

The Effects of Population Stratification and Misclassification in Studies of Gene-Environment Interactions

KF Cheng and WJ Lin
China Medical University and
National Central University

- Disease status is denoted as D with levels $d=0, 1$, indicating the presence and absence of disease, respectively.
- The true genotype is denoted as G with levels $g=0, 1, 2$, where $G=g$ means that the individual carries g copies of the high-risk candidate allele.
- The true environmental exposure is denoted as E with levels $e=0, 1$ where 1 indicates present and 0 absent.

- Assuming binary environmental exposure is only for the purpose of simplicity in presentation. All results shown in this work can be extended to any environmental exposure with finite number of levels.
- We further assume that misclassification is only present in data obtained on the genotype and environmental exposure, and that the disease status is correctly classified.

- The misclassified genotype and environmental exposure are represented by G' and E' , respectively.
- Finally, we let S denote a general stratification variable so that model (1) described subsequently is valid. S has values $s=1, \dots, K$, representing K strata.

The two important conditions used in this paper are:

Condition (1): Among the controls, an individual's true environmental exposure is independent of his/her value of stratification variable when the true genotype is given,

$$\begin{aligned} & \Pr(E = e, S = s | G = g, D = 0) \\ &= \Pr(E = e | G = g, D = 0) \Pr(S = s | G = g, D = 0), \end{aligned}$$

and

Condition (2): Among the controls, an individual's true genotype is independent his/her true environmental exposure at given value of the stratification variable,

$$\begin{aligned} & \Pr(G = g, E = e | S = s, D = 0) \\ &= \Pr(G = g | D = 0, S = s) \Pr(E = e | D = 0, S = s). \end{aligned}$$

- We postulate the following general risk model:

$$\text{logit Pr}(D = 1 | G = g, E = e, S = s) = \mu' + \alpha'_s + \beta_g + \gamma e + \delta_g e \quad (1)$$

- For identifiability, we define α'_1 , β_0 , and δ_0 to be zeroes so that $s=1$, $g=0$, and $e=0$ represent the referent subpopulation, genotype and environmental exposure, respectively.
- α'_s in model (1) represents subpopulation-specific intercept parameter to account for potential heterogeneity in disease risk between different strata.

Under Population Stratification

if one ignores population stratification then the prospective model for relating disease risk, genotype, and environmental exposure becomes:

$$\begin{aligned} & \text{logit Pr}(D = 1 | G = g, E = e) \\ &= \xi + (\beta_g + \beta_g^*) + (\gamma + \gamma^*)e + (\delta_g + \delta_g^*)e, \end{aligned} \quad (3)$$

where

$$\xi = (\mu + \mu^*) + \log \left\{ \text{Pr}(D = 0) / \text{Pr}(D = 1) \right\}$$

- The values of these new parameters $(\beta_g^*, \gamma^*, \delta_g^*)$ represent the bias levels under population stratification in case-control studies. In case-only studies, the bias levels caused by population stratification are represented by δ_g^* .
- According to their definitions, the bias levels depend on the control (G, E) frequencies across subpopulations, sampling proportions $P^*(S = s | D = d)$ of subpopulation s in cases ($d=1$) and controls ($d=0$), and subpopulation-specific intercepts α_s .

Result 1:

- Suppose that model (1) and *Condition (1)* are both true, then $\delta_1^* = \delta_2^* = \gamma^* = 0$; that is, in case-control studies using only (G, E) data, the estimated interactions and main effect of the environmental exposure will be unbiased.
- Consequently, regular tests of multiplicative interactions or main effect of E will be valid even when there exists population stratification.

Result 2:

If Conditions in *Result 1* and (G,S) - E independence in the controls are satisfied, then test of G - E independence in the cases is a valid test of no interactions in case-only studies even when population stratification is present.

Misclassification

The misclassified genotype and environmental exposure are represented by G' and E' , respectively.

Condition (3): The misclassified genotype G' and environmental exposure E' are conditionally independent,

$$\begin{aligned} & \Pr(G' = g', E' = e' | G = g, E = e, S = s, D = d) \\ &= \Pr(G' = g' | G = g, S = s, D = d) \times \Pr(E' = e' | E = e, S = s, D = d). \end{aligned}$$

Result 3:

If gene-environmental interactions are not present ($\delta_g = 0$) in the true disease-risk model (1), and **Conditions (2)** and **(3)** are satisfied, then in fitting the following more general disease-risk model

$$\begin{aligned} & \text{logit Pr}(D = 1 | G' = g, E' = e, S = s) \\ & = \theta^* + \beta_g^* + \gamma^* e + \alpha_s^* + \delta_g^* e + \omega_{(g,s)}^* + \tau_s^* e, \end{aligned}$$

the estimated gene-environment interactions are unbiased (since $\delta_g^* = 0$).

Consider the logit model:

$$\text{logit Pr}(E' = 1 | G' = g', S = s, D = 1) = \gamma' + \tau'_s + \delta'_g \quad (2)$$

Result 4:

If disease-risk model (1) is satisfied, and *Conditions (2)* and *(3)* are satisfied, then test of $G'-E'$ conditional independence based on model (2) (that is, test of $\delta'_g = 0$) is a valid test of no gene-environment interactions in case-only studies.

Condition (4):

$$\begin{aligned} & \Pr(G' = g', E' = e' \mid G = g, E = e, D = 1) \\ &= \Pr(G' = g' \mid G = g, D = 1) \times \Pr(E' = e' \mid E = e, D = 1). \end{aligned}$$

Result 5:

If Conditions in *Result 2* and (4) are satisfied, then test of no interactions in case-only studies is valid even when misclassification is present.

Simulation Results

Table 1. Empirical Type I Error Rates Under Population Stratification ^a

Population Stratification	$(\beta_1, \beta_2, \gamma)$	q ^c	t ^d	Type I error ^e		
				T ₁	T ₂ ($\frac{1}{2}$)	T ₂ (1)
No	(0, 0, 1)			0.055	0.057	0.049
Yes ^b	(0, 0, 1)	0.50	0.03	0.055	0.050	0.053
			0.05	0.060	0.051	0.051
			0.10	0.063	0.054	0.062
		0.70	0.03	0.064	0.055	0.056
			0.05	0.065	0.051	0.052
			0.10	0.056	0.049	0.051
		1.00	0.03	0.057	0.057	0.048
			0.05	0.064	0.057	0.060
			0.10	0.057	0.062	0.054

- a. The case sample size is $n_1 = 150$, and the control sample size is $n_2 = 150$. The empirical type I error rates were computed based on 2000 replications.
- b. The general population consists of two subpopulations.
- c. q is the proportion of the diseased (nondiseased) individuals sampled from the first (second) subpopulation.
- d. t is the difference of allele frequencies in two subpopulations.
- e. The computation of type I error rates for T₂($\frac{1}{2}$) (T₂(1)) is based on case data with size 150 (300).

Simulation Results

Table 1. Empirical Type I Error Rates Under Population Stratification ^a

Population Stratification	$(\beta_1, \beta_2, \gamma)$	q^c	t^d	Type I error ^e		
				T_1	$T_2(\frac{1}{2})$	$T_2(1)$
No	(1 , 2 , 1)			0.056	0.057	0.056
Yes ^b	(1 , 2 , 1)	0.50	0.03	0.053	0.056	0.048
			0.05	0.058	0.053	0.042
			0.10	0.058	0.049	0.052
		0.70	0.03	0.054	0.050	0.054
			0.05	0.060	0.046	0.049
			0.10	0.049	0.045	0.054
		1.00	0.03	0.063	0.051	0.052
			0.05	0.053	0.043	0.047
			0.10	0.049	0.060	0.052

- a. The case sample size is $n_1 = 150$, and the control sample size is $n_2 = 150$. The empirical type I error rates were computed based on 2000 replications.
- b. The general population consists of two subpopulations.
- c. q is the proportion of the diseased (nondiseased) individuals sampled from the first (second) subpopulation.
- d. t is the difference of allele frequencies in two subpopulations.
- e. The computation of type I error rates for $T_2(\frac{1}{2})$ ($T_2(1)$) is based on case data with size 150 (300).

Simulation Results

Table 2. Empirical Powers Under Population Stratification ^a

Population Stratification	$(\beta_1, \beta_2, \gamma)$	q^c	t^d	Power ^e		
				T_1	$T_2(\frac{1}{2})$	$T_2(1)$
No	(0 , 0 , 1)			0.848	0.950	1.000
Yes ^b	(0 , 0 , 1)	0.50	0.03	0.850	0.945	1.000
			0.05	0.821	0.919	1.000
			0.10	0.806	0.904	0.998
		0.70	0.03	0.843	0.943	0.999
			0.05	0.823	0.928	0.998
			0.10	0.831	0.928	0.998
		1.00	0.03	0.849	0.952	1.000
			0.05	0.842	0.955	0.999
			0.10	0.831	0.958	0.999

- a. The case sample size is $n_1 = 150$, and the control sample size is $n_2 = 150$. The empirical powers were computed based on 2000 replications.
- b. The general population consists of two subpopulations.
- c. q is the proportion of the diseased (nondiseased) individuals sampled from the first (second) subpopulation.
- d. t is the difference of allele frequencies in two subpopulations.
- e. The computation of powers for $T_2(\frac{1}{2})$ ($T_2(1)$) is based on case data with size 150 (300).

Simulation Results

Table 2. Empirical Powers Under Population Stratification ^a

Population Stratification	$(\beta_1, \beta_2, \gamma)$	q ^c	t ^d	Power ^e		
				T ₁	T ₂ ($\frac{1}{2}$)	T ₂ (1)
No	(1 , 2 , 1)			0.547	0.609	0.869
Yes ^b	(1 , 2 , 1)	0.50	0.03	0.515	0.579	0.837
			0.05	0.499	0.567	0.815
			0.10	0.463	0.507	0.783
		0.70	0.03	0.521	0.593	0.852
			0.05	0.516	0.587	0.848
			0.10	0.495	0.560	0.828
		1.00	0.03	0.542	0.619	0.874
			0.05	0.537	0.599	0.872
			0.10	0.541	0.610	0.872

- a. The case sample size is $n_1 = 150$, and the control sample size is $n_2 = 150$. The empirical powers were computed based on 2000 replications.
- b. The general population consists of two subpopulations.
- c. q is the proportion of the diseased (nondiseased) individuals sampled from the first (second) subpopulation.
- d. t is the difference of allele frequencies in two subpopulations.
- e. The computation of powers for $T_2(\frac{1}{2})$ ($T_2(1)$) is based on case data with size 150 (300).

Simulation Results

Table 3. Empirical Type I Error Under Misclassification ^a

$(\beta_1, \beta_2, \gamma)$	t^b	π^c	ε^d	Type I Error ^e				
				$T_2(\frac{1}{2})$	$T_2(1)$	T_3		
(0, 0, 1)	0.05	0.10	0.00	0.055	0.049	0.069		
			0.01	0.060	0.052	0.054		
			0.03	0.054	0.050	0.054		
		0.05	0.058	0.054	0.057			
		0.20	0.00	0.059	0.054	0.063		
			0.01	0.050	0.054	0.057		
	0.03		0.055	0.059	0.059			
	0.10	0.10	0.10	0.00	0.051	0.049	0.053	
				0.01	0.053	0.058	0.062	
				0.03	0.053	0.053	0.066	
		0.20	0.10	0.10	0.00	0.057	0.048	0.064
					0.01	0.047	0.052	0.061
					0.03	0.053	0.050	0.061
		0.20	0.10	0.10	0.00	0.054	0.057	0.056
					0.01	0.054	0.057	0.056
0.03					0.054	0.057	0.056	
0.20	0.10	0.10	0.00	0.045	0.057	0.059		
			0.01	0.045	0.057	0.059		
			0.03	0.045	0.057	0.059		

- a. The case sample size is $n_1 = 150$, and the control sample size is $n_2 = 150$. The empirical type I error rates were computed based on 2000 replications. The general population consists of two subpopulations.
- b. t is the difference of allele frequencies in two subpopulations.
- c. π is the error rate for misclassifying environmental exposure.
- d. ε is the genotyping error rate.
- e. The computation of type I error rates for $T_2(\frac{1}{2})$ ($T_2(1)$) is based on case data with size 150 (300).

Simulation Results

Table 3. Empirical Type I Error Under Misclassification ^a

$(\beta_1, \beta_2, \gamma)$	t^b	π^c	ε^d	Type I Error ^e		
				$T_2(\frac{1}{2})$	$T_2(1)$	T_3
(1, 2, 1)	0.05	0.10	0.00	0.056	0.048	0.071
			0.01	0.045	0.053	0.066
			0.03	0.052	0.046	0.057
		0.05	0.056	0.046	0.064	
		0.20	0.00	0.046	0.048	0.052
			0.01	0.059	0.043	0.061
	0.03		0.059	0.049	0.062	
	0.10	0.10	0.00	0.053	0.056	0.057
			0.01	0.059	0.051	0.067
			0.03	0.053	0.051	0.061
		0.20	0.00	0.052	0.046	0.065
			0.01	0.057	0.046	0.061
			0.03	0.045	0.050	0.059
	0.05	0.10	0.00	0.051	0.050	0.060
			0.01	0.059	0.051	0.067
			0.03	0.053	0.051	0.061
		0.20	0.00	0.052	0.046	0.065
			0.01	0.057	0.046	0.061
0.03			0.045	0.050	0.059	

- a. The case sample size is $n_1 = 150$, and the control sample size is $n_2 = 150$. The empirical type I error rates were computed based on 2000 replications. The general population consists of two subpopulations.
- b. t is the difference of allele frequencies in two subpopulations.
- c. π is the error rate for misclassifying environmental exposure.
- d. ε is the genotyping error rate.
- e. The computation of type I error rates for $T_2(\frac{1}{2})$ ($T_2(1)$) is based on case data with size 150 (300).

Simulation Results

Table 4. Empirical Powers Under Misclassification ^a

$(\beta_1, \beta_2, \gamma)$	t^b	π^c	ε^d	Power ^e		
				$T_2(\frac{1}{2})$	$T_2(1)$	T_3
(0, 0, 1)	0.05	0.10	0.00	0.700	0.945	0.530
			0.01	0.674	0.930	0.494
			0.03	0.610	0.891	0.434
			0.05	0.562	0.844	0.389
	0.10	0.10	0.00	0.414	0.724	0.292
			0.01	0.393	0.685	0.275
			0.03	0.356	0.629	0.247
			0.05	0.308	0.534	0.212
			0.00	0.640	0.933	0.481
			0.01	0.619	0.903	0.458
			0.03	0.579	0.869	0.420
			0.05	0.497	0.792	0.348
	0.20	0.20	0.00	0.374	0.656	0.263
			0.01	0.356	0.619	0.265
			0.03	0.318	0.572	0.228
			0.05	0.280	0.515	0.193

- The case sample size is $n_1 = 150$, and the control sample size is $n_2 = 150$. The empirical powers were computed based on 2000 replications. The general population consists of two subpopulations.
- t is the difference of allele frequencies in two subpopulations.
- π is the error rate for misclassifying environmental exposure.
- ε is the genotyping error rate.
- The computation of powers for $T_2(\frac{1}{2})$ ($T_2(1)$) is based on case data with size 150 (300).

Simulation Results

Table 4. Empirical Powers Under Misclassification ^a

$(\beta_1, \beta_2, \gamma)$	t^b	π^c	ε^d	Power ^e		
				$T_2(\frac{1}{2})$	$T_2(1)$	T_3
(1, 2, 1)	0.05	0.10	0.00	0.301	0.553	0.274
			0.01	0.287	0.492	0.249
			0.03	0.240	0.419	0.203
			0.05	0.198	0.343	0.155
		0.20	0.00	0.169	0.298	0.157
			0.01	0.168	0.283	0.158
			0.03	0.131	0.235	0.131
			0.05	0.115	0.181	0.113
	0.10	0.10	0.00	0.271	0.475	0.245
			0.01	0.255	0.418	0.234
			0.03	0.209	0.329	0.164
			0.05	0.172	0.293	0.115
		0.20	0.00	0.179	0.266	0.159
			0.01	0.150	0.238	0.132
			0.03	0.128	0.191	0.125
			0.05	0.108	0.159	0.095

- The case sample size is $n_1 = 150$, and the control sample size is $n_2 = 150$. The empirical powers were computed based on 2000 replications. The general population consists of two subpopulations.
- t is the difference of allele frequencies in two subpopulations.
- π is the error rate for misclassifying environmental exposure.
- ε is the genotyping error rate.
- The computation of powers for $T_2(\frac{1}{2})$ ($T_2(1)$) is based on case data with size 150 (300).