Studying the Joint Effects of Population Stratification and Sampling in Case-Control Association Studies

K.F. Cheng^{a, b} J.Y. Lee^b J.H. Chen^c

^aBiostatistics Center and College of Public Health, China Medical University, Taichung, ^bGraduate Institute of Statistics, National Central University, Chungli, and ^cBiostatistics Center and Graduate Institute of Biostatistics, China Medical University, Taichung, Taiwan, ROC

Key Words

Association • Case-control study • Population stratification • Sampling scheme • Type I error

Abstract

Population stratification (PS) is referred to the systematic difference in allele frequencies between subpopulations in a population. It could cause a false-positive conclusion in a case-control association study, where the association is due to the structure of the underlying population, not a disease-associated locus. In this paper, we study the joint effects of PS and data sampling when the genetic effect is null. The level of the PS effect depends on the variation of the baseline genotype frequency across subpopulations and matching effectiveness of the sampling. In the case of simple random sampling (SRS), the matching effectiveness equals the inverse of the variation of the disease odds, and thus the PS bias is null under constant disease risk. However, if the latter condition holds but the sampling is not SRS, the bias may still exist. The magnitude of the bias increases as the deviation between the true sampling and SRS increases. We also derive bounds for the bias. If the bounds are approximately known or estimable, we show that this information can be used to compute a conservative p value for the usual association test. We give two real examples to demonstrate the application of the new method.

Copyright © 2010 S. Karger AG, Basel

Introduction

The case-control study is a widely used approach for investigating associations between candidate genes and dichotomous disease traits. However, as has been pointed out by many authors [1–6], such studies are susceptible to the population stratification (PS) if the gene under study shows marked variation in allele frequency across subpopulations and these subpopulations also differ in their baseline disease risk. Although less mentioned, sampling design of the cases and controls is also an important factor influencing the level of PS. Some researchers showed that the PS usually has larger influence when the cases and controls are sampled from nearly different strata [7–11]. Recently, Ardlie et al. [12] found weak evidence for the PS structure in the African American sample. But the stratification can be eliminated by matching the sample based on self-reported ancestry. Campbell et al. [13] also found a SNP in the gene LCT strongly associated with height, but re-matching individuals on the basis of European ancestry also greatly reduced the false association.

The magnitude of the PS is difficult to measure or estimate. Many authors have suggested quantitative methods to control type I errors. The most popular treatments include the ‘genomic control’ method [7, 9, 13–16] and the ‘structure association’ method [17–21]. Each of the proposed methods requires typing of extra polymorphic

markers to generate an estimate of the PS which can be used to adjust the test statistic. Recently, Lee and Wang [22] proposed an interesting alternative using bounds for the PS bias to make correction for the association test. However, their bounds are valid only when the data collected in the study follow the simple random sampling (SRS) design.

In this report, we present PS bias for the association test when the genetic effect is null and the sampling design more general than the SRS design. We show that matching effectiveness of the sampling plays an important role in determining the level of the PS bias. Slight improvement of the sampling method toward perfect matching may greatly reduce the PS bias. In the case of SRS, the matching effectiveness of the sampling equals the inverse of the variation of disease odds. Thus, null PS bias is valid only when both SRS and identical disease risk conditions are satisfied. We present type I errors of the usual χ^2 test to show the presence of the PS bias when either condition fails.

We also derive bounds for the PS bias. Our bounds automatically reduce to the bounds of Lee and Wang [22] when the SRS condition holds. However, if the SRS condition fails, the latter bounds are no longer valid. We show that if the values of the genotype frequency variation and matching effectiveness are roughly known or estimable in real applications, then one can compute conservative p values for the usual association test. The conservative p values can be used to correct the usual association test in case-control studies. If the conservative p value is less than the predetermined significance level, we can safely claim that the gene-disease association is truly significant, even when there is PS. In this report, we present some properties of the corrected tests and also demonstrate applications of the new method with two examples.

Material and Methods

The Magnitude of the Bias

We focus the discussion on the joint effect of the PS and sampling design when the genetic effect is null. The disease status is denoted as D with levels $d = 1, 0$, indicating the presence and absence of disease, respectively. $G = 1 (0)$ represents the presence (absence) of the genotype of interest. S denotes the general stratification variable with values $s = 1, \dots, K$, representing K population strata. In our discussion of the bias effect, the stratification variable S is assumed to be not observable and the risk model is given by

$$\text{logit } P(D = 1|G = g, S = s) = \mu' + \alpha'_s + \beta g.$$

We define $\alpha'_1 = 0$ for the purpose of identifiability. α'_s is the subpopulation-specific intercept parameter for accounting for pos-

sible heterogeneity in disease risks between different subpopulations. The same risk model was also considered by Satten et al. [20] and Cheng and Lin [23]. Under this risk model, log-odds-ratio β measures the association between the genotype of interest and disease.

In our discussion of the bias effect, stratification variable S is assumed to be not observable in the study. We use $P(S = s|D = d)$ to denote the expected proportions of the subpopulation s in the case ($d = 1$) and control ($d = 0$) samples if the ascertainment of the cases and controls follow the SRS design. However, in practical applications, the true expectation of the sample proportions in the cases and controls of a study may be different from $P(S = s|D = d)$. For example, if the studied cases are randomly selected from hospital-based cases, but the hospital-based cases are not representative of the population-based cases, then $P(S = s|D = 1)$ may not be the expectation of the sample proportions in the case sample. Even the ascertainment of the case and control samples follow SRS from the general population, the non-response of the study subjects could also cause the expectation of the sample proportions significantly differing from $P(S = s|D = d)$. We use $P^*(S = s|D = d)$ to denote the true expectation of the sample proportions of the subpopulation s in the case and control samples. We do not require any condition on $P^*(S = s|D = d)$ and any design defining general expected proportions $P^*(S = s|D = d)$ is called the general sampling design (including the SRS design) in this paper. Under the general sampling design, we can show (the details are given in the Appendix) that the prospective risk, given only the genotype information, is

$$\text{logit } P(D = 1|G = g) = \mu + (\beta + \beta^*)g.$$

Here, the new parameters are defined by

$$\mu = \log\{K(0)\} + \log\{P(D = 1)/P(D = 0)\},$$

$$\beta^* = \log \left\{ \frac{K(1)}{K(0)} \right\}, \text{ where}$$

$$K(g) = \frac{\sum_{s=1}^k P(G = g | S = s, D = 0) P^*(S = s|D = 0) S_s}{\sum_{s=1}^k P(G = g | S = s, D = 0) P^*(S = s|D = 0)}.$$

Quantity $S_s = P^*(S = s|D = 1)/P^*(S = s|D = 0)$ measures the degree of matching with respect to subpopulation s in the study. The matching effectiveness of the sampling is defined as

$$ME^* = \min_s S_s / \max_s S_s.$$

For any quantity X_s , we also define its variation by

$$X^* = \max_s X_s / \min_s X_s.$$

If the sampling of the study follows the SRS design,

$$S_s = \frac{P(D = 1|S = s)}{P(D = 0|S = s)} \bigg/ \frac{P(D = 0)}{P(D = 1)}.$$

Thus, the matching effectiveness equals the inverse of the variation D^* of the disease odds $D_s = P(D = 1|S = s)/P(D = 0|S = s)$. β^* is the PS bias. Based on the expression for the bias, several conclusions can be reached: (1) under perfect matching (that is, $ME^* =$

1), there is no effect of PS. (2) If the baseline genotype frequencies are identical across subpopulations, then there is no effect of PS. (3) If the underlying sampling is simple and random, and if the disease rates are identical across subpopulations, then there is no effect of PS. However, if the sampling is not SRS, then the conclusion may be invalid.

Measuring Maximal Impacts of Population Stratification

The PS bias is in general not directly measurable. In the following, we present lower and upper bounds for the bias. These bounds depend on the variation G^* of the baseline genotype frequency odds $P(G=1|D=0, S=s)/P(G=0|D=0, S=s) \equiv G_s$ among the related subpopulations, and the matching effectiveness of the sampling design.

We show in the Appendix that the bias $\exp(\beta^*)$ is bounded above by

$$\max_{0 \leq w \leq 1} \frac{wG_M S_M + (1-w)G_m S_m}{(wG_M + (1-w)G_m)(wS_M + (1-w)S_m)} \\ = \frac{\sqrt{G^* ME^*} (\sqrt{G^*} + \sqrt{ME^*})^2}{(\sqrt{G^* ME^*} + G^* ME^*) (\sqrt{G^* ME^*} + 1)} \equiv U_\beta.$$

The bias is also bounded below by $L_\beta \equiv U_\beta^{-1}$. Under SRS, we note that the bounds given here are the same bounds as given in [22].

Conservative p Value

In the presence of PS, the usual χ^2 test statistic does not have a χ^2 distribution under the null hypothesis. Instead, it has a non-central χ^2 distribution, with non-centrality parameter depending on the level of the PS bias. Thus, the usual χ^2 test tends to have inflated type I errors. In the following, we suggest using bounds of the bias to compute a conservative p value. The true type I error of the usual χ^2 test based on using conservative p values is always less than or equal to the predetermined level. When there is no effect of PS, the conservative p value automatically reduces to the usual p value.

Let $\chi^2_1(\Delta)$ represent a non-central χ^2 random variable with one degree of freedom and non-centrality parameter Δ . Under the null hypothesis of no gene-disease association, the usual χ^2 statistic has a non-central χ^2 distribution with one degree of freedom and non-centrality parameter

$$\Delta_\beta = \beta^{*2} / \left(\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}} \right),$$

where n_{dg} is the number of observations with genotype g and disease status d . Thus the true type I error of the usual test with significance level α is given by $\alpha_\chi^2 = P(\chi^2_1(\Delta_\beta) \geq \chi^2_{1;1-\alpha})$, which is always $> \alpha$ when $\Delta_\beta \neq 0$. Here, $\chi^2_{1;1-\alpha}$ is the 100(1 - α) percentile of the χ^2 distribution with one degree of freedom. Based on the bounds derived here, we propose a corrected testing procedure using a conservative p value defined by $p_c = P(\chi^2_1(\Delta_\beta^u) \geq \chi^2)$, where

$$\Delta_\beta^u = \left\{ \log(U_\beta) \right\}^2 / \left(\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}} \right).$$

Table 1. The values of the bias, bounds, and true type I error rates

n	G^*	ME^*	Bias		True type I error rate	
			True	Bound*	α_χ^2	α^*
200	1.5	1	0.0000	0.0000	0.0500	0.0500
		0.5	0.0643	0.0693	0.0609	0.0485
		0.35	0.1002	0.1037	0.0766	0.0486
		0.2	0.1541	0.1544	0.1134	0.0498
	3	1	0.0000	0.0000	0.0500	0.0500
		0.5	0.1707	0.1840	0.1361	0.0436
		0.35	0.2659	0.2754	0.2618	0.0453
		0.2	0.4096	0.4108	0.5235	0.0494
	5	1	0.0000	0.0000	0.0500	0.0500
		0.5	0.2484	0.2625	0.2353	0.0431
		0.35	0.3843	0.3933	0.4806	0.0455
		0.2	0.5874	0.5878	0.8278	0.0498
500	1.5	1	0.0000	0.0000	0.0500	0.0500
		0.5	0.0643	0.0693	0.0776	0.0468
		0.35	0.1002	0.1037	0.1178	0.0474
		0.2	0.1541	0.1544	0.2122	0.0497
	3	1	0.0000	0.0000	0.0500	0.0500
		0.5	0.1707	0.1840	0.2696	0.0401
		0.35	0.2659	0.2754	0.5508	0.0428
		0.2	0.4096	0.4108	0.8910	0.0490
	5	1	0.0000	0.0000	0.0500	0.0500
		0.5	0.2484	0.2625	0.4979	0.0395
		0.35	0.3843	0.3933	0.8558	0.0431
		0.2	0.5874	0.5878	0.9958	0.0497

We suggest that the association test be declared significant, if $p_c < \alpha$. It can be shown that the type I error of the corrected test is given by $\alpha^* = P(\chi^2_1(\Delta_\beta) \geq \chi^2_{1;1-\alpha}(\Delta_\beta^u))$, which is always $\leq \alpha$. This conclusion means that the corrected test is always conservative. If the conservative test is significant, the testing result can serve as a supplementary result to confirm that the gene-disease association is truly significant, regardless of the presence of the population stratification.

Results

Type I Errors

Table 1 shows values of the PS bias β^* , bounds, and type I errors of the usual χ^2 test and corrected test, when the significance level is 0.05. We assume that there are two subpopulations ($K = 2$), baseline genotype frequency of the first subpopulation is given by $P(G = 1|D = 0, S = 1) = 0.3$, the sampling fraction of subpopulation 1 in the control sample is 0.3, and case and control sample sizes both

equal to $n = 200$ or 500 . The rest of the parameter values are determined by the variation G^* and matching effectiveness ME^* given in the table. Note that in computing the non-centrality parameters, sample observations n_{dg} are replaced by $n \times P(G = g|D = d)$.

According to table 1, we find that the reduction of the bias can be substantial when matching effectiveness is improved. For example, in all cases considered in the table, if matching effectiveness is increased from 0.2 to 0.5, the amount of bias reduction is more than 50%. We also find that the values of the bounds are close to the true bias. In some non-trivial cases such as $G^* = 1.5$, $ME^* = 0.2$ or $G^* = 5$, $ME^* = 0.35$, the values of the bias and its bound are almost identical. This implies that the type I errors of the corrected test are close to the predetermined 0.05 level. In contrast, the usual χ^2 test has type I errors much larger than 0.05. The largest type I error is 0.8278 when $n = 200$ and 0.9958 when $n = 500$. The type I error rate increases as the variation of baseline genotype frequency odds, or sample size, increases. But if matching effectiveness of the sampling can be improved, the chance of reaching a false-positive result can be reduced.

Table 2 compares the PS bias under SRS and that under general sampling design. Table 2 also compares the true type I error rates of the corrected tests based on using the bound derived here (denoted as $Bound^*$ in table 2.), and that derived by [22] under SRS design (denoted as $Bound'$). The bound of Lee and Wang [22] depends on the variations of the disease odds and baseline genotype frequency. The type I error rate of the corrected test based on using $Bound'$ is denoted by α' . For the simplicity of presentation, we define a new quantity

$$DS = \max_s \left\{ \frac{P^*(S = s|D = 1)}{P(S = s|D = 1)} \middle/ \frac{P^*(S = s|D = 0)}{P(S = s|D = 0)} \right\} \middle/ \min_s \left\{ \frac{P^*(S = s|D = 1)}{P(S = s|D = 1)} \middle/ \frac{P^*(S = s|D = 0)}{P(S = s|D = 0)} \right\}$$

to measure the deviation between the SRS design and true sampling design of the study. If the true sampling is SRS, then $DS = 1$. The values of the parameters G^* , D^* , n and DS are given in table 2. In addition, we assume that $P(G = 1|D = 0, S = 1) = 0.3$, disease rate $P(D = 1|S = 1) = 0.1$, $P(S = 1) = 0.7$ and $P^*(S = 1|D = 0) = P(S = 1|D = 0)$, so that all required parameters can be determined.

From table 2, we first find out that the sampling design greatly modifies the PS effect. For example, when $n = 200$, $G^* = 3$, $D^* = 3$, the PS bias is 0.2778 under SRS, but it becomes 0.6211 under the sampling design with $DS = 5$. In

addition, under all conditions in table 2, the values of the bias and their bounds follow an interesting ordering: $Bound' \leq \text{true bias} \leq Bound^*$. This result means that the bounds derived under the SRS condition need to be cautiously used. Since under moderate deviation from SRS (for example, $DS = 3$), the corresponding true type I error α' increases dramatically. The difference between $Bound'$ and $Bound^*$ also becomes larger when the deviation between two samplings gets larger. This implies that the performance difference in type I error between the two corrected tests becomes larger. Table 2 shows that the largest type I error α' is 0.8364 when $n = 200$ and 0.9939 when $n = 500$. In contrast, the corrected test based on the new bound maintains satisfactory performance under almost all conditions considered in table 2.

Examples

Campbell et al. [13] studied a European American case-control panel comprised of 1,057 individuals ranked in the 5th through 10th percentiles for adult height and 1,132 individuals ranked in the 90th through 95th percentiles for adult height. All individuals were born in the USA and self-described white or Caucasian, and all of their grandparents were born in either Europe or the USA. They genotyped the LCT-1910C→T polymorphism (rs4988235) and investigated the association between LCT marker and height. They found the genotype of interest strongly associated with tall stature (OR = 1.673 for genotype TT+TC versus genotype CC; confidence interval: 1.3406–2.0866; $\chi^2 = 20.7151$, $p = 6.258 \times 10^{-6}$). Campbell et al. [13] further matched the cases and controls on the basis of grandparental ancestry and found that the re-matching greatly reduced association. Their re-matching gave three subgroups: subjects with four US-born grandparents, subjects with predominantly northwestern European ancestry, and subjects with predominantly southeastern ancestry. Based on the sample proportions in the cases and controls for the three subgroups, we find that the degrees of matching S_s are respectively equal to 1.08, 0.596, and 1.953. Thus the matching effectiveness is roughly equal to $ME^* = 0.596/1.953 = 0.305$. The baseline CC genotype frequencies for the three groups can be estimated from the control samples and are respectively equal to 0.12, 0.56, and 0.14. Thus the variation of the baseline genotype frequency odds is roughly equal to $G^* = 9.20$. Using the new bound, we find that the conservative p value is $p_c = 0.7287$. Campbell et al. [13] suggested that the original conclusion of strong association may be completely or largely due to PS. Our result, on the other hand, suggests

Table 2. The values of the bias, bounds, and true type I error rates

n	Condition			Bias			True type I error rate		
	G^*	D^*	DS	True	Bound'	Bound*	α_{χ^2}	α'	α^*
200	3	1	1	0.0000	0.0000	0.0000	0.0500	0.0500	0.0500
			3	0.2847	0.0000	0.2877	0.2875	0.2875	0.0485
			5	0.4107	0.0000	0.4108	0.5246	0.5245	0.0499
		3	1	0.2778	0.2416	0.2877	0.2744	0.0711	0.0452
			3	0.5370	0.2416	0.5391	0.7508	0.4229	0.0489
			5	0.6211	0.2416	0.6372	0.8618	0.5860	0.0423
		5	1	0.4035	0.3308	0.4108	0.5055	0.0986	0.0464
			3	0.6310	0.3308	0.6372	0.8710	0.4308	0.0469
			5	0.6976	0.3308	0.7223	0.9271	0.5598	0.0387
	5	1	1	0.0000	0.0000	0.0000	0.0500	0.0500	0.0500
			3	0.4043	0.0000	0.4108	0.5185	0.5184	0.0468
			5	0.5877	0.0000	0.5878	0.8280	0.8280	0.0500
		3	1	0.3934	0.3448	0.4108	0.4950	0.0801	0.0417
			3	0.7725	0.3448	0.7734	0.9665	0.6751	0.0495
			5	0.9005	0.3448	0.9165	0.9924	0.8562	0.0425
		5	1	0.5734	0.4726	0.5878	0.8074	0.1255	0.0431
			3	0.9117	0.4726	0.9165	0.9935	0.6895	0.0476
			5	1.0148	0.4726	1.0415	0.9984	0.8364	0.0380
500	3	1	1	0.0000	0.0000	0.0000	0.0500	0.0500	0.0500
			3	0.2847	0.0000	0.2877	0.5990	0.5990	0.0477
			5	0.4107	0.0000	0.4108	0.8918	0.8918	0.0499
		3	1	0.2778	0.2416	0.2877	0.5749	0.0862	0.0426
			3	0.5370	0.2416	0.5391	0.9864	0.7417	0.0483
			5	0.6211	0.2416	0.6372	0.9979	0.9032	0.0383
		5	1	0.4035	0.3308	0.4108	0.8771	0.1395	0.0444
			3	0.6310	0.3308	0.6372	0.9983	0.7518	0.0452
			5	0.6976	0.3308	0.7223	0.9997	0.8837	0.0332
	5	1	1	0.0000	0.0000	0.0000	0.0500	0.0500	0.0500
			3	0.4043	0.0000	0.4108	0.8873	0.8872	0.0449
			5	0.5877	0.0000	0.5878	0.9958	0.9958	0.0499
		3	1	0.3934	0.3448	0.4108	0.8684	0.1030	0.0375
			3	0.7725	0.3448	0.7734	1.0000	0.9529	0.0492
			5	0.9005	0.3448	0.9165	1.0000	0.9958	0.0385
		5	1	0.5734	0.4726	0.5878	0.9940	0.1951	0.0394
			3	0.9117	0.4726	0.9165	1.0000	0.9589	0.0463
			5	1.0148	0.4726	1.0415	1.0000	0.9939	0.0322

that it is not safe to conclude the existence of the association between LCT marker and height.

The second example is from Ardlie et al. [12]. They studied the association between PPAR γ Pro12A1a polymorphism and type 2 diabetes using two case-control samples (Polish and US whites, respectively). They used the program *structure* [18, 19] and unlinked markers to detect the presence of cryptic population structure. They found that a single population best fits the US and Polish sample data, as well as the combined US/Polish sample. The odds ratio of the combined sample is OR = 0.372 (ge-

notype CC vs. genotype GC+GG; 95% confidence interval: 0.178–0.775; $\chi^2 = 6.976$, $p = 8.26e-3$). Based on these two populations, the matching effectiveness of the combined sample is $ME^* = 0.952$. The baseline C allele frequencies are 0.1 and 0.2, respectively, for the two samples and they are significantly different. Thus, the genotype frequency odds variation is $G^* = 4.125$. Using this information, we find that the conservative p value equals $p_c = 8.32e-3$, showing a strong association result. Here, we also point out that due to high matching effectiveness, the usual p value and the conservative p value are almost identical.

Discussion

In this report, we investigate the joint effect of sampling and PS. We point out that the PS bias may be small under SRS but it could become large under a different sampling design. We also point out that it is important to improve the matching effectiveness of the sampling so that the PS bias can be reduced. The SRS-based data collection in a case-control study is not the best design for removing or reducing PS bias. We show situations where the bounds derived under the SRS condition may not be bounds for the true PS bias, if the SRS condition fails. Under these situations, the association test with improper correction (using *Bound'*) tends to have inflated type I errors like the usual χ^2 test does in the presence of PS. This conclusion suggests that before using the corrected test suggested in [22], one needs to first assure that the SRS condition is satisfied.

We also obtain bounds for the PS bias without assuming any condition about the sampling design. These bounds depend on the variation of the baseline genotype frequency odds and matching effectiveness of the sampling. In the case of SRS, the matching effectiveness reduces to the inverse of the variation of the disease odds of the related subpopulations. Thus the new bounds derived here and the bounds derived by Lee and Wang [22] agree. All bounds discussed here have potential applications in constructing corrected tests, if values of the bounds are known. Tables 1 and 2 show the performance of the corrected tests. Under the same conditions and sample sizes $n_0 = n_1 = n = 200$, table 3 shows additional results about the performance of the corrected test, when the matching effectiveness is measured incorrectly. We use

$$P^* = \frac{ME_w^* - ME^*}{ME^*}$$

to represent the relative difference of the working and true matching effectiveness. For example, if true $ME^* = 0.5$ and $P^* = 0.3$, then the working ME^* value is 0.65. We use α_w^* to denote the type I error of the corrected test when working ME^* is applied. The results from table 3 show that if the relative difference is positive, then the corrected test is less conservative. Unreported results indicate that if P^* is negative, then the corrected test becomes more conservative. However, we found from table 3 that except under the extreme case with $ME^* = 0.5$ and $P^* = 0.3$, the corrected test seems to have rather stable performance (the range of type I errors is (0.0453, 0.0775)). We also have done similar sensitivity analysis for the variation of the baseline genotype frequency odds.

Table 3. True type I errors of the conservative test when matching effectiveness is measured incorrectly

Condition			Bias		True type I error rate
G^*	ME^*	P^*	True	Bound*	α_w^*
1.5	0.75	0.0	0.0252	0.0290	0.0495
		0.15	0.0252	0.0149	0.0511
		0.3	0.0252	0.0026	0.0517
	0.5	0.0	0.0643	0.0693	0.0485
		0.15	0.0643	0.0556	0.0525
		0.3	0.0643	0.0434	0.0556
3	0.75	0.0	0.0670	0.0770	0.0466
		0.15	0.0670	0.0396	0.0580
		0.3	0.0670	0.0068	0.0628
	0.5	0.0	0.1707	0.1840	0.0436
		0.15	0.1707	0.1474	0.0627
		0.3	0.1707	0.1150	0.0830
5	0.75	0.0	0.0984	0.1097	0.0453
		0.15	0.0984	0.0565	0.0669
		0.3	0.0984	0.0097	0.0775
	0.5	0.0	0.2484	0.2625	0.0431
		0.15	0.2484	0.2102	0.0727
		0.3	0.2484	0.1640	0.1094

ME^* is the true value of the match effectiveness, and ME_w^* ($\geq ME^*$) is the working value of the matching effectiveness used in the analysis. $P^* = (ME_w^* - ME^*)/ME^*$.

Unreported results show that if the relative difference of the working and true G^* is less than or equal to 0.3, then the range of the type I errors of the corrected test is (0.0389, 0.0548) when $ME^* = 0.75$. The type I error of the corrected test seems to be less sensitive to the choice of the working value of G^* .

Although the difference of the PS bias and its bound are small, the corrected test seems to be not too conservative under the conditions given in this paper. However, we point out that if there exists genetic effect, then the difference of the PS bias and the bound derived in this paper or in Lee and Wang [22] should depend on the magnitude of the genetic effect. In this scenario, the difference may be large, and it is possible that the real association signal may be missed. In the future, more study on this issue should be of importance. At this moment, we emphasize that the proposed conservative p value approach merely provides a method for helping researchers make more prudent interpretations of their (potentially biased) results.

Computing bounds requires extra information about the variation of baseline genotype frequency odds and matching effectiveness of the sampling. As noted by Lee

and Wang [22], sometimes, one may make an educated guess about the variation of the frequency odds based on one's best knowledge about the population under study, or over-exaggerate the value of G^* to obtain more conservative bounds. The value of the matching effectiveness can be obtained from the case-control data. However, one needs to identify, at least approximately, subpopulations of individuals who are genetically similar. Campbell et al. [13] found that having grandparental country-of-origin information can be valuable for dividing individuals into genetically different groups. The matching effectiveness of the sampling can be computed based on the definition of groups and case-control data. If the information from self-reported family ancestry is not certain to control for unmeasured, unknown population ancestry differences, one can apply the method by Pritchard and Rosenberg [25], for example, to check whether the level of structuring is small enough in each group. Recently, Guan et al. [26] used a modest number of genetic markers to describe a genetic similarity score matching method so that matched cases and controls have similar genetic background. Their approach also can be applied to compute ME^* .

Acknowledgements

This research was supported in part by grants from the National Science Council and Taiwan Department of Clinical Trial and Research Center of Excellence (DOH99-TD-B-111-004), and a joint research grant from China Medical University and Asia University. The authors would like to thank reviewers for their constructive comments which have improved the presentation of this paper.

Appendix

The Prospective Risk under Population Stratification

Following the usual Bayesian argument, the disease-risk model implies that

$$\Pr(G = g|S = s, D = 1)/\Pr(G = g|S = s, D = 0) = \exp(\mu' + \alpha_s + \beta g),$$

where $\alpha_s = \alpha'_s + \log\{\Pr(D = 0, S = s)/\Pr(D = 1, S = s)\}$, $s = 2, \dots, k$. As a consequence, the frequency distribution of G in the case population can be shown to be given by

$$\Pr(G = g|D = 1) = \exp(\beta g) \sum_{s=1}^k \Pr(G = g|S = s, D = 0) P^*(S = s|D = 1) \exp(\mu' + \alpha_s).$$

On the other hand, the frequency distribution of G in the control population is given by

$$\Pr(G = g|D = 0) = \sum_{s=1}^k \Pr(G = g|S = s, D = 0) P^*(S = s|D = 0).$$

Their ratio is given by

$$\Pr(G = g|D = 1)/\Pr(G = g|D = 0) = \exp(\beta_g) K(g) = \exp\{\mu^* + (\beta + \beta^*)g\}.$$

Here, we define $\mu^* = \log\{K(0)\}$,

$$\beta^* = \log\left\{\frac{K(1)}{K(0)}\right\}, \text{ where}$$

$$K(g) = \frac{\sum_{s=1}^k \Pr(G = g|S = s, D = 0) P^*(S = s|D = 0) S_s \exp(\mu' + \alpha_s)}{\sum_{s=1}^k \Pr(G = g|S = s, D = 0) P^*(S = s|D = 0)}.$$

Using the Bayesian argument again, we show that

$$\text{logitPr}(D = 1|G = g) = \mu + (\beta + \beta^*)g,$$

where $\mu = \mu^* + \log\{P(D = 1)/P(D = 0)\}$. Note that in general, $\exp(\mu' + \alpha_s)$ depends on the level of the genetic effect. Thus the PS bias also depends on β . However, under the null hypothesis of no genetic effect,

$$\Pr(G = g|S = s, D = 1)/\Pr(G = g|S = s, D = 0) = \exp(\mu' + \alpha_s),$$

and hence $\exp(\mu' + \alpha_s) = 1$ for each s , since $\Pr(G = g|S = s, D = d)$ are probability functions for $d = 0$ and 1 .

Derivation of Bounds for the Population Stratification Bias

The bias $\exp(\beta^*)$ can be expressed as

$$\frac{\sum_{s=1}^k G_s S_s w_s}{\sum_{s=1}^k G_s w_s \sum_{s=1}^k S_s w_s},$$

where the weights w_s are given by

$$w_s = \frac{P(G = 0|S = s, D = 0) P^*(S = s|D = 0)}{\sum_{s'=1}^k P(G = 0|S = s', D = 0) P^*(S = s'|D = 0)}.$$

Simple algebra shows that there exists some constant $0 \leq w_b \leq 1$, such that the following inequalities hold:

$$\begin{aligned} \exp(\beta^*) &\leq \frac{w_b G_M S_M + (1 - w_b) G_m S_m}{(w_b G_M + (1 - w_b) G_m)(w_b S_M + (1 - w_b) S_m)} \\ &\leq \max_{0 \leq w \leq 1} \frac{w G_M S_M + (1 - w) G_m S_m}{(w G_M + (1 - w) G_m)(w S_M + (1 - w) S_m)} \\ &= \frac{\sqrt{G^* ME^*} (\sqrt{G^*} + \sqrt{ME^*})^2}{\sqrt{G^* ME^*} + G^* ME^*} \left(\sqrt{G^* ME^*} + 1 \right) \equiv U_{\beta}. \end{aligned}$$

A similar approach shows that the lower bound is given by U_{β}^{-1} .

References

- 1 Lander ES, Schork NJ: Genetic dissection of complex traits. *Science (Wash. DC)* 1994; 265:2037–2048.
- 2 Ewens WJ, Spielman RS: The transmission/disequilibrium test: history, subdivision, and admixture. *Am J Hum Genet* 1995;57: 455–464.
- 3 Altschuler D, Kruglyak L, Lander ES: Genetic polymorphism and disease. *N Engl J Med* 1998;338:1626.
- 4 Witte JS, Gauderman WJ, Thomas DC: Population stratification in association studies. *Genet Epidemiol* 1998;15:538.
- 5 Khoury MJ, Beaty TH: Applications of the case-control method in genetic epidemiology. *Epidemiol Rev* 1994;16:134–150.
- 6 Khoury MJ, Yang Q: The future of genetic studies of complex human diseases: an epidemiologic perspective. *Epidemiology* 1998; 9:350–354.
- 7 Bacanu SA, Devlin B, Roeder K: The power of genomic control. *Am J Hum Genet* 2000; 66:1933–1944.
- 8 Chakraborty R: NRC report on DNA typing. *Science* 1993;260:5111.
- 9 Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 1999;55:997–1004.
- 10 Morton NE: The future of genetic epidemiology. *Ann Med* 1992;24:557–562.
- 11 Wacholder S, Rothman N, Caporaso N: Population stratification in epidemiologic studies of common genetic variants and cancer: Quantification of bias. *J Natl Cancer Inst* 2000;92:1151–1158.
- 12 Ardlie KG, Lunetta KL, Seielstad M: Testing for population subdivision and association in four case-control studies. *Am J Hum Genet* 2002;71:1478–1480.
- 13 Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altschuler D, Ardlie KG, Hirschhorn JN: Demonstrating stratification in a European American population. *Nat Genet* 2005;37:868–872.
- 14 Devlin B, Roeder K, Wasserman L: Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol* 2001;60: 155–166.
- 15 Devlin B, Roeder K, Bacanu SA: Unbiased methods for population based association studies. *Genet Epidemiol* 2001;21:273–284.
- 16 Bacanu SA, Devlin B, Roeder K: Association studies for quantitative traits in structured populations. *Genet Epidemiol* 2002;22:78–93.
- 17 Pritchard JK, Donnelly P: Case-control studies of association in structured or admixed populations. *Theor Popul Biol* 2001;60:227–237.
- 18 Pritchard JK, Stephens M, Donnelly P: Inference of population structure using multilocus genotype data. *Genetics* 2000;155:945–959.
- 19 Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: Association mapping in structured populations. *Am J Hum Genet* 2000; 67:170–181.
- 20 Satten GA, Flanders WD, Yang Q: Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet* 2001;68:466–477.
- 21 Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ESI: Dwarf8 polymorphisms associate with variation in flowering time. *Nat Genet* 2001;28:286–289.
- 22 Lee WC, Wang LY: Simple formulas for gauging the potential impacts of population stratification bias. *Am J Epidemiol* 2008;167:86–89.
- 23 Cheng KF, Lin WJ: Simultaneously correcting for population stratification and for genotyping error in case-control association studies. *Am J Hum Genet* 2007;81:726–743.
- 24 Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN: Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* 2004;74:1111–1120.
- 25 Pritchard JK, Rosenberg NA: Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 1999;65:220–228.
- 26 Guan W, Liang L, Boehnke M, Abecasis GR: Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies. *Genet Epidemiol* 2009;DOI:10.1002/gepi.20403.