Evaluating Bias due to Population Stratification in Epidemiologic Studies of Gene-Gene or Gene-Environment Interactions

Yiting Wang,1 Russell Localio,1 and Timothy R. Rebbeck1,2
1Department of Biostatistics and Epidemiology, Center for Clinical Epidemiology and Biostatistics and 2Abramson Cancer Center, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania

Abstract

Confounding by ethnicity (i.e., population stratification) can result in bias and incorrect inferences in genotype-disease association studies, but the effect of population stratification in gene-gene or gene-environment interaction studies has not been addressed. We used logistic regression models to fit multiplicative interactions between two dichotomous variables that represented genetic and/or environmental factors for a binary disease outcome in a hypothetical cohort of multiple ethnicities. Biases in main effects and interactions due to population stratification were evaluated by comparing regression coefficients in misspecified models that ignored ethnicities with their counterparts in models that accounted for ethnicities. We showed that biases in main effects and interactions were constrained by the differences in disease risks across the ethnicities. Therefore, large biases due to population stratification are not possible when baseline disease risk differences among ethnicities are small or moderate. Numerical examples of biases in genotype-genotype and/or genotype-environment interactions suggested that biases due to population stratification for main effects were generally small but could become large for studies of interactions, particularly when strong linkage disequilibrium between genes or large correlations between genetic and environmental factors existed. However, when linkage disequilibrium among genes or correlations among genes and environments were small, biases to main effects or interaction odds ratios were small to nonexistent. (Cancer Epidemiol Biomarkers Prev 2006;15(1):124–32)

Introduction

When the risk of disease varies between ethnic groups and the statistical distribution of an exposure variable (whether genetic or environmental) also varies between these groups, the association between the exposure and disease may be confounded by the ethnic background of the studied groups. This form of confounding by ethnicity is called population stratification in epidemiologic studies of genetic risk factors (1). The effect of population stratification in epidemiologic studies of disease association with a single candidate gene has been extensively studied (2-9). Although various approaches using unlinked markers have been proposed to deal with the effects of population stratification on association tests of candidate genes (10-15), epidemiologic studies are increasingly concerned with interactions between genetic and/or environmental factors. None of the prior literature explicitly addresses the pattern or extent of bias due to population stratification in association studies involving interactions between genes and environmental factors.

To evaluate bias due to population stratification, we employed two dichotomous variables to represent the genetic and/or environmental factors and used logistic regression models to fit multiplicative interactions between the two variables for a binary disease outcome in a hypothetical cohort of multiple ethnicities. We derived algebraic solutions for asymptotic biases in the maximum likelihood estimates of the interaction variables under corresponding models that ignored ethnicities and identified conditions for the biases to reach their maximum or minimum expected values. We also provided numerical examples of biases due to population stratification under a wide range of conditions that may be observed in epidemiologic studies.

Materials and Methods

Data Structure and Model Specification. Assume that the interactions between two categorical variables $V_1$ and $V_2$ are studied for a binary disease outcome $Y$ during a certain time in a closed cohort of subjects. Either $V_1$ or $V_2$ or both can be genetic or environmental variables (i.e., either gene-gene or gene-environment interactions can be modeled using the following approach). $V_1 = 0$ denotes the reference category of $V_1$, and $V_1 = 1$ denotes the comparison category. $V_2$ is coded in the same fashion as $V_1$, $Y = 1$ indicates the presence of disease, and $Y = 0$ indicates the absence of disease. Assume the cohort is comprised of $k$ ethnicities represented by indicator variables $E_j$, $j = 1, 2, \ldots, k$, so that a subject in the $i$th ethnicity will be coded by $E_i E_{j1} E_{j2} \cdots E_{jk}$, where $E_{ji} = 0, 1$, and $E_{ji} = 0$ for $i \neq j = 1, 2, \ldots, k$. Assume associations between disease and $V_1$ and $V_2$ are modeled with logistic regression as

$$
\logit(\pi) = x_1 + x_2 \times E_2 + \ldots + x_k \times E_k + \beta_1 \times V_1 + \beta_2 \times V_2 + \beta_3 \times V_1 \times V_2
$$

where $\pi$ is the conditional probability of the disease ($Y = 1$) given $V_1$, $V_2$, and ethnicity, and binomial random error is assumed. Without loss of generality, let $x_1$ specifies the log odds of disease (i.e., logit function of the baseline disease risk) in the lowest-risk ethnicity $E_1$, and $0 < x_2 < \ldots < x_k$, where $x_k$ specifies the log odds ratio (OR) of disease risks comparing ethnicity $E_k$ versus $E_1$. Similarly, $\beta_1$ specifies the log OR of disease risks comparing ethnicity $E_1$ versus $E_1$; $\beta_2$
and $\beta_2$ specify the main effects associated with the comparison of $V_1$ and $V_2$ relative to their reference categories, respectively; and $\beta_3$ specifies the multiplicative interaction between $V_1$ and $V_2$.

Population stratification may be present when the joint distributions of $V_1$ and $V_2$ are different across the ethnicities. Biases due to population stratification can be evaluated by omitting all ethnicity indicator variables from model Eq. A to fit the mis-specified model

$$\logit(\pi) = x^T + \beta_1^* \times V_1 + \beta_2^* \times V_2 + \beta_3^* \times V_1 \times V_2$$ (B)

We define the asymptotic biases due to population stratification to be $\beta_i^* - \beta_i$, $\beta_2^* - \beta_2$ for the main effects of $V_1$ $V_1$ and $V_2$, respectively, and define $\beta_3^* - \beta_3$ as the asymptotic bias for the interaction effects between $V_1$ and $V_2$. Here, we do not deal with issues of variance or precision of estimates, assuming model coefficient estimates obtained from sufficiently large samples satisfy $E(\beta_i^*) = \beta_i^*$, where $i = 1, 2, 3$.

**Two Ethnicities.** We obtained numerical estimates of large sample biases due to population stratification by fitting logistic regression models to simulated data generated under a wide range of conditions using Stata 7.0 (College Station, TX). Baseline disease risk in the low-risk ethnicity was specified to be 1%, whereas baseline risk in the high-risk ethnicity specified to be 2% or 10%. To specifically assess genotype-genotype interactions, subsequent text assumes $V_1 = G_1$ and $V_2 = G_2$ to denote the study of interactions involving two candidate genes $G_1$ and $G_2$. For simplicity, we assumed that candidate gene $G_i$ ($i = 1, 2$) had two alleles $A_i$ and $a_i$ with frequency $p_i$ and $1 - p_i$, respectively. Genotype $a_1a_1$ was coded as the reference category (i.e., $G_i = 0$), whereas genotypes $A_1A_1$ and $A_1a_2$ were coded as the comparison (i.e., “at-risk” genotype) category (i.e., $G_i = 1$). The at-risk genotype frequencies of $G_i$ were specified in the range of 5% to 95%, assuming single-locus Hardy-Weinberg equilibrium, corresponding to $p_i$ ranging from 3% to 78%.

In some cases, $G_1$ and $G_2$ may not be independently distributed due to linkage disequilibrium between the two loci, denoted here as $D$. Because $D$ between $G_1$ and $G_2$ is constrained by $D_{max}$ and $D_{min}$, where $D_{max} = \min\{p_1 \times (1 - p_2), (1 - p_1) \times p_2\}$ and $D_{min} = \max\{-p_1 \times p_2, - (1 - p_1) \times (1 - p_2)\}$ (16), we specified the degree of linkage disequilibrium between the two genes by $D = D' \times D_{max}$ or $D' \times D_{min}$, where $D'$ took on the values of 0 (no linkage disequilibrium), 0.2 (small linkage disequilibrium), 0.5 (moderate link disequilibrium), and 0.9 (strong linkage disequilibrium). Main effects and interactions of genes were specified by assigning $\beta_1 = 0$ (no effect, OR = 1), $\beta_i = \pm 0.4$ (small effect, OR = 1.49 or 0.67), $\beta_i = \pm 0.8$ (moderate effect, OR = 2.23 or 0.45), or $\beta_i = \pm 1.6$ (large effect, OR = 4.95 or 0.20). Using the expected frequency of each genotype-disease category under the specified conditions, a hypothetical cohort of 100,000 observations was generated. Disease status for each observation was determined by comparing its disease risk with the standard uniform random variable. 5000 case-control samples were randomly drawn from the hypothetical cohort. Each sample consisted of 95% of diseased individuals as cases and an equal number of nondiseased individuals as controls. Both correctly specified models and their corresponding

![Figure 1](image_url)
mis-specified models were fitted to each sample to obtain point estimates of all relevant regression coefficients. The corresponding point estimates from the 5,000 samples were averaged to obtain large sample point estimates for biases due to population stratification on both main effects and interactions. Results are presented in Figs. 1, 2, and 3.

More than Two Ethnicities. For more than two ethnicities ($k > 2$), we assumed baseline risks of $k$ ethnicities were specified to fall within certain ranges. Specifically, we assumed that baseline disease risk of the $i$th ethnicity $p_i \sim \text{Uniform}[0.01, 0.02]$ or $p_i \sim \text{Uniform}[0.01, 0.1]$ for $i = 1, \ldots, k$. Similarly, we assumed that genotype frequencies were uniformly distributed and consistent with single-locus Hardy-Weinberg Equilibrium. We considered the at-risk genotype frequency within ranges of 5% to 10%, 5% to 20%, ..., up to 5% to 95%. Linkage disequilibrium ranged from 90% of minimum possible value (i.e., $D_{\text{min}}$) to 90% of maximum possible value (i.e., $D_{\text{max}}$). Under these assumptions, we generated 5,000 sets of variables for a hypothetical cohort. Each set of disease risks and genotype frequencies of the $k$ ethnicities was randomly assigned from the distributions specified above, assuming $k = 2, 5, \text{ or } 10$, respectively, and $\beta_1 = \beta_2 = \beta_3 = 0$ (i.e., OR = 1) or $\beta_1 = \beta_2 = \beta_3 = 0.693$ (i.e., OR = 2). Next, case-control samples were randomly drawn from the hypothetical cohort under each set of variables to obtain bias estimates as described in the previous paragraph. The range and average of the 5,000 sets of bias estimates were presented in Fig. 4.

Results

Determination of Maximal Possible Bias. We have determined the maximal biases that could result from population stratification (see Appendix 1 for derivations). In the simplest case when the cohort is comprised of two ethnicities (i.e., $k = 2$: one “higher risk” ethnicity and one “lower risk” ethnicity), the maximal bias that can be attained under conditions of population stratification is determined by the difference in log odds of disease risks in the two populations (i.e., the log of OR of disease risks in the two ethnicities) being compared. Specifically, the asymptotic bias to the main effect estimates for $V_1$ or $V_2$ are bounded by $\frac{2}{\sqrt{C_0} a^2}$ and $a^2$, where $a$ is the difference in log odds of the baseline risk of disease in the higher risk population compared with the lower risk population (see Eqs. A and B and Appendix 1).

Figure 2. Biases in main effects and interactions due to population stratification (conditions c and d). Baseline disease risks in low-risk versus high-risk ethnicities were 1% versus 10%. The vertical axis represented biases in main effects or interaction estimates when ethnicities were ignored. The horizontal axis showed at-risk genotype frequencies and degrees of linkage disequilibrium $D'$. The frequency of the at-risk genotype of $G_1$ in the high-risk ethnicity equaled to the frequency of the at-risk genotype of $G_1$ in the low-risk ethnicity, going up from 5% in both ethnicities to 95% in both ethnicities. Linkage disequilibrium between the two genes was $D' \times D_{\text{max}}$ in high-risk ethnicity and $D' \times D_{\text{max}}$ in low-risk ethnicity. A. Frequency of at-risk genotype of $G_2$ varied in the high-risk ethnicity from 20% up to 80%, with a step of 20%, shown on top of each subpanel; whereas it remained fixed at 10% in the low-risk ethnicity. B. Frequency of at-risk genotype of $G_2$ varied in the low-risk ethnicity from 20% up to 80%, with a step of 20%; whereas it remained fixed at 10% in the high-risk ethnicity.
joint occurrence of $V_1 = 0$ and $V_2 = 0$ is never observed in the high-risk ethnicity.

Asymptotic biases to estimates of interactions between $V_1$ and $V_2$ are bounded by $-2\alpha_2$ and $2\alpha_2$. That is, the maximal bias due to population stratification that can be attained for an interaction between two factors is bounded by twice the log OR of the disease risks in the two populations being compared. However, these maximal biases can be reached only when all of the following four conditions hold: (a) the joint occurrence of $V_1 = 1$ and $V_2 = 0$ is never observed in the high-risk ethnicity; (b) the joint occurrence of $V_1 = 0$ and $V_2 = 1$ is never observed in the high-risk ethnicity; (c) the joint occurrence of $V_1 = 1$ and $V_2 = 1$ is never observed in the low-risk ethnicity; (d) the joint occurrence of $V_1 = 0$ and $V_2 = 0$ is never observed in the low-risk ethnicity. Similarly, the maximal bias in the other direction is $-2\alpha_k$ only when the reverse of the four conditions described above hold (see Appendix 1).

Similarly, when $k > 2$ ethnicities in the cohort, the most extreme biases that could result from population stratification are $\pm \alpha_k$ and $\pm 2\alpha_k$ to main effects and interaction estimates, respectively, where $\alpha_k$ represents the maximum of log OR among baseline risks of the $k$ ethnicities (see Eqs. A and B and Appendix 1). To our knowledge, this is the first demonstration of the bounds on the magnitude of biases to interaction estimates for genotype-genotype or genotype-environment interaction studies. However, the boundary conditions that result in the theoretical extremes are unlikely to represent the conditions observed in most studies. Therefore, we undertook numerical evaluations next to consider situations that are more likely to be encountered in actual studies.

---

**Figure 3.** Biases in main effects and interactions due to population stratification (conditions e and f). The vertical axis represented biases in main effects or interaction coefficients when ethnicities were ignored. The horizontal axis showed at-risk genotype frequencies of $G_1$, linkage disequilibrium between $G_1$ and $G_2$ were $D' \times D_{\text{max}}$ in high-risk ethnicity and $D' \times D_{\text{max}}$ in low-risk ethnicity, where $D' = 0.2, 0.5$, or $0.9$ and were shown right above the horizontal axes in three subpanels separated by solid vertical lines. Baseline disease risks in high-risk versus low-risk ethnicities were 10% versus 1% and the fraction of the high-risk ethnicity in the cohort was 0.5. A. The frequency of the at-risk genotype of $G_1$ remained 5% in low-risk ethnicity but varied from 5% up to 95% in the high-risk ethnicity. Frequency of at-risk genotype of $G_2$ remained 10% in low-risk ethnicity but varied in the high-risk ethnicity from 20% to 80%, labeled 20%, 40%, 60%, 80%, respectively. B. The frequency of the at-risk genotype of $G_1$ remained 5% in low-risk ethnicity but varied from 5% up to 95% in the low-risk ethnicity. Frequency of at-risk genotype of $G_2$ remained 10% in low-risk ethnicity but varied in the high-risk ethnicity from 20% to 80%, respectively. C. The frequency of the at-risk genotype of $G_1$ remained 5% in low-risk ethnicity but varied from 5% up to 95% in the high-risk ethnicity. Frequency of at-risk genotype of $G_2$ remained 10% in high-risk ethnicity but varied in the low-risk ethnicity from 20% to 80%, respectively. D. The frequency of the at-risk genotype of $G_1$ remained 5% in high-risk ethnicity but varied from 5% up to 95% in the low-risk ethnicity. Frequency of at-risk genotype of $G_2$ remained 10% in high-risk ethnicity but varied in the low-risk ethnicity from 20% to 80%, respectively.
Two Ethnicities. We have summarized results of large sample biases to main effects and genotype-genotype or genotype-environment interactions using six sets of conditions that may be encountered in association studies (Table 1). Note that bias can only arise when differences in baseline disease risk among ethnicities exist (1), which is a necessary condition for confounding to occur (17). Condition a represented the situation where no population stratification.

Table 1. Summary of asymptotic biases in genotype-genotype interactions under various population stratification conditions for two ethnicities

<table>
<thead>
<tr>
<th>Comparison of high-risk and low-risk ethnicities</th>
<th>Bias in estimation</th>
<th>Figure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline disease risks</td>
<td>Marginal genotype distributions for gene 1 ($G_1$)</td>
<td>Marginal genotype distributions for gene 2 ($G_2$)</td>
</tr>
<tr>
<td>≠</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>≠</td>
<td>≠</td>
<td>=</td>
</tr>
<tr>
<td>≠</td>
<td>≠</td>
<td>≠</td>
</tr>
<tr>
<td>≠</td>
<td>≠</td>
<td>≠</td>
</tr>
<tr>
<td>≠</td>
<td>≠</td>
<td>=</td>
</tr>
<tr>
<td>≠</td>
<td>≠</td>
<td>≠</td>
</tr>
</tbody>
</table>

NOTE: ≠, variables/distributions were different across the high-risk and low-risk ethnicities. =, variables/distributions were same across the high-risk and low-risk ethnicities. ±, bias was generally negligible, towards the null when the effect was nonzero; bias was absent when the effect was zero. +/−, bias was generally negligible; bias was absent when both main effects and interactions were all zero. +, bias was present. −, no bias.

*No figure was presented because (by symmetry of $G_1$ and $G_2$) similar results were presented in Fig. 2.

No figure was presented because bias (if any) was not due to population stratification and was generally negligible.

No figure was presented because no bias was observed.
existing for \( G_1 \) or \( G_2 \), when \( G_1 \) and \( G_2 \) had the same marginal distributions, and linkage disequilibrium between \( G_1 \) and \( G_2 \) was the same in each ethnicity (i.e., joint distributions of \( G_1 \) and \( G_2 \) were the same across the ethnicities). Under these conditions, ignoring ethnicity did not result in large sample bias due to population stratification in any of these estimates when \( \beta_1 = \beta_2 = \beta_3 = 0 \). When \( \beta_1 = \beta_2 = 0 \) but \( \beta_3 \neq 0 \), ignoring ethnicity did not result in large sample bias to \( \beta_1 \) or \( \beta_2 \) but resulted in negligible biases toward the null hypothesis to the interaction term \( \beta_3 \). When \( \beta_1 \neq 0 \) or \( \beta_2 \neq 0 \), a slight bias towards the null hypothesis was observed, whereas biases to \( \beta_3 \) were no longer always towards the null. Instead, the bias to \( \beta_3 \) depended on both the main and interaction effects between the two genes. In all cases, the estimation of these large sample biases was negligible and reflected nonlinearity of logistic regression model (18-20) rather than biases due to population stratification.

Condition b (Fig. 1) represented the situation where the marginal genotype distributions of both genes were same across both ethnicities, such that the frequency of the at-risk genotype of \( G_1 \) in the high-risk ethnicity was equal to the frequency of the at-risk genotype of \( G_1 \) in the low-risk ethnicity, and the frequency of the at-risk genotype of \( G_2 \) in the high-risk ethnicity was equal to the frequency of the at-risk genotype of \( G_2 \) in the low-risk ethnicity. However, condition b also specified that the joint genotype distributions of \( G_1 \) and \( G_2 \) were different due to unequal linkage disequilibrium between the two genes across ethnicities. In these circumstances, ignoring ethnicity could result in large sample biases in both main and interaction estimates. Linkage disequilibrium between \( G_1 \) and \( G_2 \) was \( D^* \times D_{\text{min}} \) in the high-risk ethnicity and \( D^* \times D_{\text{min}} \) in the low-risk ethnicity (Fig. 1). Thus, at-risk genotypes of both genes appeared together more frequently in the high-risk ethnicity than in the low-risk ethnicity. For simplicity, it was assumed that baseline disease risks and the joint genotype distributions of \( G_1 \) and \( G_2 \) were different due to unequal linkage disequilibrium between the two genes. Figure 1 suggests that there was no simple pattern of bias corresponding to the variables considered here. However, large biases in interaction estimates tended to occur when large biases in main effects also occurred.

Under condition c, the marginal genotype distribution of \( G_1 \) was constant across ethnicities, and \( G_1 \) and \( G_2 \) were in linkage equilibrium in both ethnicities (\( D = 0 \)). When \( \beta_1 = 0 \) and \( \beta_3 = 0 \), there were no large sample biases to their estimates. As shown in Fig. 2A and B for “No Linkage Disequilibrium in Either Ethnicity,” biases in \( G_2 \) main effect estimates followed the same patterns as biases to a single candidate gene under conditions of population stratification (1, 6). Bias was positive if at-risk genotype frequency of \( G_2 \) was greater in high-risk ethnicity (Fig. 2A). Bias was negative if at-risk genotype frequency of \( G_2 \) was greater in low-risk ethnicity (Fig. 2B). Condition d differed from condition c only in that linkage disequilibrium was specified to be different across ethnicities. Under condition d, biases to main effects and interactions depended on the ethnicity-specific genotype frequencies of both genes and baseline disease risks as well as the main effects of the genes and their interactions. For example, we considered the situation when baseline disease risks by ethnicity were 10% versus 1%, and both main effects were similar to those reported for condition c, but biases to interaction effect estimates did not follow simple patterns corresponding to the marginal genotype frequencies of either gene. If \( G_1 \) and \( G_2 \) were in linkage disequilibrium (condition f), biases to main effects no longer followed simple patterns. For example, in Fig. 3A, baseline disease risks by ethnicity were 10% versus 1%, and both main effects and interaction estimates were assumed to be 0. Bias depended on at-risk genotype frequencies of both genes as well as on the degree of linkage disequilibrium between the two genes. Large biases were observed even when genotype frequencies were not very different across ethnicities. Again, no simple relationship of bias was observed with respect to genotype frequency. Therefore, biases due to stratification can be large in relatively unpredictable ways when marginal genotype frequencies of both genes differ by ethnicity and the two genes are in linkage disequilibrium.

More than Two Ethnicities. When we expanded our analyses to consider cohorts consisting of \( k = 5 \) or 10 ethnicities, we observed that (on average) larger sample biases to main and interaction effects were either nonexistent or negligible (Fig. 4A-D), even if the conditions for population stratification (Table 1) were met. This result follows because we assumed that baseline disease risks and the joint genotype distributions of both genes were uncorrelated. Biases to both main effects and interaction were greatest when \( k = 2 \) (i.e., when there were only two ethnicities in the cohort) but decreased with increasing number of component ethnicities. Biases to both main effects and interaction were smaller when baseline disease risks of the \( k \) ethnicities were all within the range of 1% to 2% and larger when baseline risks of the \( k \) ethnicities were all within the range of 1% to 10%. Biases to main effects tended to increase as the differences in genotype frequencies across ethnicities increased. For example, in \( G_1 \) main effects and \( G_2 \) main effects in Fig. 4A-D, the range of biases for main effects increased as the range of at-risk genotypes of both genes across the \( k \) ethnicities increased from 5% to 10% up to 5% to 95%. Under the latter more extreme conditions, biases to main effects approached their theoretical bounds.

For example, based on our algebraic derivation of bounds to the biases of the estimates, when baseline risks of \( k \) ethnicities were within 1% to 2% (Fig. 4A and B), biases to main effects approached their bounds of \(-0.7 \) and 0.7 (on the natural log scale). Although interaction estimates were bounded by \(-1.4 \) and 1.4, the actual biases observed were far from reaching these bounds. On the other hand, biases to interaction did not show monotonic relationships corresponding to increasing or decreasing ranges of marginal genotype frequency ranges for either gene. Even when the range of marginal genotype frequencies of both genes was small across ethnicities, biases to interaction estimates could still be very large. The patterns were similar when baseline risks of \( k \) ethnicities were within 1% to 10% (Fig. 4C and D), where biases to main effects were bounded
Population Stratification for Interaction

by −2.4 to 2.4 and biases to interaction by −4.8 to 4.8. Similar patterns were observed when OR = 1 (Fig. 4A and C) and when OR = 2 (Fig. 4B and D) for both main effects and interactions.

Discussion

We have evaluated the potential for bias due to population stratification in studies of genotype-genotype or genotype-environment interactions by deriving bounds for the maximal biases that may be conferred by population stratification. We also provided numerical examples for the potential biases in main effects and interactions under a variety of conditions commonly met in association studies. Bias can occur in gene-gene interaction studies if linkage disequilibrium differs across ethnicities and the sample consists of the mixed ethnicities, even when the conditions of population stratification for either gene alone are absent. When the two genes were in linkage equilibrium, both main effects and interactions were unbiased if marginal distributions of the genes did not vary by ethnicity. When the two genes were in linkage disequilibrium, biases could be substantial for all effect estimates. Genes may seem to be in linkage equilibrium in the entire cohort, whereas within each ethnicity, they are not. Thus, when multiple genes are jointly studied across ethnicities, pooling ethnicities may induce biases that depend on the joint distributions or correlations between the genes across ethnicities, as well as their marginal distributions.

These arguments can be extended to studies of genotype-environment interactions, where correlation between the gene and environment factors exists. However, it is more likely that differences across ethnicities in the joint distributions of genetic and environmental factors result from different marginal distributions rather than from correlations between the genes and environments of interest. Although severe biases were only observed under extreme stratification conditions (e.g., large differences in linkage disequilibrium, in the marginal frequencies of two genes, and in disease risks across ethnicities), population stratification may result in larger biases in genotype-genotype interaction studies than in studies involving only one gene. Our analytic derivation showed that in most settings of interaction where the joint distributions of two factors differ across the levels of a third factor, and where disease risk also varies across those levels, the omission of that third factor as a covariate will produce biases in the estimates of interaction estimates that can be 2-fold as large as biases to the estimates of main effects. To our knowledge, this is the first time such quantitative evaluation has been addressed. We anticipate that our findings of population stratification effect on interaction studies will be useful for studies of both gene-gene interactions and gene-environment interactions (21, 22) in different populations.

Additional studies are required to address other aspects of bias due to population stratification in genotype-genotype interaction studies. For example, we have not considered the effect of deviations from Hardy-Weinberg equilibrium. Such deviations could also confer biases to interaction terms involving two or more genes. Nonetheless, the magnitude of biases would be bounded as shown in Appendix 1. Similarly, we have focused on large sample biases to point estimates from logistic regression in genotype-genotype interaction studies. Another future challenge is to assess the effect of population stratification on variance estimation and hypothesis testing. In the case of studies involving single candidate genes, Heiman et al. (8) addressed both issues by evaluating false-positive rates and comparing with confounding risk ratios due to population stratification. However, these evaluations have not been undertaken for genotype-genotype interactions. Marchini et al. (22) considered power issues due to population stratification, which have not been considered here. Finally, we have not evaluated additional situations of potential interest, such as the case of a main effect of a gene in one population but not in another.

When genotype distributions are the same across ethnicity or independent of ethnicity, ignoring ethnicity may still result in attenuation of the OR due to the nonlinearity of the logit link function in logistic regression (18-20). In the context of a single gene having same distributions in two or more ethnicities, bias is absent if the gene has no effect even when ethnicities are ignored. Otherwise, attenuation of the estimate towards the null will increase with the magnitude of the gene’s effect. The magnitude of the biases would be negligible unless disease risks were also extremely different across ethnicities. We found that when interactions between two genes are considered, the same rules applied to biases in the main effects of the two genes. However, bias may occur for interaction estimates even absent interaction (i.e., $\beta_3 = 0$). In addition, the direction of the bias is not always predictable when interaction is present (i.e., $\beta_3 \neq 0$). Nonetheless, the magnitude of these biases to main effects or interactions was generally negligible, unless the $\beta_1$, $\beta_2$, and $\beta_3$ were all large (>1.6), and the relative disease risk between the two ethnicities was >10-fold.

We have evaluated biases under relatively extreme conditions of population stratification with respect to disease risk and allele frequencies. The ranges of variables employed here were similar or more extreme than those in other studies (10, 23). In real situations, biases on genotype-genotype interactions should be smaller when ethnicity strata are more numerous, and the range of disease risk is narrower than considered here. Furthermore, our results are consistent with those of Wacholder et al. (1) and Wang et al. (6) as to the smaller potential for bias in the presence of larger numbers of ethnicities. Similar arguments hold for gene-environment interaction studies. For studies of gene-gene interactions, linkage disequilibrium patterns differ by ethnicity (13). Therefore, studies of genotype-genotype interactions should specifically consider potential linkage disequilibrium, baseline disease risks, and genotype frequency differences by ethnicity. This issue takes on special significance in light of suggestions that population-specific linkage disequilibrium might contribute to nonreplication of association study results, including studies of genotype-genotype interactions (24).

The data presented here support the hypothesis that bias due to population stratification can occur in association studies involving genotype-genotype interactions, particularly if the two genes are in strong linkage disequilibrium. However, our results show that the magnitude of potential bias is constrained by the differences in disease risk among populations. Thus, when these disease risk differences among populations are small, population stratification cannot lead to large biases. Furthermore, our empirical results show that population stratification causes relatively small biases even under extreme conditions and is unlikely to cause large biases to estimates of main effects and interactions under usual study conditions, particularly when the correlation (i.e., linkage disequilibrium) among the interacting factors is small. Therefore, if population stratification is not a major concern, studies of interaction involving unlinked genes (e.g., genes in common metabolic pathways that are located on different chromosomes) might be appropriate for case-control association studies, whereas haplotype-based approaches might be more appropriate for genes in linkage disequilibrium.
Appendix 1: Algebraic Analyses of Asymptotic Bounds on Biases Due to Population Stratification

A. Algebraic Analyses of Biases when $k = 2$ Ethnicities. The log-likelihood for a cohort sample of $N$ independent observations can be written as

$$I = \sum_{i=1}^{N} Y_i \times \ln \pi_i + \sum_{i=1}^{N} (1 - Y_i) \times \ln(1 - \pi_i).$$

where $i = 1, 2, \ldots, N$. When there are two ethnicities $E_1$ and $E_2$ in the cohort, they can be referred to as the high-risk ethnicity and low-risk ethnicity, respectively. Assume associations between disease and $V_1$ and $V_2$ are modeled with logistic regression as

$$\logit(\pi) = z_1 + z_2 \times E_2 + b_1 \times V_1 + b_2 \times V_2 + b_3 \times V_1 \times V_2 \quad (A)$$

In the presence of population stratification, we assumed the following mis-specified model was fitted:

$$\logit(\pi) = \alpha^* + \beta_1^* \times V_1 + \beta_2^* \times V_2 + \beta_3^* \times V_1 \times V_2 \quad (B)$$

Let $f_{V_1V_2E_2}$ represent the expected fraction of joint occurrence of $V_1$ and $V_2$ in ethnicity $E_1 (E_2 = 0)$ or $E_2 (E_2 = 1)$, where the subscripts $V_1$, $V_2$, and $E_2$ each take values 0 or 1. For example, $f_{000}$ represents the fraction of observations having joint reference categories of $V_1$ and $V_2$ in the high-risk ethnicity $E_1$ (i.e., $V_1 = 0$, $V_2 = 0$, $E_2 = 0$); likewise, $f_{001}$ represents the fraction of observations having the joint reference categories of $V_1$ and $V_2$ in the high-risk ethnicity $E_2$ (i.e., $V_1 = 0$, $V_2 = 0$, $E_2 = 1$). Consequently, the expected values of $D$ are $\pi_{000} = \pi_{101}$ and $\pi_{001} = \pi_{102}$ by Eq. (A), where $\pi(x)$ is the logistic function defined by $P(x) = \exp(x) / \{1+\exp(x)\}$. Let $\epsilon$ and $1 - \epsilon$ represent the proportions of the high-risk and low-risk ethnicity in the cohort, respectively, then the expected fraction of observations having the joint reference categories in the entire cohort is $f_{000} = (1 - \epsilon) \times f_{000} + \epsilon \times f_{001}$, where the “*” subscript indicates that observations are pooled over the associated index ethnicity in this case. Let $\pi_{00}$ be the estimated expected value of $D$ for these observations under the mis-specified model (Eq. B), then $\pi_{00} = P(\hat{\alpha})$, etc. Then the expected values of the maximum likelihood estimates of the variables in Eq. B are found by

$$\alpha^* = \logit\{([1 - \epsilon] \times f_{000} / f_{000}) \times P(x_1) + ([\epsilon \times f_{011} / f_{011}) \times P(x_2)]

\beta_1^* = \logit\{([1 - \epsilon] \times f_{101} / f_{101}) \times P(x_1 + \beta_1) + ([\epsilon \times f_{111} / f_{111}) \times P(x_2 + \beta_2)]

\beta_2^* = \logit\{([1 - \epsilon] \times f_{110} / f_{110}) \times P(x_1 + \beta_2) + ([\epsilon \times f_{110} / f_{110}) \times P(x_2 + \beta_2)]

\beta_3^* = \logit\{([1 - \epsilon] \times f_{110} / f_{110}) \times P(x_1 + \beta_2) + ([\epsilon \times f_{110} / f_{110}) \times P(x_2 + \beta_2)] - z^* - \beta_1 - \beta_2 - \beta_3 = 0

Because $(1 - \epsilon) \times (f_{000} / f_{000} + \epsilon \times f_{001} / f_{001}) = 1$ and $P(x)$ is monotonic function, $z_1 \leq \alpha \leq z_2$, i.e., $\alpha = \alpha_1$ only when $f_{000} = 0$ (i.e., when no observations in high-risk ethnicity fall into the joint reference categories of $V_1$ and $V_2$); and $\alpha = \alpha_2$ only when $f_{000} = 0$ (i.e., when no observations in low-risk ethnicity fall into the joint reference categories of $V_1$ and $V_2$).

Similarly, $(1 - \epsilon) \times (f_{100} / f_{100} + \epsilon \times f_{101} / f_{101}) = 1$; therefore, $\beta_1 + \beta_2 \leq \beta_1 + \beta_2$, because $z_1 \leq \beta_1 + \beta_2 \leq \beta_1 + \beta_2$, then $\alpha_1 - \beta_2 \leq \alpha_1 - \beta_2$.

As an example for the notation, here, the expected fraction of observations in the joint reference category in the entire cohort is $f_{000} = \epsilon \times f_{000} + \epsilon \times f_{001} + \ldots + \epsilon \times f_{00k}$, where the “*” subscript indicates that observations are pooled over the associated index ethnicity in this case.
Without loss of generality, assume baseline risks $\pi_{001} < \pi_{002} < \ldots < \pi_{00k}$, so that $0 < x_2 \leq x_3 \ldots \leq x_k$. Then biases on intercept term satisfy $x_1 \leq x \leq x_1 + x_2$, $x = x_1$ only when $f_{002} = \ldots = f_{00k} = 0$ (i.e., ethnicities 2 to $k$ do not have joint reference category of $V_1 = 0$ and $V_2 = 0$). In addition, $x = x_1 + x_2$ only when $f_{001} = f_{002} = \ldots = f_{00k-1} = 0$ (i.e., ethnicities 1 to $(k-1)$ do not have joint reference category of $V_1 = 0$ and $V_2 = 0$). Biases on $V_1$ effect estimates satisfy $\beta_1 - x_2 \leq \beta_1 \leq \beta_1 + x_2$, $\beta_1 = \beta_1 + x_2$ only when $f_{002} = f_{003} = \ldots = f_{00k} = 0$ and $f_{101} = f_{102} = \ldots = f_{10(k-1)} = 0$ (i.e., ethnicities 2 to $k$ do not have joint categories of $V_1 = 0$ and $V_2 = 0$, and ethnicities 1 to $(k-1)$ do not have joint categories of $V_1 = 0$ and $V_2 = 0$).

There will be no bias on $V_1$ effect estimates (i.e., $\beta_1 = \beta_1$) if $f_{001} = f_{002} = \ldots = f_{00(k-1)} = 0$ and $f_{101} = f_{102} = \ldots = f_{10(k-1)} = 0$, or if $f_{002} = f_{003} = \ldots = f_{00k} = 0$ and $f_{101} = f_{102} = \ldots = f_{10k} = 0$ (i.e., ethnicities 2 to $k$ do not have joint categories of $V_1 = 1$ and $V_2 = 0$, and ethnicities 1 to $(k-1)$ do not have joint categories of $V_1 = 0$ and $V_2 = 0$).

Acknowledgments

We thank Warren Ewens, Peter Kanetsky, Caryn Lerman, Richard Spielman, and Thomas Ten Have for their helpful comments during the development and execution of this research.

References

Evaluating Bias due to Population Stratification in Epidemiologic Studies of Gene-Gene or Gene-Environment Interactions

Yiting Wang, Russell Localio and Timothy R. Rebbeck


Updated version
Access the most recent version of this article at:
http://cebp.aacrjournals.org/content/15/1/124

Cited articles
This article cites 22 articles, 4 of which you can access for free at:
http://cebp.aacrjournals.org/content/15/1/124.full#ref-list-1

Citing articles
This article has been cited by 3 HighWire-hosted articles. Access the articles at:
http://cebp.aacrjournals.org/content/15/1/124.full#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, use this link
http://cebp.aacrjournals.org/content/15/1/124.
Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.