Original Paper

Human Heredity

Hum Hered 2004;58:30–39
DOI: 10.1159/000081454

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

Effect of Population Stratification on Case-Control Association Studies

I. Elevation in False Positive Rates and Comparison to Confounding Risk Ratios (a Simulation Study)

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

[a]Department of Epidemiology, [b]Division of Statistical Genetics, Department of Biostatistics, Mailman School of Public Health, Columbia University, and [c]Clinical-Genetic Epidemiology Unit, New York State Psychiatric Institute, New York, N.Y., USA

Key Words
Genetic association · Population stratification · Bias · Case-control studies · Reproducibility of results

Abstract
Objectives: This is the first of two articles discussing the effect of population stratification on the type I error rate (i.e., false positive rate). This paper focuses on the confounding risk ratio (CRR). It is accepted that population stratification (PS) can produce false positive results in case-control genetic association. However, which values of population parameters lead to an increase in type I error rate is unknown. Some believe PS does not represent a serious concern [1, 2], whereas others believe that PS may contribute to contradictory findings in genetic association [3]. We used computer simulations to estimate the effect of PS on type I error rate over a wide range of disease frequencies and marker allele frequencies, and we compared the observed type I error rate to the magnitude of the confounding risk ratio. ***Methods:*** We simulated two populations and mixed them to produce a combined population, specifying 160 different combinations of input parameters (disease prevalences and marker allele frequencies in the two populations). From the combined populations, we selected 5000 case-control datasets, each with either 50, 100, or 300 cases and controls, and determined the type I error rate. In all simulations, the marker allele and disease were independent (i.e., no association). ***Results:*** The type I error rate is not substantially affected by changes in the disease prevalence per se. We found that the CRR provides a relatively poor indicator of the magnitude of the increase in type I error rate. We also derived a simple mathematical quantity, Δ, that is highly correlated with the type I error rate. In the companion article (part II, in this issue) [4], we extend this work to multiple subpopulations and unequal sampling proportions. ***Conclusion:*** Based on these results, realistic combinations of disease prevalences and marker allele frequencies can substantially increase the probability of finding false evidence of marker disease associations. Furthermore, the CRR does not indicate when this will occur.

Supported by NIH grants: DK31775, DK31813, NS27941, MH48858, AA13654, MH65213, MH28274, and MH60970.

KARGER

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

Accessible online at:
www.karger.com/hhe

Gary A. Heiman
Columbia University
722 W. 168th Street, 5th Floor, Room R-502
New York, NY 10032 (USA)
Tel. +1 212 305 6607, E-Mail gah13@columbia.edu

Introduction

Many investigators use the case-control design for genetic association studies to detect an association between a genetic marker and a disease or trait. Compared to the family-based designs, data collection is easier and power greater under case-control designs [5]. However, case-control studies from different laboratories often yield contradictory results [6]. This lack of reproducibility is often attributed to marked variation of disease prevalence and marker allele frequency within subpopulations, so-called PS. PS can lead to false evidence for an association [6–10]. The extent to which PS, the mixing of individuals from heterogeneous backgrounds, biases case-control genetic association studies has been debated in the literature. Some believe PS does not represent a serious concern [1, 2], while others believe that PS may contribute to the contradictory findings and thus represent a serious problem for association studies [11]. In order to try and predict when PS would be a concern, Wacholder et al. used the CRR [1]. These authors reported, for example, that for bladder cancer in European-Americans, the CRR indicated only moderate distortion due to PS. However, the CRR indicates only the magnitude of the distortion in the *relative risk*; it tells us nothing about the probability of finding a false indication of genetic association. The type I error rate, on the other hand, provides a direct measure of how often the null hypothesis (i.e. no association between the genetic marker and disease) is falsely rejected. Other expressions describing the extent of confounding exist [12], but the publication by Wacholder using the CRR has had a strong influence on current literature [13–16]. In this work, we use computer simulations to determine the effect of PS on the type I error rate over a wide range of disease frequencies and marker allele frequencies.

This and the companion article [4] determine the effect of PS on the type I error rate. This first article uses computer simulations over a wide range of disease frequencies and marker allele frequencies for two equal-sized populations. The second article uses a theoretical analysis to extend this work to multiple sub-populations and unequal sampling proportions [4]. This first article has three goals: (1) Determine which combinations of population parameters elevate the type I error rate and in particular, the magnitude of the effect that disease prevalence has on the type I error rate. (2) Determine how the CRR compares to the type I error rate in describing the extent of bias in association studies. (3) Derive a mathematical expression that predicts the magnitude of the type I error rate.

Methods

Combinations of Population Parameters

Using our genome simulator [17], we simulated two populations in equal proportions (POP1 and POP2). The sizes of these populations ranged from 700,000 to 10 million. These populations were simulated with varying disease prevalences and marker allele frequencies (see below). We randomly mixed the two populations to produce a combined population, from which we selected case-control datasets, thus simulating the situation in which a sample is drawn from a mixed population. From this combined population, we randomly sampled 5000 datasets, each with 50, 100, or 300 cases and controls. We determined how many of these samples yielded a χ^2 value ≥ 3.84 (nominal critical value for test significance of 5%). In all simulations, there was no association between the marker and disease in the subpopulations, i.e., the marker allele was unrelated to disease. Therefore, the proportion of datasets yielding $\chi^2 \geq 3.84$ gave the type I error rate or false positive rate.

We examined 160 parameter combinations. Let K_i represent disease prevalence, and p_i, marker allele frequency, in population i. In the first 80 combinations (Set A), POP1 had a 'high' disease prevalence of $K_1 = 0.10$ and a marker allele frequency $p_1 = 0.2$. Note that for marker allele frequency p_i, the carrier frequency (i.e., the frequency of individuals in the population who carry that marker allele) is $p_i + 2\,p_i(1 - p_i)$, under the assumption of Hardy-Weinberg equilibrium. POP2 had 80 different combinations of K_2 and p_2, representing all pairings of ten disease prevalence ratios $K_2/K_1 = 0.1, 0.2, ..., 1.0$ with eight marker allele frequency differences $p_2 - p_1 = 0.0, 0.1, ..., 0.7$. For the second 80 combinations (Set B), POP1 had a 'low' disease prevalence of $K_1 = 0.01$; marker allele frequency was again $p_1 = 0.2$. POP2's parameters were chosen to represent the same disease prevalence ratios and marker allele frequency differences as for Set A. We included the full range of disease prevalence ratios and marker allele frequency differences to illustrate situations in which the type I error rate is inflated. We included the extreme ends (e.g., marker allele frequency differences = 0.7 in different subpopulations) to give some idea of upper bounds on possible type I error rate. Table 1 shows the actual values for the 160 input parameter combinations. For each of these 160 combinations, we selected samples of 50, 100, and 300 cases and controls.

Appendix A gives full details of how we generated and mixed the populations and created case and control samples.

Confounding Risk Ratio

We compared the observed type I error rate with the observed CRR [1], defined as the ratio of the crude odds ratio (OR) to the adjusted OR:

$$CRR \equiv \frac{RR_{un}}{RR_{adj}},$$

where RR_{un} and RR_{adj} are the unadjusted (for the confounder) and adjusted risk ratios, respectively. Denoting the confounder (i.e. subpopulation) by C, and the presence of marker allele by M and its absence by $\overline{M}$, the above expression becomes

$$CRR \equiv \frac{\Sigma_k \Pr\{C = k \mid M\} RR_k}{\Sigma_k \Pr\{C = k \mid \overline{M}\} RR_k},$$

where RR_k is the risk of the disease in subpopulation k given M, divided by the risk of the disease in subpopulation k given $\overline{M}$ [1].

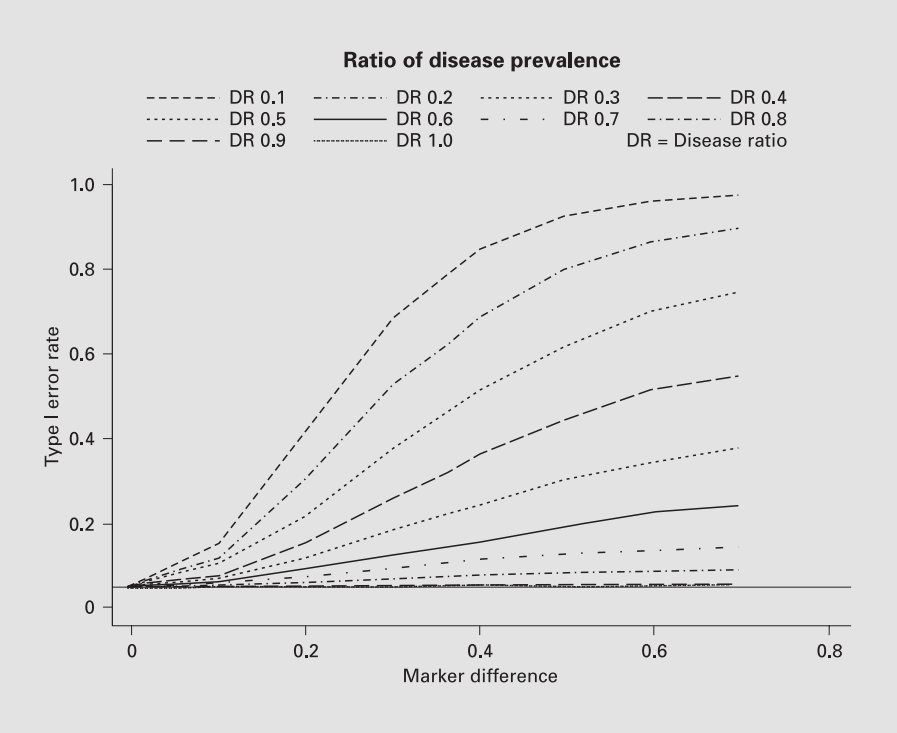

Fig. 1. Plots of type I error rate for each ratio of disease prevalence for Set 1 (population 1 disease prevalence = 10% and marker allele frequency = 0.2) and 100 cases and controls. DR = Disease ratio.

Table 1. Listing of 160 input parameter combinations

Set A Population 1: Disease prevalence K_1 is fixed at 0.1, and marker allele frequency p_i is fixed at 0.2 (carrier frequency = 0.36). Population 2: Parameters include all pairings of the following 10 values of K_2 with the following 8 values of p_2:

10 values of K_2										
Ratio	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
K_2	0.10	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01
8 values of p_2										
Diff.	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7		
p_2	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9		
Carr.	0.36	0.51	0.64	0.75	0.84	0.91	0.96	0.99		

Set B Population 1: Disease prevalence K_1 is fixed at 0.01, and marker allele frequency p_i is fixed at 0.2 (carrier frequency = 0.36). Population 2: Parameters include all pairings of the following 10 values of K_2 with the following 8 values of p_2:

10 values of K_2										
Ratio	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
K_2	0.010	0.009	0.008	0.007	0.006	0.005	0.004	0.003	0.002	0.001
8 values of p_2										
Diff.	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7		
p_2	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9		
Carr.	0.36	0.51	0.64	0.75	0.84	0.91	0.96	0.99		

Ratio = disease prevalence ratio; Diff. = marker allele frequency difference; Carr. = carrier frequency.

For the observed crude OR, we calculated the average OR of all 5000 case-control datasets in each simulation. For the observed adjusted OR, we calculated the average Mantel-Haenszel adjusted OR [18].

Relationship between Marker Allele Difference, Prevalence, and Type I Error Rate

We define a population stratification statistic, Δ, which is a standardized measure of the difference between the proportions of affected (denoted as s) and unaffected (denoted as t) individuals with the marker ($\Pr\{M|D\}$ and $\Pr\{M|\overline{D}\}$, respectively) in the total population:

$$s = \frac{d_1 m_1 + d_2 m_2}{d_1 + d_2},\ t = \frac{d_1 m_1 + d_2 m_2 - m_1 - m_2}{d_1 + d_2 - 2} \quad (1)$$

Appendix B gives the mathematical details. In this formula, d_1 and d_2 are the disease prevalences, and m_1 and m_2, the carrier frequencies in population 1 and population 2, respectively. The populations and the number of cases and controls are assumed to be of equal size (n cases and n controls). Then we define:

$$\Delta = |s - t| \sqrt{\frac{2n}{(s + t)(2 - s - t)}}. \quad (2)$$

When either the disease prevalences or marker allele frequencies are equal in the two populations, Δ becomes 0. As the disease prevalences and marker allele frequencies increasingly differed in the two populations, Δ increases correspondingly. The $s - t$ (expression 2) is well-known in the literature as the case-control effect and has been used to quantify the extent of bias [19]. Our expression is based on the same principles as the expression reported in 1999 [12], but this one is simplified and standardized. In the companion article (part II, in this issue) [4], we extend this formula to any number of subpopulations and unequal sampling proportions.

Results

Combinations of Population Parameters

Table 2A–C shows the type I error rates for the 80 combinations of disease prevalences K and marker allele frequencies p in Set A (POP1 disease prevalence = 10%), for sample sizes 50, 100, and 300 cases and controls. As expected, when either the disease prevalences or the marker allele frequencies were equal in both populations (i.e., left column and bottom row of the tables), the type I error rate was not inflated over the nominal value of 5%. For the other parameter combinations, the type I error rate increased as the parameters increasingly differed between the two populations and could easily exceed 50% or even be considerably higher. Figure 1 shows the type I error rates as a function of marker allele frequency difference and disease prevalence ratio for Set A with sample size of 100 cases and controls.

Table 3A–C shows the same combinations for Set B (POP1 disease prevalence = 1%). The results of these combinations were similar to those in Set A. Thus, the type I error rate is not substantially affected by changes in the disease prevalence between 1 and 10%.

Table 2. Type I error rate (false positive rate) for Set A (population 1 disease prevalence = 10% and marker allele frequency = 0.2)

Disease prevalence ratio	Marker frequency difference							
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
A Sample size = 50 cases and controls								
0.1	0.05	0.10	0.25	0.41	0.58	0.69	0.76	0.80
0.2	0.06	0.09	0.18	0.30	0.42	0.51	0.58	0.63
0.3	0.05	0.08	0.14	0.21	0.30	0.37	0.43	0.46
0.4	0.05	0.07	0.11	0.16	0.22	0.26	0.29	0.32
0.5	0.04	0.06	0.09	0.12	0.15	0.18	0.20	0.21
0.6	0.05	0.06	0.08	0.09	0.11	0.12	0.14	0.14
0.7	0.05	0.06	0.07	0.07	0.08	0.08	0.09	0.09
0.8	0.05	0.06	0.06	0.07	0.07	0.07	0.07	0.07
0.9	0.05	0.06	0.06	0.06	0.06	0.06	0.05	0.05
1.0	0.05	0.05	0.06	0.06	0.06	0.05	0.05	0.05
B Sample size = 100 cases and controls								
0.1	0.05	0.15	0.42	0.68	0.85	0.93	0.96	0.98
0.2	0.05	0.12	0.30	0.53	0.69	0.81	0.87	0.90
0.3	0.05	0.10	0.22	0.38	0.51	0.62	0.70	0.75
0.4	0.05	0.08	0.16	0.26	0.36	0.45	0.51	0.55
0.5	0.05	0.07	0.12	0.18	0.24	0.31	0.35	0.38
0.6	0.05	0.06	0.09	0.12	0.16	0.19	0.23	0.24
0.7	0.05	0.06	0.07	0.09	0.12	0.13	0.13	0.15
0.8	0.05	0.05	0.06	0.07	0.08	0.08	0.08	0.09
0.9	0.05	0.06	0.06	0.05	0.05	0.05	0.05	0.06
1.0	0.05	0.06	0.06	0.06	0.06	0.05	0.05	0.05
C Sample size = 300 cases and controls								
0.1	0.05	0.36	0.85	0.99	1.00	1.00	1.00	1.00
0.2	0.05	0.25	0.68	0.92	0.99	1.00	1.00	1.00
0.3	0.05	0.17	0.50	0.78	0.93	0.98	0.99	1.00
0.4	0.05	0.14	0.36	0.60	0.78	0.88	0.93	0.95
0.5	0.05	0.10	0.23	0.40	0.56	0.69	0.77	0.81
0.6	0.05	0.08	0.16	0.27	0.37	0.46	0.54	0.59
0.7	0.05	0.06	0.10	0.16	0.22	0.26	0.31	0.34
0.8	0.04	0.05	0.07	0.09	0.11	0.13	0.15	0.17
0.9	0.04	0.05	0.06	0.07	0.07	0.07	0.08	0.08
1.0	0.05	0.05	0.05	0.06	0.05	0.05	0.06	0.05

Confounding Risk Ratio

Table 4 shows the inverse of the observed CRR in Set A (POP1 disease prevalence = 10%) for sample size 100 cases and controls. We took the inverse of the observed CRRs for ease of interpretation. These values are similar to those found by Wacholder [1]. While the observed

Table 3. Type I error rate for Set B (population 1 disease prevalence = 1% and marker allele frequency = 0.2)

Disease prevalence ratio	Marker frequency difference							
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
A Sample size = 50 cases and controls								
0.1	0.05	0.10	0.22	0.37	0.51	0.63	0.71	0.76
0.2	0.05	0.09	0.17	0.28	0.39	0.48	0.55	0.58
0.3	0.05	0.08	0.12	0.19	0.26	0.32	0.38	0.40
0.4	0.05	0.07	0.10	0.15	0.19	0.23	0.26	0.29
0.5	0.05	0.07	0.09	0.11	0.14	0.17	0.19	0.20
0.6	0.05	0.06	0.08	0.09	0.10	0.12	0.12	0.13
0.7	0.04	0.05	0.06	0.07	0.07	0.07	0.08	0.08
0.8	0.05	0.06	0.06	0.06	0.07	0.06	0.07	0.06
0.9	0.05	0.06	0.06	0.05	0.05	0.05	0.05	0.05
1.0	0.05	0.06	0.06	0.06	0.05	0.05	0.05	0.05
B Sample size = 100 cases and controls								
0.1	0.05	0.14	0.38	0.64	0.81	0.91	0.94	0.96
0.2	0.05	0.11	0.28	0.46	0.64	0.75	0.82	0.86
0.3	0.05	0.09	0.20	0.33	0.47	0.58	0.66	0.69
0.4	0.05	0.08	0.14	0.22	0.30	0.39	0.45	0.48
0.5	0.05	0.08	0.12	0.17	0.22	0.27	0.32	0.34
0.6	0.05	0.06	0.08	0.11	0.15	0.17	0.19	0.21
0.7	0.05	0.07	0.08	0.08	0.10	0.11	0.12	0.12
0.8	0.05	0.05	0.06	0.06	0.07	0.07	0.07	0.07
0.9	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
1.0	0.05	0.05	0.05	0.06	0.05	0.05	0.05	0.05
C Sample size = 300 cases and controls								
0.1	0.05	0.32	0.80	0.97	1.00	1.00	1.00	1.00
0.2	0.05	0.25	0.66	0.89	0.98	1.00	1.00	1.00
0.3	0.05	0.17	0.47	0.73	0.89	0.96	0.98	0.99
0.4	0.05	0.13	0.32	0.53	0.72	0.82	0.89	0.91
0.5	0.05	0.11	0.23	0.37	0.53	0.65	0.73	0.78
0.6	0.05	0.08	0.14	0.23	0.34	0.43	0.50	0.54
0.7	0.05	0.07	0.09	0.13	0.18	0.21	0.26	0.28
0.8	0.05	0.06	0.07	0.09	0.10	0.11	0.13	0.14
0.9	0.05	0.05	0.05	0.06	0.06	0.06	0.07	0.07
1.0	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05

CRR values were only 'moderately' inflated over 1.00, the corresponding type I error rate can be quite large. For example, while the CRR is 'only' 1.46 when the disease ratio = 0.5 and the marker allele frequency difference = 0.4, the corresponding type I error rate is 24% (see discussion). Thus, for this combination of parameters, 24% of data sets would show statistically significant evidence of a disease-marker association as opposed to 5%. The CRR value of 1.46, while apparently reflecting only a modest increase in relative risk, does not indicate the almost five-fold increase in the probability of falsely finding evidence for association.

Table 4. Inverse of the confounding risk ratio (CRR) for Set A (population 1 disease prevalence = 10% and marker allele frequency = 0.2); Sample size = 100 cases and controls

Disease prevalence ratio	Marker frequency difference							
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
0.1	1.01	1.34	1.68	2.06	2.43	2.78	3.08	3.27
0.2	1.00	1.26	1.53	1.79	2.07	2.31	2.52	2.66
0.3	1.01	1.21	1.41	1.61	1.81	2.00	2.16	2.26
0.4	1.00	1.16	1.31	1.46	1.61	1.75	1.86	1.93
0.5	1.00	1.12	1.24	1.35	1.46	1.56	1.64	1.70
0.6	1.00	1.09	1.18	1.26	1.35	1.43	1.49	1.53
0.7	1.00	1.06	1.12	1.18	1.24	1.29	1.34	1.38
0.8	1.00	1.04	1.08	1.12	1.16	1.20	1.23	1.26
0.9	1.00	1.02	1.04	1.06	1.08	1.11	1.13	1.14
1.0	1.00	1.00	1.01	1.01	1.02	1.03	1.04	1.05

Relationship between Marker Allele Difference, Prevalence, and Type I Error Rate

As noted above, the type I error rate increased as the disease prevalence and/or marker allele frequency increasingly differed in the two populations. For Set A with 100 cases and control, the type I error rate and Δ are highly correlated (correlation = 0.98) and the proportion of variation of type I error rate explained by Δ is also high ($R^2 = 96\%$). Figure 2 illustrates the sigmoidal relationship between type I error rate and Δ. When we take the logit of the type I error rate, the relationship is linear (logit type I error rate = $1.62\Delta - 3.13$) with a concordant index of c = 85.8% [20]. Figure 3 shows this line. Thus, Δ statistic is highly predictive of the type I error rate. Elsewhere, we have derived a general formula to incorporate any number of subpopulations and unequal population sizes [4].

Discussion

The main goals of this study were (1) to show to what extent population stratification can inflate the type I error rate; (2) to show how the CRR compares to the type I error rate in describing the extent of bias in association studies, and (3) to derive a mathematical expression that predicts the magnitude of the type I error rate.

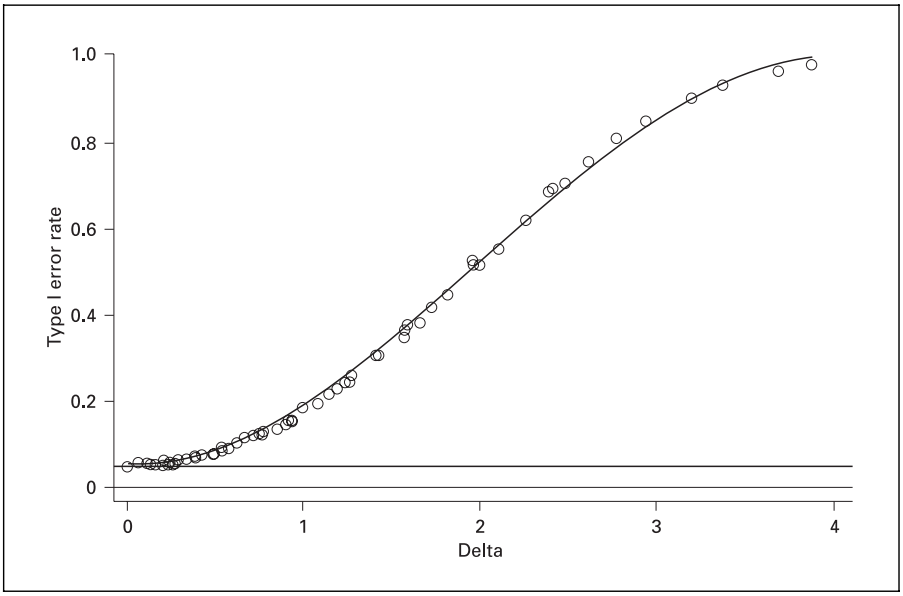

Fig. 2. Δ vs. type I error rate. Scatterplot of type I error rate and Δ and for Set A (population 1 disease prevalence = 1% and marker allele frequency = 0.2) with 100 cases and controls. Polynomial fitted to datapoints.

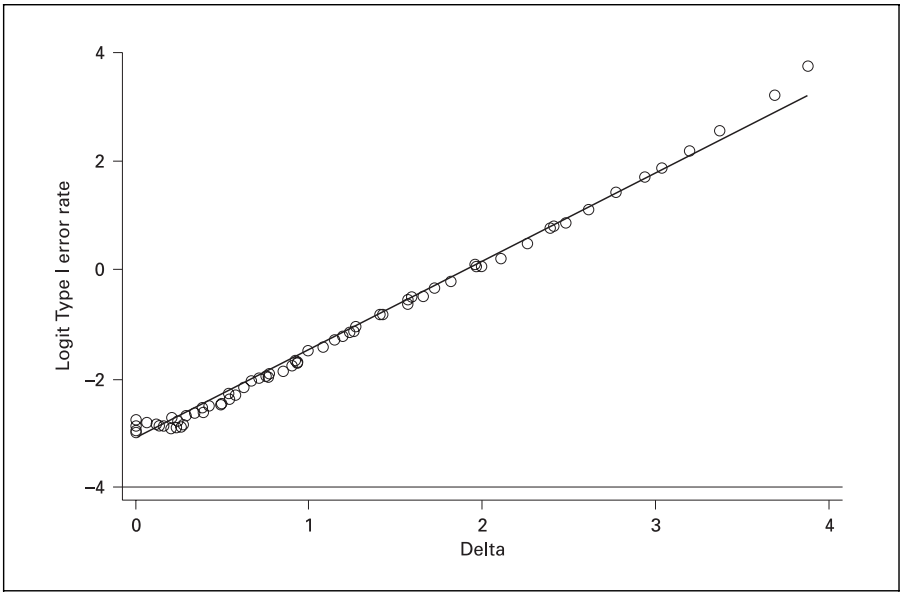

Fig. 3. Δ vs. logit type I error rate. Scatterplot of Δ and logit type I error rate for Set A (population 1 disease prevalence = 1% and marker allele frequency = 0.2) with 100 cases and controls. Line fitted to datapoints.

Inflation of Type 1 Error Rate

We have shown that even with moderate sample sizes (100 cases and 100 controls) and with disease prevalences of 1 and 10%, many combinations of disease prevalence ratios and marker allele frequency differences between the component populations can elevate the type I error rate. The type I error rate is not elevated when the disease prevalence ratio equals 1.0 or when the marker allele frequency difference is 0 in the two populations. As expected, when type I error rate differences exist (i.e., when the prevalence ratio is not 1.0 and the marker allele difference is not 0) the type I error rate also increases with increasing sample sizes (e.g., 300 cases and 300 controls) [21]. Studies that do not account for ethnic background may exhibit artificial association. For example, one study found an association between monoamine oxidase and manic-depressive illness using a case-control design but did not find an association using a family-based design [22]. While newer methods exist to adjust for population stratification [19], many studies do not account for ethnicity, which may have led to inconsistency in the literature [6].

Table 5. Type I error rate for Set C (population 1 disease prevalence = 1% and marker allele frequency = 0.1); Sample size = 100 cases and controls

Disease prevalence ratio	Marker frequency difference							
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
0.1	0.05	0.22	0.59	0.85	0.96	0.99	1.00	1.00
0.2	0.05	0.17	0.43	0.67	0.83	0.92	0.96	0.98
0.3	0.05	0.13	0.30	0.49	0.66	0.79	0.87	0.91
0.4	0.05	0.11	0.21	0.34	0.49	0.60	0.69	0.75
0.5	0.05	0.08	0.15	0.24	0.33	0.42	0.49	0.55
0.6	0.05	0.07	0.10	0.15	0.21	0.27	0.32	0.35
0.7	0.05	0.06	0.08	0.11	0.14	0.17	0.19	0.22
0.8	0.05	0.06	0.07	0.08	0.10	0.11	0.12	0.13
0.9	0.05	0.05	0.05	0.06	0.06	0.07	0.07	0.07
1.0	0.06	0.05	0.05	0.05	0.06	0.06	0.06	0.05

Note that the type I error rate is not substantially affected by changing the disease prevalence in population 1 from 1 to 10%. To further test the effect of prevalence, we increased the disease prevalence to 40% in population 1 with little change in results (data not shown). On the other hand, the marker allele frequency in population 1 does make a difference. We tested 80 parameter combinations (Set C) with $p_1 = 0.1$ (and $K_1 = 0.10$) for a sample size of 100 cases and controls. The lower marker allele frequency notably raised the type I error rate, as shown in table 5 (compare with table 2, part B).

Comparison of FRP with CRR as a Predictor of the Effect of PS on Association Results

The CRRs derived from the parameters used in our simulation were similar in magnitude to those found by Wacholder (2000). However, the type I error rate corresponding to small or moderate increases in CRR were often highly elevated, a fact not easily discerned by merely looking at the magnitude of the CRR. Thus, while an inflation of the observed risk ratio from, for example, 1.0 to 1.4 may not appear severe, it may in fact represent a quite substantial increase in the chance of a finding being a false positive. Therefore, the CRR provides a relatively poor indicator of the magnitude of the increase in type I error rate. The CRR answers a different question: What is the magnitude of the distortion? In addition, since hypothesis testing approaches are used to evaluate whether or not a specific marker is associated with a disease, an elevation of the type I error rate rather than the CRR is the appropriate measure of distortion. Studies of genetic association generally report whether or not a specific marker is statistically different in cases and controls (i.e., hypothesis testing). However, many of these studies are not large enough to reach statistical significance for genes of modest effect [6, 23] and may have contributed to the lack of consistency [6–8]. Instead, attention should be on the effect size and confidence interval [24].

The fundamental difference between the CRR and the type I error rate corresponds to that between estimation and hypothesis testing. The CRR gives an idea of the magnitude of bias due to population stratification, but no indication of how meaningful that magnitude is for the given sample size. The type I error rate indicates 'statistical significance' of the bias but does not indicate its magnitude. Analogously, if we were comparing the means of two normal distributions (e.g., mean height between two groups of people), a mean height difference of, say, 2.5 inches, would not be meaningful by itself, without either a hypothesis test or a confidence interval. Thus, whereas Wacholder says that 'bias is remarkably small even when the range of genotype frequencies is extremely large' [1], we have shown that relatively small CRR values can actually represent highly inflated type I error rate. Moreover, for any given CRR value, the type I error rate becomes more inflated as sample size increases.

We certainly do not want to exaggerate the effects of population stratification on association studies. If anything, we were originally predisposed the other way, i.e., toward believing that other people may have exaggerated the concerns about population stratification. We were surprised by our findings in this work. We now believe the situation is more nuanced than we had originally thought. Wacholder's work has made people question the danger and magnitude of population stratification, which is good. But for some, the pendulum has swung back the other way, i.e., toward, 'it turns out that population stratification need not be a concern at all'. Unfortunately, this is also an oversimplification. This is what we have tried to convey in these papers.

Mathematical Expression

We have derived a formula, Δ, which is a standardized measure of the difference between the proportions of affected and unaffected individuals with the marker in the total population. This statistic is highly correlated with and predictive of the type I error rate.

Conclusions

What combination of disease prevalence and marker allele frequency difference can we expect in an actual study of disease? While PS can elevate the type I error rate, the question remains: Are disease prevalence differences or marker frequency differences in different populations so great that PS is actually a problem? The difference in marker allele frequencies in the two populations simulated ranged from 0 to 0.70, with the lowest marker allele frequency in one population being 0.1. Just how often these values are found in realistic situations is unclear. Variations in allele frequency are reported in the literature [25]. A recent paper found significant evidence for population stratification in real populations [26]. To get an idea of the variation in the frequency of marker alleles in different population, we reviewed the ALelle FREquency Database (ALFRED) [27], which lists the frequency distribution of alleles (including mutations in known genes, CA repeats, and SNPs) at a number of loci in different populations. While many loci have allele frequencies that are similar across populations (e.g. allele frequency for PV92 Alu insertion of the CDH13 locus: Bretons = 0.27; Cypriot Greeks = 0.25; Spaniards = 0.25) other alleles had highly variable frequencies (e.g., TPA25 Alu insertion: Ingush = 0.22; Bretons = 0.32; Finns 0.65; French 0.72). Even within a supposedly geographically homogeneous population, different studies reported population allele frequencies at variance with one another. For example, the allele frequencies for the ACE Alu insertion ranged from 0.27 to 0.48 in three independent samples of Yucatan Mayas and 0.64 to 0.96 in three independent samples of Malaysians. Reported variation of intrapopulation allele frequencies suggests that either sample sizes from which the frequencies were estimated were too small or there may be unexpectedly high variation in what are thought to be homogeneous populations. Whatever the reason, there is the potential for significant differences in allele frequencies in different populations and even, apparently, within the same population. The values reported above are for known genes. How SNPs will vary across populations remains to be determined.

What about realistic variations of disease prevalence? The ratio of disease prevalence generated in these simulations ranged from 1.0 down to 0.1. While some diseases, such as epilepsy [28], appear uniform throughout various populations, there are notable examples of diseases with markedly different prevalence rates in different populations. For example, the worldwide prevalence of many cancers varies substantially [29] as well as type 1 diabetes [30]. Even within the United States, cancer rates vary substantially by state [31]. For example, the US Caucasian male bladder cancer rate ranges from 4.34/100,000 person-years (UT) to 8.58/100,000 person-years (RI) – a disease ratio of 0.51. Moreover, not only may estimates of disease prevalence be based on small sample sizes, but the definition of the disease, the diagnostic criteria for the disease, the precision of the diagnosis, and cultural factors in disease definition may all play a role to make comparison of some diseases among populations most problematic.

Thus, it is difficult to determine what range of differences to expect in allele frequencies of marker alleles among populations, although that may eventually be determined. Determining disease frequencies for the so-called common complex diseases will be more difficult. More importantly, the extent of mixing of different populations in the large cities of the developed world, where many disease studies take place, is difficult to determine.

In this first paper, we limited ourselves to two specifics: First, the findings are based on only two component subpopulations. Most studies may well include individuals from more than two subpopulations. Wacholder (2000) concluded that the PS is expected to be less of a concern as more subpopulations are included (e.g. four to eight ethnic groups). In the second article [4], we use a theoretical analysis to extend the number of sub-populations and relaxed the assumption of equal sampling proportions. In that paper, we show that bias can sometimes be quite substantial even with a very large number of subpopulations. Secondly, we have addressed only the question of mixtures of 'intact' populations. These observations do not address the problem of population admixture, which we define as the offspring of parents from different populations, which may, in fact, be the more important question when assessing disease-marker association than is population stratification. More simulations are needed to determine if similar elevations in type I error rate in admixed populations.

Recent methods, known as genomic control, have been proposed to adjust for population stratification. These methods use a panel of unlinked markers to determine if PS exists and if so, provide a means to adjust. While these methods are intuitively appealing, to date none have been rigorously tested to determine their efficacy. Recent data suggest that genomic control may not adjust appropriately (i.e., over or under-conservative) [21, 32]. Thus, the performance of these methods is currently unknown. Further evaluation of these methods is necessary before they are adopted [33].

Acknowledgments

We thank the members of the Genetics of Complex Disorders Training Program and Drs. Martina Durner and Dvora Shmulewitz for useful discussions on this paper.

Appendix A

We created our simulated samples in the following three steps:

Step 1. Generate the Populations. Beginning with Set A, we generated between 700,000 and 1 million random individuals for POP1 ($K_1 = 0.10$, $p_1 = 0.2$). Then for each combination of K_2 and p_2 we created a POP2, by generating the same number of random individuals as POP1. This yielded 81 populations for Set A (1 POP1 and 80 versions of POP2). We then repeated the procedure for Set B ($K_1 = 0.01$, $p_1 = 0.2$), yielding 81 more populations, except we generated between 3.2 and 10 million individuals in each population.

Step 2. Create 160 Pairs of Case Files and Control Files. Beginning with Set A, for each of the 80 population pairs, we mixed the populations as follows: Generate a random number to choose either POP1 or POP2 with equal probability. From the chosen population, take the 'next' individual (in sequence as those populations were originally generated). Put that individual into either the Case or Control File, depending on whether the individual is affected or unaffected. Repeat the process until one or the other population is depleted. This procedure yielded 80 pairs of Case Files and Control Files for Set A. We then repeated the procedure for Set B, yielding 80 more pairs of Case and Control Files. All Case Files contained at least 40,000 random individuals (many over 100,000), and all Control Files contained at least 600,000 individuals.

Step 3. Create Samples from the Case and Control Files. For each of the 160 pairs of Case and Control Files, we randomly sampled individuals (with replacement) from the mixed population to create samples of size 50, 100, or 300 cases and controls. Every simulation consisted of 5,000 samples.

Appendix B

Derivation of Δ

Notation and Terminology

Let there be two subpopulations of the same size. Define the events: D = person in population has disease; M = genotype in population has marker; C_i = person is selected from subpopulation i ($i = 1, 2$).

We define the following probabilities: $d_i = \Pr\{D|C_i\}$; $m_i = \Pr\{M|C_i\}$; $\Pr\{C_1\} = \Pr\{C_2\} = 1/2$. By assumption, M and D are not associated within any subpopulation, hence

$$\Pr\{\mathrm{M} \cap \mathrm{D}|\mathrm{C_i}\} = m_i d_i$$

Derivation

The population proportion of affected individuals in the population with the marker is:

$$s = \Pr\{M|D\} =$$

$$\frac{\Pr\{\mathrm{M} \cap \mathrm{D}|\mathrm{C_1}\}\Pr\{\mathrm{C_1}\} + \Pr\{\mathrm{M} \cap \mathrm{D}|\mathrm{C_2}\}\Pr\{\mathrm{C_2}\}}{\Pr\{\mathrm{D}|\mathrm{C_1}\}\Pr\{\mathrm{C_1}\} + \Pr\{\mathrm{D}|\mathrm{C_2}\}\Pr\{\mathrm{C_2}\}} \text{ (by Bayes' Rule)}$$

$$= \frac{\frac{m_1 d_1}{2} + \frac{m_2 d_2}{2}}{\frac{d_1}{2} + \frac{d_2}{2}} = \frac{m_1 d_1 + m_2 d_2}{d_1 + d_2}.$$

Similarly, the proportion of unaffected individuals in the population with the marker is

$$t = \Pr\{M|\overline{D}\} = \frac{m_1(1 - d_1) + m_2(1 - d_2)}{2 - (d_1 + d_2)}.$$

Let $\hat{s}$ and $\hat{t}$ represent the observed proportions of cases and controls, respectively, that have the marker allele (cf., e.g., (1) in text). We define the $\hat{\Delta}$ statistic as the standardized difference between these two proportions in the sample:

$$\hat{\Delta} = \left| \frac{\hat{s} - \hat{t}}{SE(\hat{s} - \hat{t})} \right| = \frac{|\hat{s} - \hat{t}|}{\sqrt{\left(\frac{\hat{s} + \hat{t}}{2}\right)\left(1 - \frac{\hat{s} + \hat{t}}{2}\right)\left(\frac{2}{n}\right)}}$$

$$= |\hat{s} - \hat{t}| \sqrt{\frac{2n}{(\hat{s} + \hat{t})(2 - \hat{s} - \hat{t})}}.$$

For n large, the sample statistics $\hat{s}$, $\hat{t}$ and $\hat{\Delta}$ approach the corresponding population values s, t and Δ respectively, so that the magnitude of the standardized difference between the proportions of affected and unaffected individuals with marker in the total population is approximately:

$$\Delta = |s - t| \sqrt{\frac{2n}{(s + t)(2 - s - t)}},$$

which is equation (2) in the text.

References

1 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(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 Gorroochurn P, Hodge SE, Heiman G, Greenberg DA: Effect of population stratification on case-control association studies. II. False-positive rates and their limiting behavior as number of subpopulations increases. Hum Hered 2004;58:40–48.

5 Risch N, Teng J: The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. Genome Res 1998;8(12): 1273–1288.

6 Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K: A comprehensive review of genetic association studies. Genet Med 2002; 4(2):45–61.

7 Ioannidis JP, Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG: Genetic associations in large versus small studies: An empirical assessment. Lancet 2003;361(9357):567–571.

8 Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG: Replication validity of genetic association studies. Nat Genet 2001; 29(3):306–309.

9 Tabor HK, Risch NJ, Myers RM: Opinion: Candidate-gene approaches for studying complex genetic traits: Practical considerations. Nat Rev Genet 2002;3(5):391–397.

10 Cardon LR, Palmer LJ: Population stratification and spurious allelic association. Lancet 2003;361(9357):598–604.

11 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.

12 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.

13 Redden DT, Allison DB: Nonreplication in genetic association studies of obesity and diabetes research. J Nutr 2003;133(11):3323–3326.

14 Worrall BB, Azhar S, Nyquist PA, Ackerman RH, Hamm TL, DeGraba TJ: Interleukin-1 receptor antagonist gene polymorphisms in carotid atherosclerosis. Stroke 2003;34(3):790–793.

15 Knight JA, Onay UV, Wells S, Li H, Shi EJ, Andrulis IL, et al: Genetic variants of GPX1 and SOD2 and breast cancer risk at the Ontario site of the Breast Cancer Family Registry. Cancer Epidemiol Biomarkers Prev 2004;13(1): 146–149.

16 Schaid DJ: Disease-Marker Association; in Elston R, Olson JM, Palmer L (eds): Biostatistical Genetics and Genetic Epidemiology. Indianapolis, John Wiley & Sons, 2002, pp 206–217.

17 Greenberg DA, MacCluer JW, Spence MA, Falk CT, Hodge SE: Simulated data for a complex genetic trait (problem 2 for GAW11): How the model was developed, and why. Genet Epidemiol 1999;17(suppl 1):S449–S459.

18 Mantel N, Haenszel W: Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959;22:719–748.

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

20 Armitage P, Berry G, Matthews JNS: Statistical Methods in Medical Research, ed 4. Oxford, Blackwell Science LTD, 2001.

21 Marchini J, Cardon LR, Phillips MS, Donnelly P: The effects of human population structure on large genetic association studies. Nat Genet 2004.

22 Parsian A, Todd RD: Genetic association between monoamine oxidase and manic-depressive illness: Comparison of relative risk and haplotype relative risk data. Am J Med Genet 1997;74(5):475–479.

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

24 Cohen J: The earth is round ($p < 0.05$). Am Psychologist 1994;49(12):997–1003.

25 Bowcock AM, Kidd JR, Mountain JL, Hebert JM, Carotenuto L, Kidd KK, et al: Drift, admixture, and selection in human evolution: A study with DNA polymorphisms. Proc Natl Acad Sci USA 1991;88(3):839–843.

26 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.

27 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.

28 Kotsopoulos IA, van Merode T, Kessels FG, de Krom MC, Knottnerus JA: Systematic review and meta-analysis of incidence studies of epilepsy and unprovoked seizures. Epilepsia 2002; 43(11):1402–1409.

29 Parkin DM, Pisani P, Ferlay J: Global cancer statistics. CA Cancer J Clin 1999;49(1):33–64.

30 Silink M: Childhood diabetes: A global perspective. Horm Res 2002;57(suppl 1):1–5.

31 Devesa SS, Grauman DJ, Blot WJ, Pennello G, Hoover RN, Fraumeni JFJ: Atlas of cancer mortality in the United States, 1950–94. 1999. Washington, DC, US Govt Print Off [NIH Publ No. (NIH) 99–4564].

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

33 Spence MA, Greenberg DA, Hodge SE, Vieland VJ: The emperor's new methods. Am J Hum Genet 2003;72(5):1084–1087.