Original Paper

Human Heredity

Hum Hered 2004;58:40–48
DOI: 10.1159/000081455

Received: April 20, 2004
Accepted after revision: August 6, 2004

Effect of Population Stratification on Case-Control Association Studies

II. False-Positive Rates and Their Limiting Behavior as Number of Subpopulations Increases

Prakash Gorroochurn[a,d] Susan E. Hodge[a,c] Gary Heiman[b]
David A. Greenberg[a,c]

Division of Statistical Genetics, Departments of [a]Biostatistics and [b]Epidemiology, [c]Clinical-Genetic Epidemiology Unit, New York State Psychiatric Institute, New York, N.Y., USA, and [d]University of Mauritius, Reduit, Mauritius

Key Words
Population stratification · Association studies · Type I error · Bias

Abstract
There has been considerable debate in the literature concerning bias in case-control association mapping studies due to population stratification. In this paper, we perform a theoretical analysis of the effects of population stratification by measuring the inflation in the test's type I error (or false-positive rate). Using a model of stratified sampling, we derive an exact expression for the type I error as a function of population parameters and sample size. We give necessary and sufficient conditions for the bias to vanish when there is no statistical association between disease and marker genotype in each of the subpopulations making up the total population. We also investigate the variation of bias with increasing subpopulations and show, both theoretically and by using simulations, that the bias can sometimes be quite substantial even with a very large number of subpopulations. In a companion simulation-based paper (Heiman et al., Part I, this issue), we have focused on the CRR (confounding risk ratio) and its relationship to the type I error in the case of two subpopulations, and have also quantified the magnitude of the type I error that can occur with relatively low CRR values.

Introduction

There is now an abundance of studies conducted to detect genetic association between marker genotypes and complex diseases [5]. The designs usually involve case-control samples. The essential premise of these gene mapping studies is that, since in affected individuals (the cases) there are significantly more disease genes than in the unaffected individuals (the controls), then any marker in linkage disequilibrium a disease gene would also be expected to show significant genotype frequency differences between the two groups.

Case-control studies, however, have been subjected to much criticism because they can show associations between disease and marker loci even when in fact there is no true association. This can occur in situations where there is considerable population stratification or admixture [1, 2, 5–8]. Population stratification refers to the existence of subpopulations within a population, such that

KARGER

Fax + 41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

Accessible online at:
www.karger.com/hhe

Prakash Gorroochurn
Columbia University, Division of Statistical Genetics
722 W. 168th Street, Room R-620
New York, NY 10032 (USA)
Tel. +1 212 342 1263, Fax +1 212 342 0484, E-Mail pg2113@columbia.edu

genotype frequencies differ systematically across loci. Population stratification confounds the association between marker genotype and disease. This is because if either the disease or any marker genotype has higher frequency in any one subpopulation, then there will be an apparent association between the marker genotype and the disease at the population level, even though there is no true association at the subpopulation level. In the statistical literature, this effect has long been known under the name of Simpson's paradox [9]. It arises here because case-control studies, by their very nature, give rise to non-randomized data. Some reported examples of such false-positive associations include the alleged associations of non-insulin-dependent diabetes mellitus in the Pima and Papago tribes (Native Americans) with a haplotype at the immunoglobin G locus [10], and of prostate cancer in African-American populations with a polymorphism in the *CYP344* gene [11].

One way to tackle the issue is to match the ethnic backgrounds of the cases and controls, a technique often used by epidemiologists. However, this might not altogether get rid of all population stratification, as some 'cryptic stratification' may always remain [7]. A more popular alternative is to use family-based tests of association that avoid the problem of population substructure altogether [12, 13]. Undoubtedly, the most famous of these tests is the transmission/disequilibrium test (TDT) [6, 13], which uses the McNemar statistic.

In spite of the popularity of the TDT, some authors have argued that case-control studies have several advantages over the TDT [7, 14, 15]. For example, family-based samples can be difficult, expensive and time-consuming to collect, and family-based tests are often less powerful than case-control tests. Thus, many authors have continued to use case-control studies but have tried to adjust for population stratification. One method, the so-called genomic control method, uses unlinked markers in the genome to quantify, and then correct for, population stratification [see 7, 14, 16–18]. However, the power of these methods remains unclear [19, 20].

Other authors have also argued that population stratification does not cause serious bias in case-control studies [1, 2]. For instance, Wacholder et al. [1] use the confounding risk ratio (CRR) to quantify the bias resulting from population stratification in a case-control design where marker genotype was unrelated to disease in each subpopulation. The CRR is the ratio of the crude (unadjusted for population stratification) to the adjusted relative risk. An important conclusion from their paper is that, as the number of subpopulations (K) increases, confounding from each subpopulation will tend to cancel, since some subpopulations contribute positive confounding and others contribute negative confounding. Thus the overall bias is expected to decrease. In both their empirical and theoretical studies, Wacholder et al. [1] use information on 'first ancestry' based on the first ethnicity reported to the US census [21] and consider $K = 8$ ethnicities. They conclude that, whereas confounding due to stratification might be important for $K = 2$ or 3, the CRR is close to 1 when $K =$ 4–8, implying that the bias is very small.

In this paper, we theoretically investigate the extent of bias due to population stratification by using the type I error. In a companion paper (part I, this issue), Heiman et al. [4] have performed a related simulation-based study in the case of two subpopulations. There we argue that the type I error is more appropriate than the CRR in a significance-testing framework and, like the CRR, is also distorted by bias. Substantial bias will usually result in an inflated type I error compared to the nominal value of the statistical test. The type I error is a function of both population parameters and the sample (the number of cases and controls). Given necessary population parameters and the number of cases and controls, we obtain an exact expression for the type I error, and we give mathematical criteria that determine the extent of bias in association studies. We also show that, under certain conditions, bias can be substantial even with a large number of subpopulations. Finally, when bias does decrease considerably with increasing subpopulations, we investigate the rate of decrease under different distributions of disease and marker genotypes.

The Statistical Model

We base our mathematical model on the same principles as Pritchard and Rosenberg's model of stratified sampling [22]. Consider a sample of m cases (disease) and n controls (non-disease or unaffected) sampled from a population consisting of K subpopulations (or ethnicities) of relative sizes π_i ($i = 1,2, ..., K$), where

$$\sum_{i=1}^{K} \pi_i = 1.$$

We define the following events for $i = 1, 2, ..., K$: C_i: a person belongs to subpopulation i; D: a person has the disease; M: a person has the marker genotype (i.e. has a genotype containing the marker allele).

Table 1. Contingency table for the number of marker genotypes in cases and controls

	M (marker genotype)	$\overline{M}$ (other genotype(s))	
D (case)	X	$m - X$	m
$\overline{D}$ (control)	Y	$n - Y$	n
	$X + Y$	$(m + n) - (X + Y)$	$m + n$

Within the ith subpopulation, we denote the disease prevalence by d_i and the marker genotype frequency by r_i. Then

$$\pi_i = \Pr\{C_i\}, d_i = \Pr\{D|C_i\}, r_i = \Pr\{M|C_i\}, i = 1, 2, ..., K. \quad (1)$$

In each subpopulation, we assume that D is independent of M, i.e.

$$\Pr\{D \cap M|C_i\} = \Pr\{D|C_i\}\Pr\{M|C_i\} = d_i r_i.$$

The Type I Error

Given the numbers of cases (m) and controls (n), the subpopulation relative sizes $\boldsymbol{\pi} = \{\pi_i\}$, the disease prevalences $\mathbf{d} = \{d_i\}$ and the marker genotype frequencies $\mathbf{r} = \{r_i\}$, an exact expression for the expected type I error can be obtained. Consider the following contingency table (table 1) for cases and controls.

Let X and Y denote the number of marker genotypes in the cases and controls, respectively. Conditional on the row totals (m and n), we see that X and Y are independent binomial variates:

$$\Pr\{X = x, Y = y\} = \Pr\{X = x\}\Pr\{Y = y\} \quad (2)$$

and

$$X \sim Bin(m, s), Y \sim Bin(n, t), \quad (3)$$

where, by using Bayes' Theorem,

$$s \equiv \Pr\{M|D\} = \frac{\sum_{i=1}^{K} \pi_i d_i r_i}{\sum_{i=1}^{K} \pi_i d_i}, \mathrm{t} \equiv \Pr\{M|\overline{D}\} = \frac{\sum_{i=1}^{K} \pi_i (1 - d_i) r_i}{1 - \sum_{i=1}^{K} \pi_i d_i}. \quad (4)$$

Thus, s and t represent the proportions of diseased and non-diseased individuals, respectively, in the total population carrying the marker genotype.

Suppose we did not know the actual values of s and t, but we had only an observed contingency table (with realized values x and y of X and Y, respectively). A test of no association between D (disease) and M (marker genotype) is really a test of the hypothesis:

$$H_0: s = t \quad \text{vs.} \quad H_a: s \neq t. \quad (5)$$

Let the nominal type I error of the statistical test T (such as the Fisher's exact test or the chi-squared test) be set to α. The test's actual type I error, as a function s, t, m and n, is then the sum of $\Pr\{X = x, Y = y\}$ over all possible values of x and y such that the realized contingency table is significant at α, i.e. has a significance level (or p value) given by $pval \leq \alpha$. Writing the type I error as $p(s,t; m,n)$, we thus have

$$p(s,t; m,n) = \sum_{y=0}^{n} \sum_{x=0}^{m} I(pval \leq \alpha|x,y,m,n)\Pr\{X = x\}\Pr\{Y = y\}$$

$$= \sum_{y=0}^{n} \sum_{x=0}^{m} I(pval \leq \alpha|x,y,m,n) \times$$

$$\binom{m}{x}\binom{n}{y} s^x t^y (1 - s)^{m-x}(1 - t)^{n-y}. \quad (6)$$

In the above, $I(\cdot)$ is the indicator variable defined by conditioning on x, y, m, n:

$$I(pval \leq \alpha) = \begin{cases} 1 & if\ pval \leq \alpha, \\ 0 & otherwise. \end{cases} \quad (7)$$

Eq. (6) shows that the type I error depends on the population parameters (through $\boldsymbol{\pi}$, $\mathbf{d}$ and $\mathbf{r}$) and on the sample (only through m and n). We expect: (i) An inflated type I error (i.e. $p(s,t; m,n) >> \alpha$) when there is substantial bias due to population stratification; (ii) an increase in the type I error when the sample size increases (see Discussion); (iii) a correct type I error (i.e. $p(s,t; m,n) = \alpha$) when all d_i's or all r_i's are the same because then there is no bias due to population stratification.

Choice of the Statistical Test *T*

Which statistical test T for categorical (or count) data should we use to evaluate $p(s,t; m,n)$ in Eq. (6)? We need to determine $p(s,t; m,n)$ quite accurately because we are measuring the extent of bias due to population stratification from the false positive rate. An accurate determination of $p(s,t; m,n)$ is made even harder because the cell frequencies in table 1 are small when x, y are very small or very large in the summation in Eq. (6). In such situations, it is usually recommended to use Fisher's exact test [23, 24]. There are several ways in which a two-sided p value

Table 2. The type I error using four different tests ($\alpha = 0.05$)

	Type I error			
	Fisher's two-sided p value (i)	Fisher's two-sided mid-p value (ii)	chi-squared p value with Yates' correction	uncorrected chi-squared p value
$\boldsymbol{\pi} = (0.5, 0.5)$, $m = n = 25$ $\mathbf{d} = (0.6, 0.5)$, r = (0.3, 0.3)	0.0229	0.4480	0.0224	0.0537
$\boldsymbol{\pi} = (0.05, 0.5)$, $m = n = 50$ $\mathbf{d} = (0.1, 0.9)$, r = (0.5, 0.5)	0.0352	0.0569	0.0352	0.0569
$\boldsymbol{\pi} = (0.2, 0.8)$, $m = 25$, $n = 50$ $\mathbf{d} = (0.1, 0.5)$, r = (0.9, 0.9)	0.0132	0.0347	0.0121	0.0451
$\boldsymbol{\pi} = (0.4, 0.6)$, $m = n = 100$ $\mathbf{d} = (0.2, 0.6)$, r = (0.2, 0.2)	0.0328	0.0477	0.0322	0.0513

for the latter test can be calculated, but two of the most recommended ones are [25]: (i) calculate the probability of the given table, add this probability to the probabilities of all tables more extreme (i.e. less likely) than the given table, then multiply by 2 to give the two-sided p value; (ii) calculate the probability of the given table, add half of this probability to the probabilities of all tables more extreme than the given table, then multiply by 2 to give the two-sided mid-p value. Procedure (i) gives p values very close to those of the corrected (Yates') chi-squared test, whereas procedure (ii) gives p values very close to that of the uncorrected chi-squared test. However, (i) is over-conservative and leads to type I errors consistently much less than the nominal value α of the test. Some simple calculations can illustrate this. We take two subpopulations ($K = 2$) and let the marker genotype frequencies be the same in both (alternatively, we could have let the disease prevalences to be the same). In such a situation, there can be no bias due to population stratification, so the type I error (as calculated from Eq. (6)) should very close to α.

Table 2 shows that Fisher's two-sided p value and the corrected chi-squared p value are close to each other and are both quite conservative in the sense that the tests consistently reject at a rate much less than the nominal alpha (i.e. $p(s,t;\ m,n)$ consistently <0.05). On the other hand, Fisher's two-sided mid-p value and the uncorrected chi-squared p value are also close to each other and perform quite well (i.e. $p(s,t;\ m,n) \approx 0.05$). Since Fisher's two-sided mid-p value is slightly more conservative of these two, we shall use it in our future calculations (as recommended in [25] and [24]).

Population Stratification and the Magnitude $|s - t|$

The magnitude $|s - t|$ measures the average difference in marker genotype frequencies between affected and unaffected people in the total population. The larger this magnitude, the greater the bias due to population stratification (for given m and n). When $s = t$ there is no bias due to population stratification, no matter what the values of m and n, and we expect $p(s,t;\ m,n) = \alpha$. This then means that independence between D and M at the *subpopulation level* translates into independence between D and M at the *population level*. Using Eq. (4), we see that

$$|s - t| = \left| \frac{\sum_{i=1}^{K} \pi_i d_i r_i - \sum_{i=1}^{K} \pi_i d_i \sum_{i=1}^{K} \pi_i r_i}{\left(\sum_{I=1}^{K} \pi_i d_i \right) \left(1 - \sum_{i=1}^{K} \pi_i d_i \right)} \right| . \tag{8}$$

We see that $s = t$ if and only if

$$\sum_{i=1}^{K} \pi_i d_i r_i - \sum_{i=1}^{K} \pi_i d_i \sum_{i=1}^{K} \pi_i r_i = 0. \tag{9}$$

If all K subpopulations are of the same size ($\pi_i = 1/K$ for $i = 1, 2, ..., K$), then Eq. (9) becomes

$$K \sum_{i=1}^{K} d_i r_i - \sum_{i=1}^{K} d_i \sum_{i=1}^{K} r_i = 0, \tag{10}$$

which means that the correlation coefficient of the data points (d_i, r_i), $i = 1, 2, ..., K$ (viewed as realizations of the random variables d and r), is zero. Eq. (10) gives the necessary and sufficient condition for no bias due to population stratification in a population divided into subpopulations of the same size as long as there is no association in

any of the subpopulations. More generally, from Eq. (9), the necessary and sufficient condition for no bias due to population stratification in subpopulations of arbitrary size π_i ($i = 1, 2, ..., K$) is that the *weighted* correlation coefficient of (d_i, r_i) is zero, where the π_i's are the weights.

There are some special cases of Eq. (9). For example, if $d_i = d$ or $r_i = r$ for $i = 1, 2, ..., K$ then $s = t$ and there is no bias due to population stratification. As another example, with two subpopulations of the same size (i.e. $K = 2$, $\pi_i = \pi_2 = 1/2$), then

$$|s - t| = \left| \frac{(d_2 - d_1)(r_2 - r_1)}{(d_1 + d_2)(d_1 + d_2 - 2)} \right|. \quad (11)$$

A feature of Eq. (11) is that, whereas the extent of bias depends on the marker allele frequencies only through their differences, the dependence on disease prevalences is somewhat more complex. This asymmetry in the type of dependence results from the fact that our contingency table is conditional on the rows (cases and controls) rather than on the columns (marker genotype and other genotype(s)). Another feature of Eq. (11) is that $|s - t|$, and hence the bias, is more sensitive to marker genotype frequency differences than to disease prevalence differences. Finally, we have shown a standardized version of Eq. (11) to be highly predictive of the type I error [4] in the case of two subpopulations of the same size.

Bias when the Number of Subpopulations Increases

What is the effect of increasing the number of subpopulations on the magnitude of the bias due to population stratification? We are here interested in the limit

$$\lim_{K\to\infty} (s - t) = \lim_{K\to\infty} \frac{\Sigma_i \pi_i d_i r_i - \Sigma_i \pi_i d_i \Sigma_i \pi_i r_i}{\Sigma_i \pi_i d_i (1 - \Sigma_i \pi_i d_i)}. \quad (12)$$

Let us assume that d_i, r_i re realizations of the random variables d, r in the interval (0, 1), respectively, according to some density functions. By assumption in the Section 'The Statistical Model', d and r are independent. Also, let π_i^* be realizations of the random variable π^* in the interval (0, 1), such that $\pi_i = \pi_i^* / \Sigma_j \pi_i^*$. Then, for large K,

$$\sum_i \pi_i d_i r_i = \frac{\Sigma_i \pi_i^* d_i r_i}{\Sigma_j \pi_j^*} \approx \frac{E(\pi^* dr)}{E\pi^*},$$

$$\sum_i \pi_i d_i = \frac{\Sigma_i \pi_i^* d_i}{\Sigma_j \pi_j^*} \approx \frac{E(\pi^* d)}{E\pi^*},$$

$$\sum_i \pi_i r_i = \frac{\Sigma_i \pi_i^* r_i}{\Sigma_j \pi_j^*} \approx \frac{E(\pi^* r)}{E\pi^*}.$$

The limit in Eq. (12) then becomes

$$\lim_{K\to\infty} (s - t) = \frac{E\pi^* E(\pi^* dr) - E(\pi^* d) E(\pi^* r)}{E(\pi^* d) \{E\pi^* - E(\pi^* d)\}}. \quad (13)$$

We see that

$$\lim_{K\to\infty} (s - t) = 0$$

if and only if

$$E\pi^* E(\pi^* dr) - E(\pi^* d) E(\pi^* r) = 0, \quad (14)$$

which is true when d, r and π^* are mutually independent. The latter is the sufficient condition for no bias due to population stratification in a population divided into 'infinitely-many' subpopulations. For example, d, r and π^* could be mutually independent uniform variates on (l_d, u_d), (l_r, u_r) and (0, 1), respectively, for convenient values of l_d, u_d, l_r, u_r such that $0 < l_d, u_d, l_r, u_r < 1$.

It is interesting to note the difference between Eqs. (9) and (14). They both give conditions for no bias due to population stratification. However, Eq. (9) is a 'sample result', expressed in terms of realizations of the random variables d, r and π^*. (Note that d, r and π^* are values in the subpopulations, which are biological populations rather than statistical populations, so they are random variables rather than parameters [see 26]. On the other hand, Eq. (14) is a '(statistical) population result', expressed in terms of the *expectations* of the above-mentioned random variables, i.e. averaged over randomly selected subpopulations. Therefore, in Eq. (9), we are concerned with the *data points* (π_i^*, d_i, r_i), whereas, in Eq. (14) we are concerned with the *statistical distribution* of π^*, d and r. If the latter are mutually independent, Eq. (14) holds exactly but Eq. (9) holds approximately, the approximation getting better as K increases.

We demonstrated the variation in bias with increasing number of subpopulations numerically by performing four sets of simulations[1], based on the following mutually independent distributions: (i) $d \sim U(0, 1)$, $r \sim U(0,1)$, $\pi^* \sim U(0, 1)$; (ii) $d \sim U(0, 0.1)$, $r \sim U(0, 1)$, $\pi^* \sim U(0, 1)$; (iii) $d \sim U(0, 0.1)$, $r \sim U(0.4, 0.7)$, $\pi^* \sim U(0, 1)$; (iv) $d \sim$ beta(2, 20), $r \sim U(0, 1)$, $\pi^* \sim U(0, 1)$. For example, in (i), for a given value of K, we simulated K values for each of the variates d, r and π^* from their respective uniform distributions. We took $m = n = 100$ and we calculated the type I error from Eq. (6), using $\alpha = 0.05$. We repeated the simulation 100 times and averaged out the 100 values of type I error that we obtained. We then did the same for

[1] Using Maple V release 4.00a.

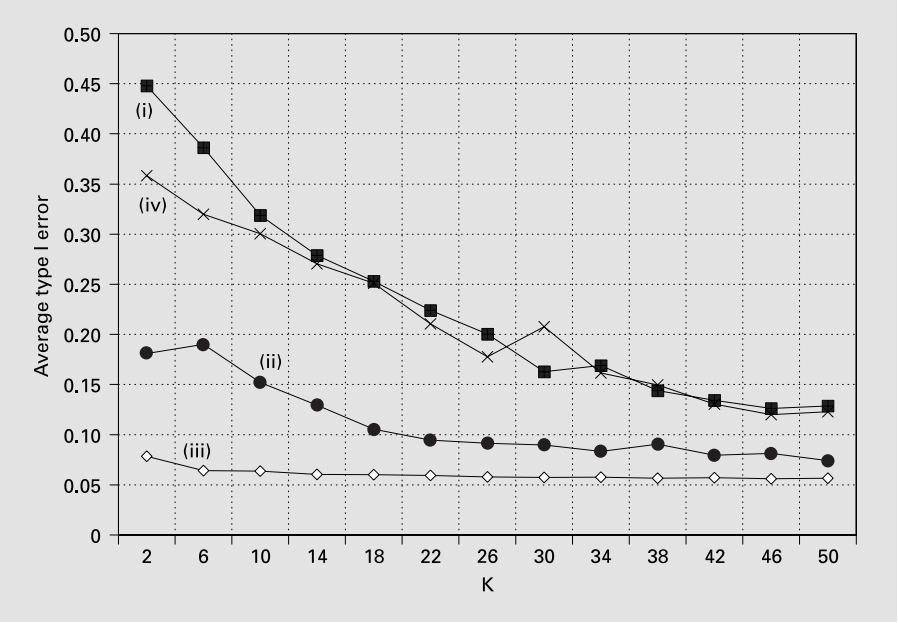

Fig. 1. The variation of average type I error with number of subpopulations (K). (i) $d \sim$ U(0, 1), $r \sim$ U(0, 1), $\pi^* \sim$ U(0, 1); (ii) $d \sim$ U(0, 0.1), $r \sim$ U(0, 1), $\pi^* \sim$ U(0, 1); (iii) $d \sim$ U(0, 0.1), $r \sim$ U(0.4, 0.7), $\pi^* \sim$ U(0, 1); (iv) $d \sim$ beta(2, 20), $r \sim$ U(0, 1), $\pi^* \sim$ U(0, 1). In all cases, test size $\alpha = 0.05$.

different values of K. The same procedure was used in (ii) and (iii) except for the changes in the range of the distributions. In (iv) we used a beta(2, 20) distribution for d (this distribution is highly concentrated around its mean 2/22 = 0.09). Figure 1 shows the variation of the average type I error with K in each of the four cases.

From figure 1, we see that, in all four cases, the type I error decreases with increasing number of subpopulations (though at different rates), since disease, marker genotype and subpopulation size have been assumed to be independent. The type I error decays most quickly in case (iii) when both the ranges of the distributions of d and r are quite narrow. On the other hand, in cases (i) and (iv), when the ranges are as wide as possible, a very large number of subpopulations are required before the type I error gets even close to 0.05. For example, even with 50 subpopulations, the type I error remains over 10%.

The simulations above assume that d, r and π^* are mutually independent. However, it is also possible for disease prevalences to depend on the sizes of subpopulations, so that

$$\lim_{K \to \infty} (s - t) \neq 0,$$

and the bias can stay large with increasing number of subpopulations. For instance, consider the so-called founder effect [27, 28]: very small population sizes create bottlenecks, resulting in a serious loss of heterozygozity and elevated disease prevalence. The effect is more pronounced for rare diseases. Well-documented examples include Tay-Sachs disease in Ashkenazi Jews, Ellis-van Creveld syndrome in the old-order Amish in Pennsylvania, and variegate porphyria among the Afrikaners in South Africa [29, 30].

We demonstrated numerically that bias can be quite substantial with increasing number of subpopulations with three sets of simulations, based on the following distributions: (i) $d \sim$ U(0, 1), $r \sim$ U(0, 1), $\pi^* \sim$U(0, 1); (iii) $d \sim$ U(0, 0.1), $r \sim$ U(0.4, 0.7), $\pi^* \sim$ U(0, 1); (v) $d \sim$ U(0, 1), $r \sim$ U(0, 1), $\pi_i^* = 1/d_i$; (vi) $d \sim$ U(0, 0.1), $r \sim$ U(0.4, 0.7), $\pi_i^* = 1/d_i$. Cases (i) and (iii) are the same as before, but cases (5) and (6) have d and π^* dependent. The results for the four sets of experiments are shown in figures 2a and b.

We see that, in cases (i) and (iii), when we assume mutual independence of d, r and π^*, the type I error decays fairly rapidly close to the nominal α (= 0.05) with increasing number of subpopulations. However, in cases (v) and (vi), when mutual independence no longer holds, the type I error remains substantial even when K is large. The situation is very drastic for case (v), where the ranges are as wide as possible and the disease prevalence is inversely related to the subpopulation size: the type I error decreases erratically but always remains high (nearly 40% with 30 subpopulations and almost 30% with 400 subpopulations!).

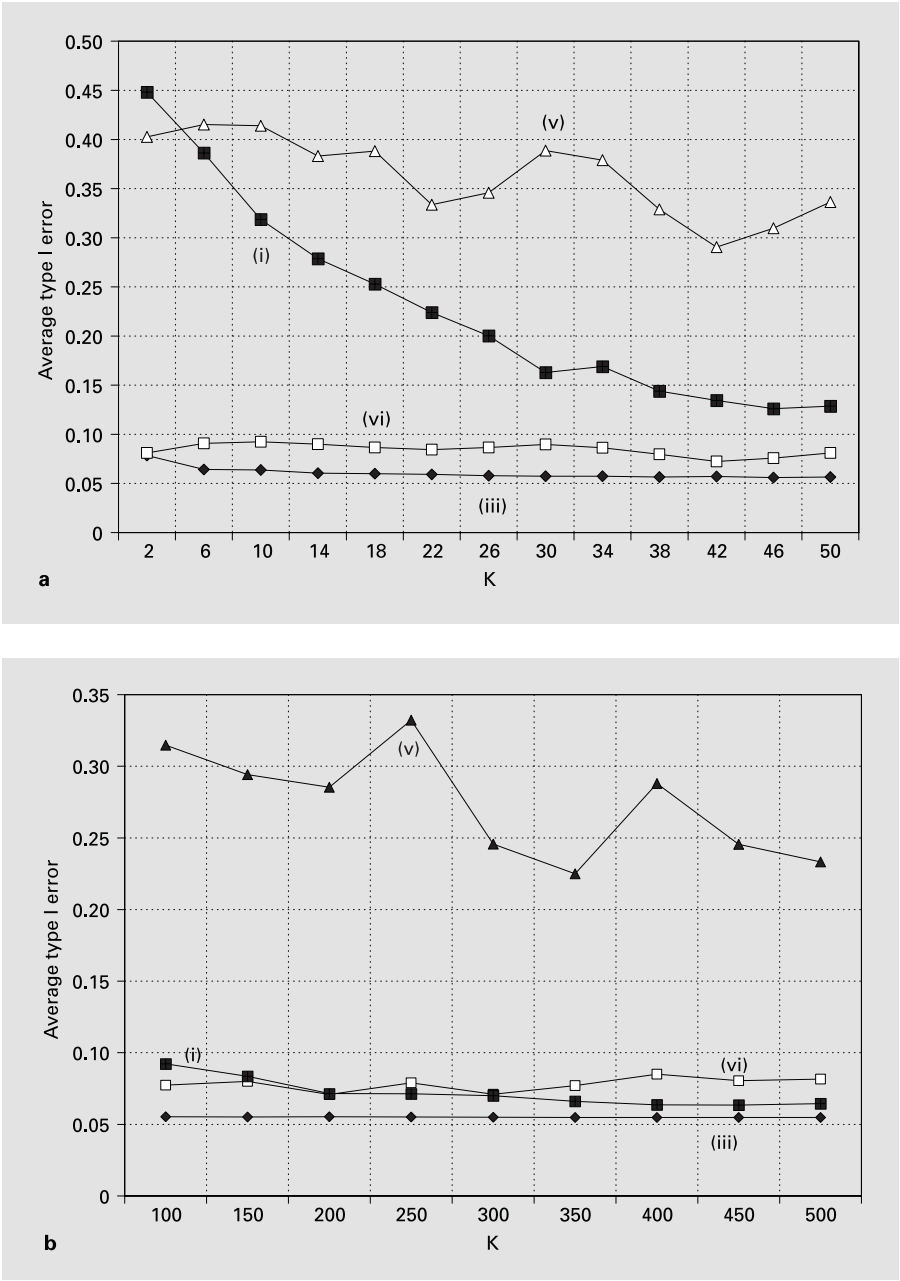

Fig. 2. a The variation of average type I error with number of subpopulations (K) for K = 2, 6, 10, ..., 50. (i) $d \sim U(0, 1)$, $r \sim U(0, 1)$, $\pi^* \sim U(0, 1)$; (iii) $d \sim U(0, 0.1)$, $r \sim U(0.4, 0.7)$, $\pi^* \sim U(0, 1)$; (v) $d \sim U(0, 1)$, $r \sim U(0, 1)$, $\pi^* 1/d_i$; (vi) $d \sim U(0, 0.1)$, $r \sim U(0.4, 0.7)$, $\pi^* = 1/d_i$ (i.e. all random variables not mutually independent). In all cases, test size $\alpha = 0.05$. **b** The variation of average type I error with number of subpopulations (K) for K = 100, 150, ..., 200. (i) $d \sim U(0, 1)$, $r \sim U(0, 1)$, $\pi^* \sim U(0, 1)$; (iii) $d \sim U(0, 0.1)$, $r \sim U(0.4, 0.7)$, $\pi^* \sim U(0, 1)$; (v) $d \sim U(0, 1)$, $r \sim U(0, 1)$, $\pi^* = 1/d_i$; (vi) $d \sim U(0, 0.1)$, $r \sim U(0.4, 0.7)$, $\pi^* = 1/d_i$ (i.e. all random variables not mutually independent). In all cases, test size $\alpha = 0.05$.

Discussion

Our analysis shows that population stratification in case-control association mapping studies may pose a more serious problem than argued by Wacholder et al. [1, 2]. In the absence of any statistical association between disease and marker genotype within the individual subpopulations, population stratification does not give rise to any bias when the subpopulation size-weighted correlation coefficient between disease prevalences and marker allele frequencies is zero (Eq. (9)). The subpopulation size-weighted correlation coefficient is normally approximately zero *either* when (a) the ranges over which the distributions of marker genotype frequency and disease prevalence are defined become narrower (fig. 1), *or* when (b) the distributions of disease prevalence, marker genotype frequency and subpopulation size are mutually independent (fig. 2a). Note in particular that we can still have

negligible bias when condition (a) holds but condition (b) does not hold. That is, we can have disease prevalence, marker genotype frequency and subpopulation size not mutually independent (as in fig. 2a); however, by strongly restricting the ranges of these random variables any correlation can be made to vanish so that Eq. (9) is approximately true. When either condition (a) or (b) holds, then the bias is negligible when the number of subpopulations is even as small as 4–8, as claimed by Wacholder et al. [1, 2]. However, when either (or both) of the ranges becomes wider, the bias increases and can become substantial. For example, when the ranges are as wide as possible, we have an type I error above 30% even with 10 subpopulations (fig. 1, case (i)). The same happens when disease prevalence, marker genotype frequency and subpopulation size become more and more pairwise correlated. For example, with disease prevalence inversely related to subpopulation size, we have an type I error of almost 40% with 30 subpopulations (fig. 2a, case (v)).

We note another way to view the null hypothesis in Eq. (5): Imagine a situation in which we knew or assumed that the disease and marker are *not* associated, but what we wanted to test is whether there is population stratification. Then H_0: $s = t$ would represent a hypothesis of 'no spurious association,' i.e., 'no bias due to population stratification'; and now the test statistical test T (whether Fisher's exact test or the chi-squared test) would represent the *correct* test for this hypothesis. Thus, the sum of probabilities in Eq. (6) would represent *power*, rather than a false positive rate. It is well known that power increases with increasing sample size. Looking at Eq. (6) this way explains why that sum of probabilities, and hence the type I error in our study, has to increase as sample size increases.

With a large number of subpopulations, there is negligible bias due to population stratification when the disease prevalence, marker genotype frequency and subpopulation size are independently distributed (Eq. 14). However the decrease in bias is substantial only when the ranges of marker genotype frequency and disease prevalence are narrow (fig. 2b). When the ranges are wide, there is a decrease in bias but a large number of subpopulations (~ 150) are necessary before the bias is substantially reduced. When disease prevalence, marker genotype frequency and subpopulation size are not independent, bias stays large even with a large number of subpopulations. For example, with the disease prevalence inversely related to the subpopulation size, we have an type I error of almost 30% with 400 subpopulations (fig. 2b, case (v))! Moreover, when the disease prevalence and subpopulation size are inversely related, the bias remains relatively large even though the ranges over which marker genotype frequency and disease prevalence are defined are narrow, and the number of subpopulations is extremely large. For example, with disease prevalence inversely related to subpopulation size and with the ranges of disease and marker quite restricted, we have an type I error of about 8% even with 500 subpopulations (fig. 2b, case (vi)).

Why do the results of Wacholder et al. [1, 2] suggest 'bias from population stratification is unlikely to be substantial'? In the light of the previous paragraphs, there are at least three reasons for this. First, the genotype frequency ranges considered by these authors are quite narrow, but in real data those frequency ranges are not always narrow. For example, from the ALelle FREquency Database [31], genotype frequencies ranges can be quite wide (e.g. for ACE Alu insertion of the ACE locus and the TPA25 Alu insertion of the PLAT locus, the genotype frequency range was 0–1; for the PV92 Alu insertion of the CDH13 locus and the APO Alu insertion of the APOA1 locus, the range was 0.02–1). Further references are given in Thomas and Witte [3]. Second, the authors have not considered the possibility that disease prevalences can depend on the size of subpopulations. Even Thomas and Witte [3], who warn that population stratification 'is of sufficiently serious concern' in case-controls studies, seem to concede that the resulting bias becomes negligible with a large number of subpopulations. However, it is very plausible for disease prevalences to be negatively correlated to subpopulation sizes. For example, three founder mutations have been observed in Ashkenazi Jewish breast and ovarian cancer patients [32]: the *BRCA2* 6174delT mutation (with a frequency of 0.9–1.5%); the *BRCA1* 185delAG mutation (in both Ashkenazi and Sephardic Jews, with a frequency of 0.8–1.1%); and the *BRCA1* 5382insC mutation (with a frequency of 0.13–0.3%). The population prevalences for these three mutations combined is 2–2.5%, which is approximately 10–50 times higher than the allele frequency in the general population. A third possible reason is that Wacholder et al. [1, 2] focus on the CRR instead of the type I error and, as demonstrated by Heiman et al. [4] (part I, this issue), the CRR is somewhat insensitive to bias due to population stratification. Note, however, that when the first two conditions are relaxed, our results do indeed agree with those of Wacholder et al. [1, 2] (fig. 1, 2a, b, case (iii)). Thus, in essence, our results do not really contradict those of the authors, but, if anything, further extends them by exploring different distributions of and dependencies between marker genotype, disease frequency and subpopulation size.

Our method assumes no a priori association between marker genotype and disease in each subpopulation, and then investigates the false association arising in the total population due to population stratification. It would be interesting to see whether the conclusions we reached in this paper would hold even when there existed some a priori real association between genotype and disease within each subpopulation, i.e. whether the effects of population stratification are additive on association studies. This remains an important component of our current research.

Acknowledgments

We thank Drs. Martina Durner and Dvora Shmulewitz for useful discussions on this paper. This work was supported in part by NIH grants AA13654, DK31813, MH48858, DK31775, NS27941, T32 MH65213, MH28274, MH60970.

References

1 Wacholder S, Rothman N, Caporaso N: Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J Natl Can 2000;92(14):1151–1158.

2 Wacholder S, Rothman N, Caporaso N: Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. Cancer Epidemiol Biomarkers Prev 2002;11(6):513–520.

3 Thomas DC, Witte JS: Point: population stratification: A problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prev 2002;11(6):505–512.

4 Heiman GA, Hodge SE, Gorroochurn P, Zhang J, Greenberg DA: Effect of population stratification on case-control studies. I. Elevation in false positive rates and comparison to confounding risk ratios (a simulation study). Hum Hered 2004;58:30–39.

5 Risch NJ: Searching for genetic determinants in the new millennium. Nature 2000;405:847–856.

6 Ewens WJ, Spielman RS: The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet 1995;57(2):455–464.

7 Devlin B, Roeder K: Genomic control for association studies. Biometrics 1999;55(4):997–1004.

8 Colhoun HM, McKeigue PM, Davey SG: Problems of reporting genetic associations with complex outcomes. Lancet 2003;361:865–872.

9 Simpson EH: The interpretation if interaction in contingency tables. J R Stat Soc Ser B 1951; 13:238–241.

10 Knowler WC, Williams RC, Pettitt DJ, Steinberg AG: Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am J Hum Genet 1988;43(4):520–526.

11 Paris PL, Kupelian PA, Hall JM, Williams TL, Levin H, Klein EA, et al: Association between a CYP3A4 genetic variant and clinical presentation in African-American prostate cancer patients. Cancer Epidemiol Biomarkers Prev 1999;8(10):901–905.

12 Falk CT, Rubinstein P: Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 1987;51(Pt 3):227–233.

13 Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52(3):506–516.

14 Reich DE, Goldstein DB: Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol 2001;20(1):4–16.

15 Pritchard JK, Donnelly P: Case-control studies of association in structured or admixed populations. Theor Popul Biol 2001;60(3):227–237.

16 Devlin B, Roeder K, Wasserman L: Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 2001;60(3): 155–166.

17 Marchini J, Cardon LR, Phillips MS, Donnelly P: The effects of human population structure on large genetic association studies. Nat Genet 2004; published online: 28 March 2004; doi:10.1038/ng1337.

18 Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, et al: Assessing the impact of population stratification on genetic association studies. Nat Genet 2004; 36(4):388–393.

19 Chen HS, Zhu X, Zhao H, Zhang S: Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. Ann Hum Genet 2003;67:250–264.

20 Shmulewitz D, Zhang J, Greenberg DA: Case-control studies in mixed populations: Correcting using genomic controls. Hum Hered 2004 (in press).

21 US Bureau of the Census. 1990 census of populations. Vol. 2. Social and economic characteristics. United States. Washington D.C.: U.S. Government Printing Office, 1993.

22 Pritchard JK, Rosenberg NA: Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 1999;65(1):220–228.

23 Stuart A, Ord JK, Arnold A: Advanced Theory of Statistics, Volume 2A: Classical Inference and the Linear Model, ed 6. London, 1999.

24 Agresti A: Categorical Data Analysis, ed 2. Oxford, Blackwell, 2002.

25 Armitage P, Berry G, Matthews JNS: Statistical Methods in Medical Research, ed 4. Blackwell Science LTD, 2002.

26 Rousset F: Inference in spatial population genetics; in Balding DJ, Bishop M, Cannings C (eds): Handbook of Statistical Genetics. Chichester, Wiley, 2001.

27 Holgate P: A Mathematical Study of Founder Principle of Evolutionary Genetics. J Appl Prob 1966;3(1):115–128.

28 Nei M: Molecular Evolutionary Genetics. New York, Columbia University Press, 1975.

29 Gelehrter T, Collins F, Ginsburg D: Principles of Medical Genetics ed 2. Baltimore, Williams & Wilkins, 1998.

30 Hartl D, Clark A: Principles of Population Genetics, ed 3. Sunderland, Sinauer, 1997.

31 Rajeevan H, Osier MV, Cheung KH, Deng H, Druskin L, Heinzen R, et al: ALFRED: the ALelle FREquency Database. Update. Nucleic Acids Res 2003;31(1):270–271.

32 Neuhausen SL: Founder populations and their uses for breast cancer genetics. Breast Cancer Res 2000;2(2):77–81.