial	β_{ICC}[c] $= 0$	1	Case-control-I
	Yes	Yes	No	No	Binomial	β_{ICO}[d] $= 0$	1	Case-only-I
Both genetic effect and *G* × *E* interaction	Yes	Yes	Yes	Yes	Binomial	β_{GCC}[c] $= 0$, β_{ICC}[c] $= 0$	2	Binomial-GI
	Yes	No	Yes	Yes	Multinomial	β_{G1}[e] $= 0$, β_{G2}[e] $= 0$	2	Multinomial-GI

[a] β_{GM}, genetic regression coefficient of the marginal genetic effect model (equation 6).

[b] β_{GA}, genetic regression coefficient of the adjusted model (equation 7).

[c] β_{GCC} and β_{ICC}, genetic and *G* × *E* interaction regression coefficients of the full model (equation 8).

[d] β_{ICO}, *G* × *E* interaction regression coefficient of the case-only model (equation 9).

[e] β_{G1} and β_{G2}, regression coefficients of the multinomial model (equation 10).

$$\operatorname{logit} P(D = 1 | G) = \beta_0 + \beta_{GM} G. \tag{6}$$

The null hypothesis of no association between G and D ($\beta_{GM} = 0$) is tested with a 1-df likelihood ratio test (LRT).

Adjusted genetic association test (referred to as adjusted-G). When E is available in both cases and controls, it is possible to adjust on E without accounting for $G \times E$ interaction. The adjusted model is expressed as

$$\operatorname{logit} P(D = 1 | G, E) = \beta_0 + \beta_E E + \beta_{GA} G. \tag{7}$$

The null hypothesis of no association between G and D ($\beta_{GA} = 0$) is tested with a 1-df LRT.

G × E interaction test in a case-control design (referred to as case-control-I). Again, when information on E is available in both cases and controls, the full logistic model is expressed as

$$\operatorname{logit} P(D = 1 | G, E) = \beta_0 + \beta_E E + \beta_{GCC} G + \beta_{ICC} GE. \tag{8}$$

The null hypothesis of no $G \times E$ interaction ($\beta_{ICC} = 0$) is tested with a 1-df LRT.

G × E interaction test in a case-only design (referred to as case-only-I). Piegorsch et al. (12) present this design as a more powerful alternative to test for $G \times E$ interaction when the 2 factors are independent in the population. The procedure consists of considering exposure the dependent variable on which the genotype is regressed. The corresponding case-only logistic model is expressed as

$$\operatorname{logit} P(E = 1 | G) = \beta_0 + \beta_{ICO} G. \tag{9}$$

The $G \times E$ interaction ($\beta_{ICO} = 0$) is tested with a 1-df LRT.

Combined genetic and G × E interaction test (referred to as binomial-GI). Using the full logistic model (equation 8), Kraft et al. (9) suggest jointly testing the association between G and D and the $G \times E$ interaction with a combined 2-df LRT of the null hypothesis: $\beta_{GCC} = \beta_{ICC} = 0$. This test requires information on E in both cases and controls.

Multinomial logistic regression model

The multinomial logistic regression model is a logistic regression model extended for a nominal outcome variable with more than 2 categories (14, 15). Widely used in traditional epidemiologic studies, it has recently been shown to be a powerful approach in genetic studies of disease subphenotypes (16). It is usually used to contrast subcategories of phenotypes against the control group and can therefore be extended to our particular context, where environmental information is not collected in the control group. For that purpose, the D and E dichotomous variables are combined into a single 3-class multinomial variable D_M taking on the value of 0 in controls, 1 in unexposed cases, and 2 in exposed cases. The multinomial model can be defined as

$$\begin{aligned} \operatorname{logit} P(D_M = j | G) &= \log[P(D_M = j | G)/P(D_M = 0 | G)] \\ &= \beta_{0j} + \beta_{Gj} G, \end{aligned} \tag{10}$$

with j taking the values 1 and 2. The 2 corresponding logit equations are used simultaneously to estimate the 2 genetic parameters, β_{G1} and β_{G2}, that maximize the overall likelihood. The β_{G1} and β_{G2} parameters represent the genetic odds ratios in unexposed and exposed cases, respectively, as compared with the whole control group. A combined test (referred to as multinomial-GI) of the genetic association and $G \times E$ interaction can be carried out by testing the null hypothesis $\beta_{G1} = \beta_{G2} = 0$ with a 2-df LRT.

Computations, simulations, and evaluation criteria

We modeled the disease probability conditional on G and E using a logit function as follows:

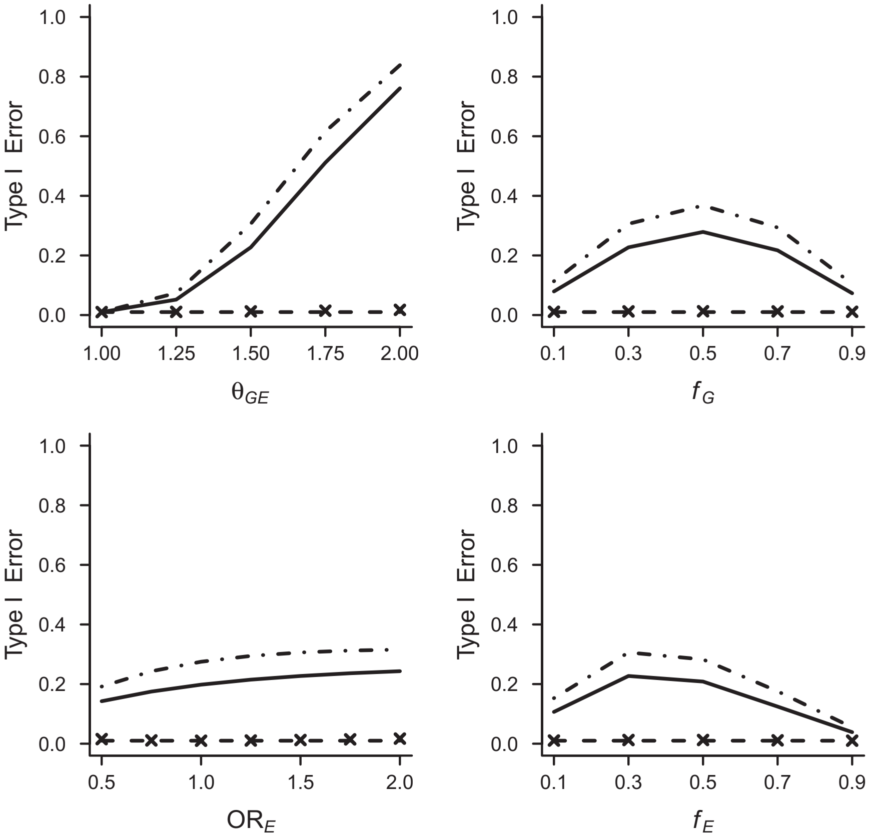

Figure 1. Type I error rates as a function of gene-environment correlation, the frequency of risk genotypes, the environmental main effect, and the frequency of environmental exposure. The base model from which these parameters vary is the following: environmental odds ratio (OR), $OR_E = 1.5$; environmental exposure frequency, $f_E = 0.3$; dominant genetic model with genetic odds ratio, $OR_G = 1$; risk genotype frequency, $f_G = 0.3$; gene-environment interaction odds ratio, $OR_I = 1$; gene-environment correlation, $\theta_{GE} = 1.5$; disease prevalence, $f_D = 0.1$; nominal type I error rate of 0.01, 2-sided test; sample of 500 cases and 500 controls. The marginal-G test is represented by crosses (×), the case-only-I test is represented by alternate dots and dashes (· – · – ·), the binomial-GI test is represented by dashed lines (– – –), and the multinomial-GI test is represented by solid lines (——).

$$\operatorname{logit} P(D = 1 | G, E) = \beta_0 + \beta_E E + \beta_G G + \beta_I GE, \quad (11)$$

where $\beta_0 = \log(B/1 - B)$, with B being the baseline risk of disease (i.e., the probability of disease given $G = 0$ and $E = 0$), $\beta_G = \log(OR_G)$, $\beta_E = \log(OR_E)$, and $\beta_I = \log(OR_I)$.

Using Bayes' theorem and equations 1–3, 5, and 11, the expected probabilities of all of the categories of $G = i$ and $E = j$ conditional on $D = k$ are

$$\begin{aligned} P(G = i, E = j | D = k) = {} & P(D = k | G = i, E = j) \\ & \times P(G = i) \times P(E = j | G = i) \\ & / P(D = k), \quad (12) \end{aligned}$$

where $P(D = k | G = i, E = j)$ is derived from equation 11 and $P(E = j | G = i)$ is a function of θ_{GE}, f_E, and f_G. The disease prevalence f_D was set to 0.1 and, to mimic a reference control panel design, a misclassification bias was considered in controls, assuming that 10% (f_D) of the controls are in fact affected.

We considered values of f_G and f_E ranging from 0.1 to 0.9 in intervals of 0.2, values of OR_G, OR_E, and OR_I ranging from 0.5 to 2.0 (to 3.0 for OR_I) in intervals of 0.25, and values of θ_{GE} ranging from 1.0 to 2.0 in intervals of 0.25. For each set of parameters, 1,000 samples of 500 cases and 500 controls were simulated.

Asymptotic type I error and power were estimated through the use of noncentral χ^2 distributions with corresponding degrees of freedom at a nominal type I error rate of 0.01. They were also estimated by means of the proportion of simulated replicates with a P value less than or equal to 0.01. Since both asymptotic and simulation-based results were similar, only asymptotic results are reported. The squared bias, the variance, and the coverage probability (probability that the 95% confidence interval of the estimates contained the theoretical value) of the different parameters were also computed for the simulated data sets.

All computations and simulations were carried out in R (17), and statistical analyses were conducted using the "logit" and "mlogit" functions of Stata (18).

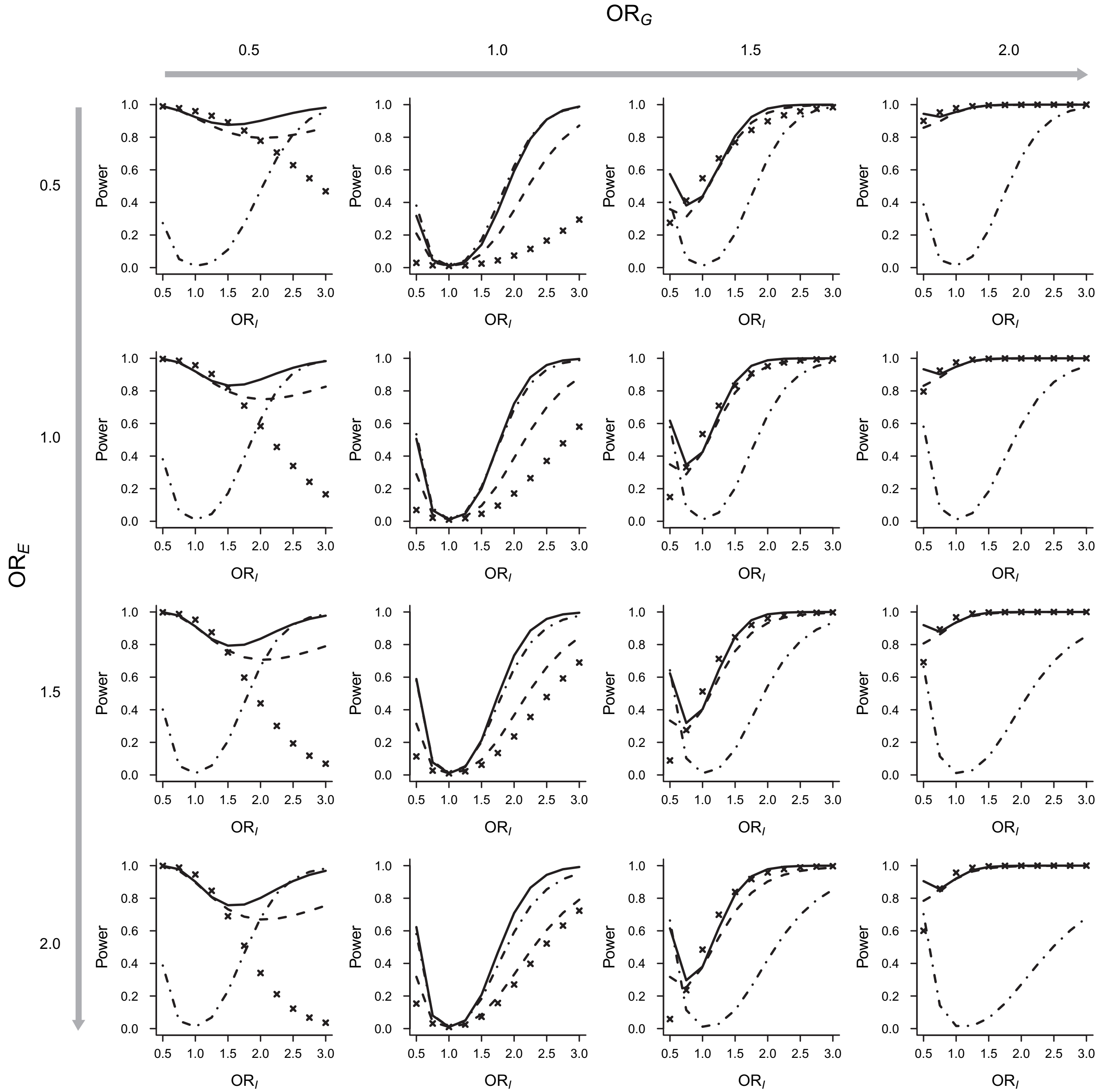

Figure 2. Asymptotic power as a function of the odds ratio (OR) for gene-environment interaction (OR$_I$) for a range of genetic main effects (OR$_G$) and environmental main effects (OR$_E$). Fixed parameters: environmental exposure frequency, $f_E = 0.3$; dominant genetic model with risk genotype frequency, $f_G = 0.3$; gene-environment correlation, $\theta_{GE} = 1.0$; disease prevalence, $f_D = 0.1$; nominal type I error rate of 0.01, 2-sided test; sample of 500 cases and 500 controls. The marginal-G test is represented by crosses (×), the case-only-I test is represented by alternate dots and dashes (· – · – ·), the binomial-GI test is represented by dashed lines (– – –), and the multinomial-GI test is represented by solid lines (——).

RESULTS

For better readability, we present here only the results comparing the multinomial-GI test with the marginal-G, case-only-I, and binomial-GI tests. The comparisons with the 2 other tests are presented in Web Figures 1–4, which are posted on the *Journal*'s Web site (http://aje.oxfordjournals.org/). A comparison with additive models is presented in Web Figure 5.

Type I error rates

Type I error rates obtained under the null hypothesis of no *G* effect and no $G \times E$ interaction are presented in Figure 1.

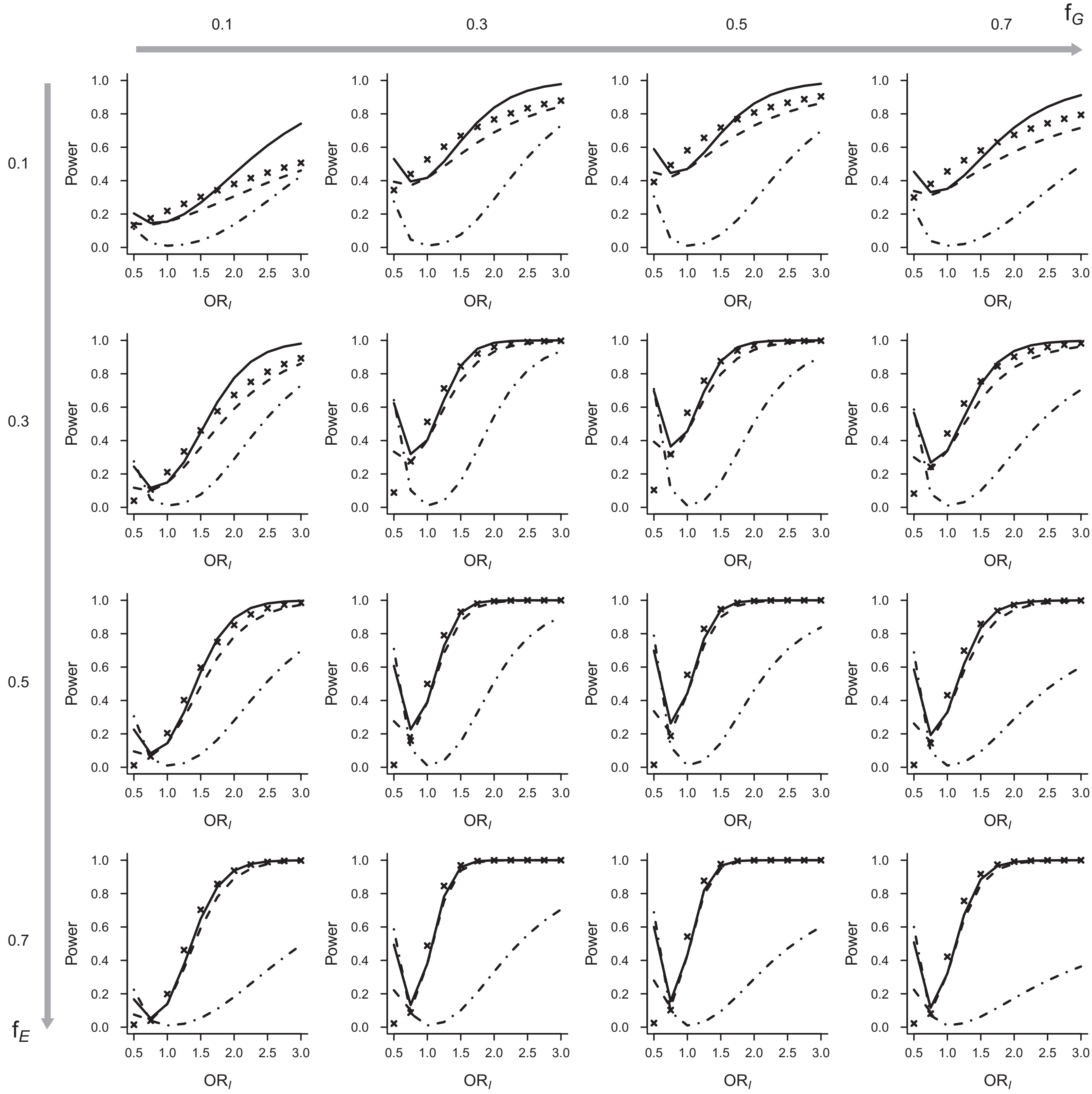

Figure 3. Asymptotic power as a function of the odds ratio (OR) for gene-environment interaction (OR_I) for a range of risk genotype frequencies (f_G) and environmental exposure frequencies (f_E). Fixed parameters: environmental odds ratio, $OR_E = 1.5$; dominant genetic model with genetic odds ratio, $OR_G = 1.5$; gene-environment correlation, $\theta_{GE} = 1.0$; disease prevalence, $f_D = 0.1$; nominal type I error rate of 0.01, 2-sided test; sample of 500 cases and 500 controls. The marginal-G test is represented by crosses (×), the case-only-I test is represented by alternate dots and dashes (· – · – ·), the binomial-GI test is represented by dashed lines (– – –), and the multinomial-GI test is represented by solid lines (——).

In the absence of gene-environment correlation ($\theta_{GE} = 1$), all tests have a type I error rate consistent with the nominal value of 0.01. In the presence of correlation ($\theta_{GE} > 1$), type I errors of the binomial-GI test remain at the nominal value but those of the case-only-I and multinomial-GI tests are significantly increased, reaching values up to 89.7% and 84.9%, respectively, when $\theta_{GE} = 2$, $OR_E = 2$, $f_G = 0.5$, and $f_E = 0.3$ (Web Table 1). The magnitude of the inflation depends mainly on the value of θ_{GE} (the more it deviates from 1, the higher) and the values of f_G and f_E (with a maximum inflation for values between 0.3 and 0.5). The type I errors of

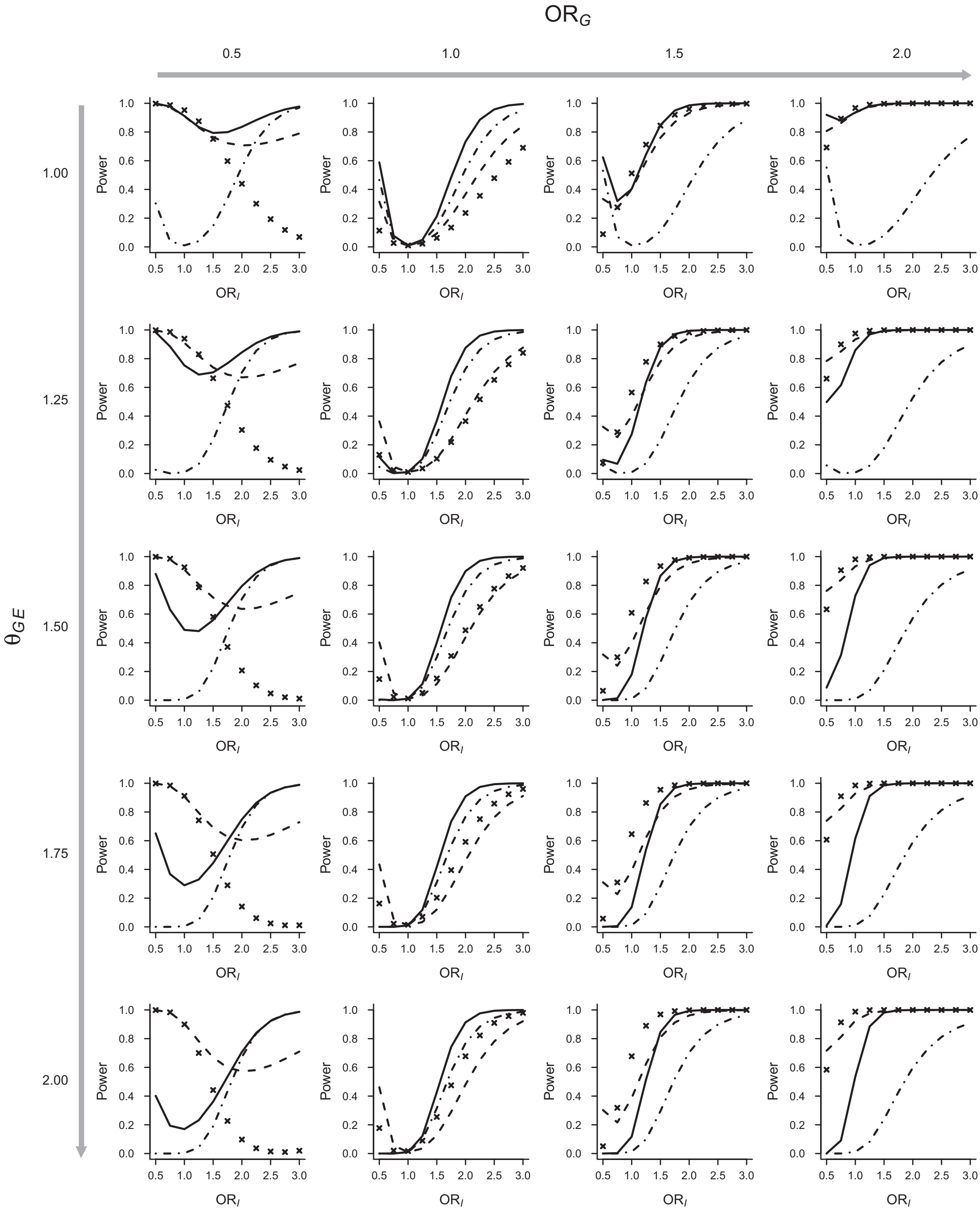

Figure 4. Asymptotic power as a function of the odds ratio (OR) for gene-environment interaction (OR_I) for a range of genetic main effects (OR_G) and gene-environment correlations (θ_{GE}). Fixed parameters: environmental odds ratio, $OR_E = 1.5$; environmental exposure frequency, $f_E = 0.3$; dominant genetic model with risk genotype frequency, $f_G = 0.3$; disease prevalence, $f_D = 0.1$; nominal type I error rate of 0.01, 2-sided test; sample of 500 cases and 500 controls. The marginal-G test is represented by crosses (×), the case-only-I test is represented by alternate dots and dashes (· – · – ·), the binomial-GI test is represented by dashed lines (– – –), and the multinomial-GI test is represented by solid lines (——).

Table 2. Variation in the Percentage of Asymptotic Power in the Presence of Gene-Environment Correlation[a]

OR_G[b]	θ_{GE}[c]	Adjusted-G		Marginal-G		Binomial-GI		Multinomial-GI		Case-Control-I		Case-Only-I	
		OR_I[d]		OR_I		OR_I		OR_I		OR_I		OR_I	
		0.5	2.0	0.5	2.0	0.5	2.0	0.5	2.0	0.5	2.0	0.5	2.0
1.0	1.00	9.68[e]	15.20[e]	11.35[e]	23.62[e]	31.43[e]	36.67[e]	58.86[e]	73.30[e]	27.77[e]	27.39[e]	57.84[e]	65.25[e]
	1.25	2.89	3.51	1.23	12.05	4.86	4.05	–47.43	14.28	2.18	0.95	–52.78	11.24
	1.50	5.71	6.62	1.68	22.10	8.80	7.11	–58.37	16.58	3.60	1.34	–57.81	12.35
	1.75	8.40	9.39	1.68	30.05	12.03	9.49	–58.84	17.58	4.51	1.37	–57.83	12.64
	2.00	10.96	11.86	1.46	36.30	14.71	11.37	–58.85	18.07	5.05	1.19	–57.83	12.44
1.5	1.00	10.30[e]	92.60[e]	8.88[e]	96.04[e]	33.34[e]	93.22[e]	62.39[e]	98.57[e]	29.23[e]	20.60[e]	64.26[e]	54.49[e]
	1.25	–2.39	1.39	–1.63	2.12	–0.76	1.29	–52.79	0.83	2.20	0.71	–58.19	13.65
	1.50	–4.08	2.27	–3.25	2.98	–1.54	2.08	–62.21	0.84	3.63	1.01	–64.22	15.69
	1.75	–5.31	2.88	–4.60	3.37	–2.32	2.59	–62.38	0.82	4.55	1.05	–64.25	16.60
	2.00	–6.22	3.31	–5.65	3.58	–3.08	2.95	–62.38	0.79	5.10	0.93	–64.25	16.84

Abbreviation: OR, odds ratio.

[a] Fixed parameters: environmental odds ratio, $OR_E = 1.5$; environmental exposure frequency, $f_E = 0.3$; dominant genetic model with risk genotype frequency, $f_G = 0.3$; disease prevalence, $f_D = 0.1$; sample of 500 cases and 500 controls.

[b] OR_G, genetic odds ratio of the penetrance model (equation 11).

[c] θ_{GE}, gene-environment correlation (equation 3).

[d] OR_I, gene-environment interaction odds ratio of the penetrance model (equation 11).

[e] Asymptotic power in the absence of gene-environment correlation ($\theta_{GE} = 1$). Other values represent variations in percentage from this reference value: either an increase when the value is positive or a decrease when the value is negative.

the marginal-G test are also slightly inflated but do not compare with the inflation of the other 2 methods.

Power in the absence of gene-environment correlation ($\theta_{GE} = 1$)

As Figure 2 and Figure 3 show, the multinomial-GI test performs better than or as well as all other tests in the presence of a $G \times E$ interaction. Compared with the marginal-G test, the multinomial-GI test slightly loses power in the absence of $G \times E$ interaction, since stratifying the case sample brings little information and increases both the variance of parameters and the degrees of freedom of the test. The difference in power is at most 11.25% when $OR_G = 1.5$ and $f_G = 0.5$. In contrast, in the presence of $G \times E$ interaction, the gain in power is much higher, particularly for pure interaction effects in the absence of main effects of G and E. In such situations, the multinomial-GI test improves statistical power by 42%.

Compared with the case-only-I test, the multinomial-GI test performs much better when G has a main effect ($OR_G \neq 1$), with a power gain that can reach 98.4%, and it performs similarly when $OR_G = 1$.

Finally, the most striking point evidenced in Figure 2 is that, over all of the studied models, the multinomial-GI test outperforms the binomial-GI test, with a gain in power of 56% in some situations.

Additional results presented in Web Figure 2 show that adjusting on exposure (adjusted-G) globally decreases power in all situations, except when either the E effect or the $G \times E$ interaction point in opposite directions than the G effect (see the first line and the first column of Web Figure 2), whereas the case-control-I test has the lowest power.

Power in the presence of gene-environment correlation ($\theta_{GE} > 1$)

Power computations under scenarios including a gene-environment correlation ($\theta_{GE} > 1$) were adjusted for inflated type I error rates by using a corrected threshold instead of the central χ^2 threshold. These corrected thresholds were computed using a noncentral χ^2 distribution with a noncentrality parameter equal to the value of the test under the corresponding scenario with the same θ_{GE} value but with no G and $G \times E$ interaction effects (null hypothesis).

With increasing values of θ_{GE}, an important decrease in power can be observed for the 2 methods that had an inflated type I error (case-only-I and multinomial-GI tests), especially for small $G \times E$ interactions and under flip-flop scenarios in which $OR_I < 1$ (Figure 4). Using the power when $\theta_{GE} = 1$ as the reference value, Table 2 quantifies variations in power for increasing θ_{GE}. For strong $G \times E$ interaction, all 6 tests have an increase in power when θ_{GE} increases, but there is a high power loss for both the multinomial-GI test and the case-only-I test under flip-flop scenarios ($OR_I < 1$).

Similar results are observed with an additive model, with an overall increase in the power of the tests (Web Figure 5).

Bias and variance of genetic and $G \times E$ interaction estimators

The bias, variance, and coverage probability of the different genetic parameters in the absence of gene-environment correlation are shown in Table 3. The bias associated with the effect of G was measured by taking OR_G as the expected

Table 3. Squared Bias, Variance, and Coverage Probability of the Different Estimators of the Genetic Parameter in the Absence of Gene-Environment Correlation[a]

OR_G[b]	OR_I[c]	OR_E[d]	Squared Bias × 10^2				Variance × 10^2				Coverage Probability × 10^2			
			β_{GM}[e]	β_{GA}[f]	β_{GCC}[g]	β_{G1}[h]	β_{GM}	β_{GA}	β_{GCC}	β_{G1}	β_{GM}	β_{GA}	β_{GCC}	β_{G1}
0.5	1.0	1.0	0.18	0.17	0.19	0.18	2.30	2.31	3.30	2.88	95.7	95.6	94.6	95.1
		2.0	0.41	0.38	0.20	0.21	2.29	2.35	3.80	3.39	91.3	91.5	93.9	93.9
	2.0	1.0	8.20	7.51	0.22	0.19	2.12	2.14	3.40	2.98	50.0	52.9	94.9	93.9
		2.0	14.44	10.17	0.14	0.13	2.06	2.16	3.97	3.55	23.4	40.9	95.2	96.2
1.0	1.0	1.0	0.00	0.00	0.00	0.00	1.91	1.91	2.74	2.32	95.1	94.9	95.6	94.9
		2.0	0.00	0.00	0.00	0.00	1.91	1.96	3.08	2.67	94.5	94.4	94.9	94.3
	2.0	1.0	5.04	4.39	0.00	0.00	1.84	1.86	2.84	2.43	62.2	67.3	94.8	95.6
		2.0	7.53	4.81	0.01	0.00	1.83	1.91	3.25	2.84	46.9	65.7	94.8	95.0
1.5	1.0	1.0	0.19	0.18	0.21	0.18	1.81	1.81	2.58	2.17	94.0	94.2	92.3	94.2
		2.0	0.26	0.23	0.15	0.13	1.81	1.85	2.88	2.47	93.2	93.6	94.5	95.2
	2.0	1.0	2.42	1.99	0.27	0.26	1.77	1.79	2.69	2.28	79.0	81.5	93.2	93.2
		2.0	3.00	1.57	0.26	0.29	1.77	1.84	3.04	2.63	73.9	84.7	92.7	93.9
2.0	1.0	1.0	0.94	0.92	0.94	0.97	1.77	1.77	2.53	2.12	87.9	88.1	91.0	88.8
		2.0	1.31	1.18	0.58	0.66	1.77	1.81	2.81	2.39	84.8	86.4	92.1	92.2
	2.0	1.0	1.19	0.91	0.67	0.72	1.76	1.77	2.65	2.23	88.7	91.1	92.1	91.8
		2.0	0.92	0.34	0.77	0.78	1.76	1.82	2.95	2.54	88.0	92.5	91.5	90.4

Abbreviation: OR, odds ratio.

[a] Fixed parameters: environmental exposure frequency, $f_E = 0.3$; dominant genetic model with risk genotype frequency, $f_G = 0.3$; disease prevalence, $f_D = 0.1$; gene-environment correlation, $\theta_{GE} = 1$; sample of 500 cases and 500 controls.

[b] OR_G, genetic odds ratio of the penetrance model (equation 11).

[c] OR_I , *G* × *E* interaction odds ratio of the penetrance model (equation 11).

[d] OR_E, environmental odds ratio of the penetrance model (equation 11).

[e] β_{GM}, genetic regression coefficient of the marginal model (equation 6).

[f] β_{GA}, genetic regression coefficient of the adjusted model (equation 7).

[g] β_{GCC}, genetic regression coefficient of the full model (equation 8).

[h] β_{G1}, genetic regression coefficient of the multinomial model (equation 10).

value. This is actually not the expected value of $\log(\beta_{G1})$ in the multinomial model, but it truly measures the effect of *G* in the absence of any other factor. The bias of β_{G1} is found to be very similar to that of β_{GCC} in the full logistic model and lower than the biases of β_{GM} and β_{GA} in the marginal and adjusted models. The variance of β_{G1} is lower than that of β_{GCC} but higher than the variances of β_{GM} and β_{GA}, β_{GM} having the lowest variance of all genetic parameters. The coverage probabilities of both β_{G1} and β_{GCC} confidence intervals are very close to the expected 95% value but decrease with increasing OR_G, whereas those of β_{GM} and β_{GA} are lower, particularly for elevated values of OR_E and OR_I.

A *G* × *E* interaction coefficient estimator can be derived from the ratio of both parameters of the multinomial model: β_{G2}/β_{G1}. This estimator has exactly the same bias, variance, and coverage probability patterns as the parameter derived from the logistic model of the case-only design (β_{ICO}). Compared with the estimator of the full model using the case-control design (β_{ICC}), β_{ICO} and β_{G2}/β_{G1} have a higher bias and lower coverage probability when OR_I increases and a lower variance over all models (Table 4).

Web Table 2 and Web Table 3 show the same results in the presence of gene-environment correlation ($\theta_{GE} = 1.5$).

DISCUSSION

The multinomial logistic regression model is a simple and efficient model for comparing exposed and unexposed cases with controls when the exposure information is available in cases only. It allows a combined test of genetic association and *G* × *E* interaction, merging together a marginal test of genetic association regardless of exposure and a case-only *G* × *E* interaction test regardless of the genotype distribution in controls. By combining these 2 designs into a unified approach, we show how it is possible to maintain satisfactory statistical power while fulfilling the 2 concerns of detecting a potential genetic factor and not missing it because it interacts with a particular environmental factor. This is well exemplified by the fact that our approach is similar to or only slightly less powerful than tests of *G* effect only in the absence of *G* × *E* interaction and than the case-only-I test in the presence of pure *G* × *E* interaction (no main *G* effect)—situations in which these 2 tests are respectively the most powerful. With a better variance and a similar precision, the estimators of the *G* effect and of the *G* × *E* interaction of the multinomial model have better coverage probabilities than estimators obtained with methods that account for the exposure status of controls.

Table 4. Squared Bias, Variance, and Coverage Probability of the Different Estimators of the Gene-Environment Interaction Parameter in the Absence of Gene-Environment Correlation[a]

OR_I[b]	OR_G[c]	OR_E[d]	Squared Bias × 10^2			Variance × 10^2			Coverage Probability × 10^2		
			β_{ICC}[e]	β_{ICO}[f]	β_{G2}/β_{G1}[g]	β_{ICC}	β_{ICO}	β_{G2}/β_{G1}	β_{ICC}	β_{ICO}	β_{G2}/β_{G1}
0.5	1.0	1.0	0.22	0.21	0.21	11.00	6.41	6.41	94.7	94.7	94.7
		2.0	0.71	0.53	0.53	9.53	4.91	4.91	93.0	93.1	93.1
	2.0	1.0	0.83	0.67	0.67	9.51	4.90	4.90	92.6	93.1	93.1
		2.0	0.84	0.58	0.58	8.45	3.84	3.84	93.9	93.1	93.1
1.0	1.0	1.0	0.04	0.00	0.00	9.17	4.60	4.60	95.5	94.2	94.2
		2.0	0.01	0.00	0.00	8.52	3.91	3.91	95.3	93.8	93.8
	2.0	1.0	0.00	0.00	0.00	8.50	3.91	3.91	95.6	95.4	95.4
		2.0	0.78	0.62	0.62	7.94	3.35	3.35	94.3	91.7	91.7
1.5	1.0	1.0	0.17	0.16	0.16	8.65	4.05	4.05	93.8	95.6	95.6
		2.0	0.29	0.48	0.48	8.26	3.65	3.65	92.7	92.8	92.8
	2.0	1.0	0.53	0.63	0.63	8.25	3.66	3.66	93.4	92.2	92.2
		2.0	3.49	3.97	3.97	7.87	3.26	3.26	88.4	81.0	81.0
2.0	1.0	1.0	0.63	0.80	0.80	8.43	3.82	3.82	94.0	93.6	93.6
		2.0	1.63	2.22	2.22	8.18	3.57	3.57	92.6	87.6	87.6
	2.0	1.0	2.44	2.33	2.33	8.16	3.57	3.57	91.7	86.1	86.1
		2.0	9.64	9.86	9.86	7.84	3.25	3.25	79.8	59.9	59.9

Abbreviation: OR, odds ratio.

[a] Fixed parameters: environmental exposure frequency, $f_E = 0.3$; dominant genetic model with risk genotype frequency, $f_G = 0.3$; disease prevalence, $f_D = 0.1$; gene-environment correlation, $\theta_{GE} = 1$; sample of 500 cases and 500 controls.

[b] OR_I, gene-environment ($G \times E$) interaction odds ratio of the penetrance model (equation 11).

[c] OR_G, genetic odds ratio of the penetrance model (equation 11).

[d] OR_E, environmental odds ratio of the penetrance model (equation 11).

[e] β_{ICC}, $G \times E$ interaction regression coefficient of the full model (equation 8).

[f] β_{ICO}, $G \times E$ interaction regression coefficient of the case-only model (equation 9).

[g] β_{G2}/β_{G1}, $G \times E$ interaction regression coefficient of the multinomial model (equation 10).

Interestingly, over all of the $G \times E$ interaction models explored, the multinomial-GI test outperformed the binomial-GI test proposed by Kraft et al. (9), which makes use of the environmental statuses of cases and controls. This gain in power is due to the smaller variance of the parameter estimates under the multinomial model. This unexpected result suggests that, as long as independence between G and E holds, exposure information on controls is not mandatory for exploring G and its interaction with E. However, exposure information on controls becomes more important when there is gene-environment correlation in the underlying population. In this situation, such information becomes crucial in order to avoid false-positive detection of $G \times E$ interactions, as shown by the inflated type I errors of the multinomial-GI and case-only-I tests. The nonrobustness of the case-only design to the presence of gene-environment correlation is well-known in the literature and has sometimes limited its use for detection of $G \times E$ interaction. However, being aware of this problem, we think the multinomial-GI test is worth considering when exposure information is absent in controls, as is often the case in large-scale genome-wide association scans where a reference control panel is used. One could then either rely on previous studies in the same population to exclude an underlying genetic correlation with the environmental factor being studied or study the association between both factors in a second step in a more specific control group where exposure information is available. This strategy is also discussed in the case-only design literature (19–22). It might be even more problematic in the context of genome-wide association scans with a reference control panel, since in that case controls might not be very well matched to the cases, but several methods for selecting the best-matched controls have been proposed in the literature (3, 23).

Throughout this study, we only explored positively correlated risk factors ($\theta_{GE} > 1$). Scenarios in which the G and E factors are negatively correlated but are positively interacting on the disease seem very unlikely. We also concentrated on dominant genetic models, but similar trends with overall lower power values would be expected if one were assuming a recessive genetic model. We also explored flip-flop $G \times E$ interactions in which the risk genotype becomes protective when the exposure status changes. In this situation, the multinomial-GI test proves to be very interesting while outperforming other methods. Flip-flop interactions are rarely described in real data sets, probably because this is not a common situation, but also perhaps because traditional logistic approaches do not have enough power to detect them.

A similar approach to the multinomial-GI test was proposed by Umbach and Weinberg (24) using log-linear models. The relations between the multinomial and log-linear equations are shown in the Web Appendix. The main difference between the log-linear modeling and the logistic modeling resides in the choice of the risk factor estimator, either the relative risk or the odds ratio, respectively (25). An advantage of the logistic model is the possibility of including continuous covariates in the model.

Easy to implement in most existing statistical packages, the multinomial model is flexible and can also handle diseases categorized into subphenotypes (16), multinomial environmental exposures, and a combination of the two. As in the standard logistic regression analysis, adjusting on specific covariates such as age, sex, or recruitment category is possible. We believe it could help in improving our understanding and appraisal of $G \times E$ interactions in genetic association studies, particularly at the genome-wide level, provided that we could select a few ''interesting'' exposures to test for (26).

ACKNOWLEDGMENTS

Author affiliations: Faculté de Médecine, Université Paris-Sud 11, Le Kremlin Bicêtre, Paris, France (Rémi Kazma); U946 (Variabilité Génétique et Maladie Humaine), Institut National de la Santé et de la Recherche Médicale, Paris, France (Rémi Kazma, Marie-Claude Babron, Emmanuelle Génin); and Institut Universitaire d'Hématologie, Université Paris Diderot, Paris, France (Marie-Claude Babron, Emmanuelle Génin).

Dr. Rémi Kazma's doctoral work was supported by grants from the Université Paris-Sud President and from the Fondation pour la Recherche Médicale.

Conflict of interest: none declared.

REFERENCES

1. Lander ES, Schork NJ. Genetic dissection of complex traits. *Science*. 1994;265(5181):2037–2048.
2. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest*. 2008;118(5):1590–1605.
3. Luca D, Ringquist S, Klei L, et al. On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. *Am J Hum Genet*. 2008; 82(2):453–463.
4. Nelson MR, Bryc K, King KS, et al. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am J Hum Genet*. 2008; 83(3):347–358.
5. GAIN Collaborative Research Group. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat Genet*. 2007;39(9): 1045–1051.
6. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447(7145):661–678.
7. Wichmann HE, Gieger C, Illig T. KORA-gen—resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen*. 2005;67(suppl 1):S26–S30.
8. Selinger-Leneman H, Genin E, Norris JM, et al. Does accounting for gene-environment (G x E) interaction increase the power to detect the effect of a gene in a multifactorial disease? *Genet Epidemiol*. 2003;24(3):200–207.
9. Kraft P, Yen YC, Stram DO, et al. Exploiting gene-environment interaction to detect genetic associations. *Hum Hered*. 2007; 63(2):111–119.
10. Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. *Am J Epidemiol*. 2009;169(2):219–226.
11. Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! *Am J Epidemiol*. 1996; 144(3):207–213.
12. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med*. 1994;13(2):153–162.
13. Lindström S, Yen YC, Spiegelman D, et al. The impact of gene-environment dependence and misclassification in genetic association studies incorporating gene-environment interactions. *Hum Hered*. 2009;68(3):171–181.
14. Dobson AJ. Nominal and ordinal logistic regression. In: *An Introduction to Generalized Linear Models*. Boca Raton, FL: Chapman & Hall, Inc; 2002:135–150.
15. Kleinbaum DG, Klein M. Polytomous logistic regression. In: *Logistic Regression*. New York, NY: Springer-Verlag; 2002: 267–292.
16. Morris AP, Lindgren CM, Zeggini E, et al. A powerful approach to sub-phenotype analysis in population-based genetic association studies. *Genet Epidemiol*. 2010;34(4):335–343.
17. R Development Core Team. *R Version 2.9.1*. Vienna, Austria: R Foundation for Statistical Computing; 2009.
18. Stata Corporation. *Stata Statistical Software, Release 10*. College Station, TX: StataCorp LP; 2007.
19. Albert PS, Ratnasinghe D, Tangrea J, et al. Limitations of the case-only design for identifying gene-environment interactions. *Am J Epidemiol*. 2001;154(8):687–693.
20. Gatto NM, Campbell UB, Rundle AG, et al. Further development of the case-only design for assessing gene-environment interaction: evaluation of and adjustment for bias. *Int J Epidemiol*. 2004;33(5):1014–1024.
21. Goldstein AM, Andrieu N. Detection of interaction involving identified genes: available study designs. *J Natl Cancer Inst Monogr*. 1999;(26):49–54.
22. Schmidt S, Schaid DJ. Potential misinterpretation of the case-only study to assess gene-environment interaction. *Am J Epidemiol*. 1999;150(8):878–885.
23. Guan W, Liang L, Boehnke M, et al. Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies. *Genet Epidemiol*. 2009; 33(6):508–517.
24. Umbach DM, Weinberg CR. Designing and analysing case-control studies to exploit independence of genotype and exposure. *Stat Med*. 1997;16(15):1731–1743.
25. Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. *Stat Med*. 2002;21(1):35–50.
26. Kazma R, Bonaïti-Pellié C, Norris JM, et al. On the use of sibling recurrence risks to select environmental factors liable to interact with genetic risk factors. *Eur J Hum Genet*. 2010; 18(1):88–94.