Asymptotic Bias and Efficiency in Case-Control Studies of Candidate Genes and Gene-Environment Interactions: Basic Family Designs

John S. Witte,1 W. James Gauderman,2 and Duncan C. Thomas2

Case-control designs that use population controls are compared with those that use controls selected from their relatives (i.e., siblings, cousins, or "pseudosibs" based on parental alleles) for estimating the effect of candidate genes and gene-environment interactions. The authors first evaluate the asymptotic bias in relative risk estimates resulting from using population controls when there is confounding due to population stratification. Using siblings or pseudosibs as controls completely addresses this issue, whereas cousins provide only partial protection from population stratification. Next, they show that the conventional conditional likelihood for matched case-control studies can give asymptotically biased effect estimates when applied to the pseudosib approach; the asymptotic bias is toward the null and disappears with disease rarity. They show how to reparameterize the pseudosib likelihood so this approach gives consistent effect estimates. They then show that the designs using population or pseudosib controls are generally the most efficient for estimating the main effect of a candidate gene, followed in efficiency by the design using cousins. Finally, they show that the design using sibling controls can be quite efficient when studying gene-environment interactions. In addition to asymptotic bias and efficiency issues, family-based designs might benefit from a higher motivation to participate among cases' relatives, but these designs have the disadvantage that many potential cases will be excluded from study by having no available controls. Am J Epidemiol 1999;149:693-705.

A major thrust of genetic epidemiology is exploring the effects of candidate genes and gene-environment \((G \times E)\) interactions on disease risks. Most of these studies look for association using the case-control design, particularly when investigating rare diseases. Although one treats candidate genes just like exposures of interest in a conventional epidemiologic study, many of the usual concerns (e.g., recall bias, temporal ambiguity) do not arise. Appropriate selection of controls, however, remains a primary concern.

A well-established tenet of case-control study design stipulates that cases and controls should be selected from the same source population (1). When studying potentially hereditary diseases, this implies that cases and controls should have similar genetic backgrounds. Because such information is usually not available at the time of subject selection, controls are often matched instead on race or ethnicity, so they are more representative of the source population of cases. For example, a recent population-based study of prostate cancer (2) required that controls have three of four grandparents from the same ethnic group as the case to which they were matched. Nevertheless, the broad categories of race or ethnicity generally used for matching (e.g., Caucasian, African-American, Latino, Asian) still leave room for potential confounding by different genetic backgrounds within ethnic groups. Geneticists call such confounding "population stratification," and examples of false positives due to this phenomenon indicate that it should be carefully thought about in any genetic epidemiologic association study (3, 4).

For example, Knowler et al. (5) show that confounding by population stratification led to a purported association between variants in the immunoglobulin genes \(Gm\) and non-insulin-dependent diabetes mellitus in American Indians. In particular, among residents of the Gila River (Arizona) Indian Community, diabetes was inversely associated with the haplotype \(Gm^{3,5,13,14}\).
When restricting their analysis to ethnically homogeneous subjects (i.e., full heritage Pima-Papago Indians), however, this association disappeared. The confounding by population stratification occurred because the Gm$^{3,5,13,14}$ haplotype serves as a marker of Caucasian heritage, and the risk of diabetes varies inversely with the level of this ancestry. Fortunately, Knowler et al. (5) were able to estimate the fraction of Indian heritage, and thus adjust for the confounding arising from studying an ethnically heterogeneous population.

Another example where there might be confounding due to population stratification is in the purported association between the A1 allele at the D2 dopamine receptor locus (DRD2) and alcoholism. Numerous studies give equivocal results: initial reports strongly suggested an association, while further studies fail to support this finding. Gelernter et al. (6) evaluated published studies attempting to replicate the initial association observed by Blum et al. (7), and found much greater heterogeneity among studies than differences between alcoholics and controls. Population stratification could explain this finding, as there are large ethnic differences in DRD2 alleles—and a wide range of corresponding allele frequencies among potential population-based controls—as well as ethnic differences in incidence of alcoholism (6). Unfortunately, unlike the Knowler et al. (5) report, most of the DRD2-alcoholism studies did not measure ethnicity and the few studies that were restricted to ethnically homogeneous populations did not observe an association (6).

Although these examples imply that population stratification might be a serious concern in genetic association studies, the potential magnitude of the bias resulting from this phenomenon remains unclear.

Within the framework of case-control studies, one approach to dealing with population stratification entails the use of family members (siblings or cousins) as controls. In fact, Lander and Schork (3) note that in light of the equivocal results for the DRD2-alcoholism association, it is crucial that future studies use familial controls to address the potential for confounding by population stratification. Each matched set will then arise from similar genetic source populations, essentially eliminating the potential for bias due to population stratification. Another family-based approach uses cases' parents as controls. Here, in each matched set one compares a case's genotype with those of a fictitious set of "pseudosibs" assigned the parent's genotypes that were not transmitted to the case (8, 9).

Using family members as controls can provide additional benefits, such as ease of contact and increased willingness to participate in genetic epidemiologic studies. However, these designs also have some weakness, such as the limited pool of potential controls (or parents in the pseudosib approach), and the possibility of lower efficiency due to "overmatching" on genotype and environment.

Case-control designs and family studies have been extensively discussed in the literature (e.g., see references 10–16). Nevertheless, there has been little explicit comparison of conventional population-based with family-based case-control designs in terms of bias and efficiency when estimating the effects of candidate genes and G × E interactions on disease. Here we compare case-control designs that use population controls with those that use controls who are related to the cases (siblings, cousins, or pseudosibs).

METHODS

Overview of case-control designs considered here

In the conventional case-control study, one randomly selects controls from the source population of cases. When selecting controls matched on potential confounders (e.g., age, sex, race, or ethnicity), one should be guided by the principles of risk-set sampling (17) to help ensure that the resulting matched odds ratio is a consistent estimator of the incidence rate ratio. In particular, controls should be sampled from the set of subjects meeting any matching criteria who have attained the age at which the case was diagnosed (the "reference age"), and were still disease-free at that age. Exclusion of controls who later develop the disease is neither necessary nor desirable; indeed, such exclusions will tend to bias an odds ratio away from the null (17).

Instead of population controls, one could match each case to his or her non-diseased sibling(s), and analyze the resulting matched data by using the standard conditional likelihood. Here, risk-set sampling with prospective ascertainment of cases will, depending on the time frame for case enrollment, effectively restrict controls to older siblings, because one cannot guarantee that younger siblings will remain disease-free until the age at which the case was diagnosed, nor predict their covariate histories up to that age. The exposure period of older siblings will not, however, be comparable to that of the cases in terms of calendar time or birth order effects. Thus, in comparison with population controls, older siblings are more susceptible to confounding by secular trends in environmental (i.e., non-genetic) exposures and to recall bias due to differences in the interval between exposure and interview. In principle, one can address both of the difficulties with the use of younger siblings by: 1) constructing a likelihood which allows for the possibility that the sib-

Am J Epidemiol Vol. 149, No. 8, 1999
ling will develop the disease before the case’s age; and 2) choosing a reference date that is sufficiently prior to the age at diagnosis of the case to include the entire exposure period of the control. Constructing the likelihood, however, would involve the baseline age-specific rates and will be considered elsewhere (B. Langholz, personal communication, 1998).

One could also match each case to an unaffected first cousin. Compared with using siblings as controls, this might allow for closer matching on age. In addition, there will usually be more cousins than siblings available as potential controls, so a higher proportion of cases may be informative for study, thereby reducing concerns about an unrepresentative case series. Nevertheless, in comparison with using siblings, one might expect a reduced participation rate for cousin controls due to lower motivation and geographic constraints. Furthermore, cousin controls do not provide the absolute protection from population stratification that sibling controls do, since cousins each have one parent that typically did not descend from a common ancestor.

Another design matches each case to hypothetical controls (pseudosibs) having the possible combinations of parental alleles not inherited by the case. A number of analytic approaches for evaluating case-parental allele data have been proposed, including the haplotype relative risk (HRR) approach (18–21) and the transmission/disequilibrium test (TDT) approach (9, 22). Here we investigate the case-control comparison of the case’s full genotype against those of the three possible pseudosibs, using conditional logistic regression for 1:3 matched sets (8, 23–25). More specifically, at a particular diallelic locus, the three pseudosib controls are defined by the genotypes not transmitted to the case. Then the three pseudosibs are defined as having genotypes A/D, B/C, and B/D, respectively. For example, assume that a case’s parents have genotype A/B and C/D, respectively, and that the genotype A/C was transmitted to the case. For example, assume that a case’s parents have genotype A/B and C/D, respectively, and that the genotype A/C was transmitted to the case. Then the three pseudosibs are defined as having genotypes A/D, B/C, and B/D, respectively. This approach allows one to fit any dominance model. Under a multiplicative dominance model, the maximum-likelihood estimate (MLE) and score test from the conditional logistic regression approach are equivalent to the McNemar odds ratio (OR) and chi-square test from the conventional (i.e., two-allele) TDT (8, 23–25). Thus, TDT is a special case of the conditional logistic regression approach evaluated here (25). (Weinberg et al. (26) recently investigated a similar pseudosib approach using a Poisson regression model.)

Assessment of asymptotic bias and efficiency

To compare the case-control designs outlined above, we first evaluate the potential asymptotic bias due to population stratification in estimating the main effect of a candidate gene. We then compare the designs’ asymptotic relative efficiency in estimating a main genetic effect, and an interaction between this candidate gene and an environmental factor. Our focus of inference is on the relative risk parameters, not the absolute penetrances. In our efficiency comparisons, we assume that there is no residual or population stratification confounding. Using the methods outlined below, we calculate infinite sample parameter values for our comparisons.

Let \( d_i \) represent the disease status of member \( j \) in case-control matched set \( i \). Since the overall score and Fisher information is simply the sum of contributions from each independent set, we assess the asymptotic bias (i.e., consistency) and efficiency of designs by considering the probability distribution of contributions from single case-control pairs; hence, we omit the matched-set subscript \( i \) in most of what follows. Let \( f \) and \( m \) represent the father and mother of the case in this matched set. We assume that the candidate gene under study \( g \) is diallelic with “high risk” allele \( A \) and “low risk” allele \( a \), having population allele frequencies \( q \) and \( 1 - q \), respectively. We denote the genotypes (AA, Aa, or aa) by \( g_i \) for the sample individuals (cases and controls), and by \((g_f, g_m)\) for the parents. We let \( G'_i = \{1, \Delta, 0\} \) denote a coding of the effect of the gene on penetrance, where \( \Delta \) indicates the assumed dominance for heterozygotes (Aa): 0 if the A allele is recessive, 1 if it is dominant, and \( \frac{1}{2} \) if the two alleles have multiplicative effects on the relative risk. For the sake of brevity and focus, we treat the dominance parameter \( \Delta \) as fixed (i.e., we restrict our investigation to recessive, multiplicative, and dominant models). Nevertheless, one could leave the parameter \( \Delta \) free in a codominant model. Finally, we let \( \chi \) denote a dichotomous environmental risk factor with population prevalence \( p \).

We base our calculations on the following logistic model for penetrance

\[
\text{logit}[\Pr(d = 1\mid g,s)] = \alpha + \beta G_s + \gamma x + \delta G_s x,
\]

where \( \alpha \) is the baseline risk for non-exposed, non-carriers in matched set \( i \), the coefficients \( \beta \) and \( \gamma \) are log odds ratios corresponding to the effects of genotype and environment, respectively, and \( \delta \) is the log of the ratio of odds ratios corresponding to the genotype-environment interaction. For shorthand, let \( \theta = (\beta, \gamma, \delta) \) denote the vector of
odds ratio parameters (estimating the relative risks (RRs)), and let \( \theta_0 \) denote their true values. In models without gene-environment (G \( \times \) E) interactions, we set the last two terms in equation 1 to zero. Note that one can incorporate G \( \times \) E interactions in the case-pseudosib design by comparing the genetic RRs between exposed and unexposed cases, or equivalently, by including an interaction term (but not a main effect) in the logistic model. As discussed further below, however, this approach assumes that the genetic and environmental exposures are independently distributed within families.

We assume that the disease outcomes are conditionally independent between family members, given their genotypes and measured environmental factors. Each design is analyzed by maximum likelihood, fitting a correctly specified logistic regression model appropriate to the design. For example, in a matched case-control study, the conditional likelihood is

\[
l(\theta) = \prod_i \frac{\exp(\beta G_{g_i} + \gamma x_{i1} + \delta G_{g_i} x_{i1})}{\sum_{j \in M_i} \exp(\beta G_{g_j} + \gamma x_{i1} + \delta G_{g_j} x_{i1})},
\]

where \( j = 1 \) denotes the case and \( M_i \) denotes the \( i \)th case-control set. Note that the conditional likelihood does not depend on the intercept terms \( \alpha \) from equation 1, so one does not have to assume they are equal across families nor model or otherwise estimate them. Nevertheless, conditional logistic regression does assume that all unmeasured family-specific factors can be aggregated into a single family-specific baseline risk, against which the measured factors have a multiplicative effect.

To investigate the potential asymptotic bias in the candidate gene effect \( \beta \), we compute the expected score statistic, evaluated at \( \theta_0 \), over all possible study outcomes. As an example, consider a main effects model for a candidate gene using a sibling-matched case-control design. Let \( U_\beta(\theta|g) \) denote the score function (i.e., the derivative of the log-likelihood) with respect to \( \beta \), evaluated at \( \theta \), that would be computed for a sibship with genotypes \( g = (g_f, g_m, g_r, g_2) \). Then the expected score is

\[
E[U_\beta(\theta_0)] = \sum_g U_\beta(\theta_0|g) \Pr(g|d_1 = 1, d_2 = 0; \theta_0, \alpha, q)
\]

where

\[
\Pr(g|q) = \Pr(g_f|q) \Pr(g_m|q) \Pr(g_r|g, q) \Pr(g_2|g_m, g_r). \]

For a design to be asymptotically unbiased—or more precisely, “Fisher consistent”—the corresponding expected score evaluated at the true parameter value must equal zero \( E[U_\beta(\theta)] = 0 \), or, equivalently, the maximum likelihood estimate must converge to the true value in large samples. For consistent designs, one can show this by demonstrating that the expected score equals zero for each possible genotype configuration. To quantify the consistency of alternative models, we have evaluated the expected score over all possible genotypes numerically for selected values of the parameters \( \theta, \alpha, \Delta, q, \) and \( p \), and computed the asymptotic expectation of the maximum likelihood estimate \( \hat{\beta} \) by a direct search for the value \( \hat{\beta} \) at which \( E[U_\beta(\hat{\beta})] = 0 \).

In the comparisons that follow, we vary the penetrances and allele frequencies such that the population attributable risk AR = \((\overline{R} - R_\infty)/\overline{R}\) and the population average disease rate \( \overline{R} = \sum_i R \Pr(g|q) \) (where \( R = \Pr(d = 1|g) \) is the penetrance and \( \Pr(g|q) \) is given by Hardy-Weinberg Equilibrium: \((1 - q)^2, 2q(1 - q), q^2\)), for genetic effects remain fixed. We determine the allele frequency by solving the equation for AR as a function of \( q \), and determine the baseline penetrance by an iterative search for \( \alpha \) such that

\[
\overline{R} = \sum_g \sum_x \Pr(g|q) \Pr(x|p) \Pr(d = 1|g, x). \]

To investigate the bias from population stratification, we consider the case of two strata \( s = 1, 2 \), with allele frequencies \( q_s \), and baseline risks (on a logit scale) \( \alpha_s \) and vary the ratios \( q_1/q_2 \) and \( e^{\alpha_s}/e^{\alpha_t} \) (holding the average allele...
frequencies $\tilde{q} = (q + q_2)/2$ and population rates $\tilde{R}$ fixed). For this purpose, we also consider the asymptotic bias under the null hypothesis $RR = 1$.

The asymptotic relative efficiency of alternative designs is defined as the inverse of the ratio of the variances of their estimated parameters, or, equivalently, as the inverse of the ratio of the sample sizes needed to attain the same precision and power, as evaluated at the true values of the parameters. We compute these variances as the inverse of the Fisher information matrix $I(\theta|g)$: the negative of the expected value of the matrix of second derivatives of the log likelihood. One can calculate the expected information in the same way as the expected score, replacing $U(\theta|g)$ by $I(\theta|g)$ in equation 3.

When evaluating the designs' asymptotic relative efficiency for estimating $G \times E$ interactions, we include a moderate true $G \times E$ interaction effect and shared environment within families. For the $G \times E$ interaction, we set $\delta = \ln(2)$, where $\delta$ is the interaction coefficient from equation 1 (i.e., corresponding to an odds ratio of 2). We allow for the shared environment factors using a regressive model (27). Specifically, we assume logistic models for husband-wife (HW), parent-offspring (PO), and sibling-sibling (SS) sharing of the environmental factor, such that

$$
\text{logit}[\Pr(x_f = 1)] = \eta_0 = \text{logit}(p)
$$

$$
\text{logit}[\Pr(x_m = 1|x_f)] = \eta_0 + \eta_{HW}(x_f - p)
$$

$$
\text{logit}[\Pr(x_1 = 1|x_m, x_m) = \eta_0 + \eta_{PO}(x_f + x_m - 2p)]
$$

$$
\text{logit}[\Pr(x_2 = 1|x_m, x_1) = \eta_0 + \eta_{PO}(x_f + x_m - 2p) + \eta_{SS}(x_1 - p)].
$$

(4)

Here, for the shared environment, we set $p = 0.5$ and varied $\eta_{HW} = \eta_{PO} = \eta_{SS}$ over a range of values from zero to $\ln(4)$ (i.e., corresponding to a range of environmental concordance). The expected score and information is then obtained by summing terms of the form given in equation 3 over all possible combinations of environment as well as genotype. Results are shown for an environmental main effect $e^\prime = 2$, but are quite insensitive to this choice.

RESULTS

Figure 1 shows the asymptotic bias in genetic RR estimates resulting from use of a conventional population-based case-control design in the presence of population stratification under a low penetrance, additive model. These infinite sample parameter values were obtained from the expected score statistic, as outlined above (i.e., using equation 3). The direction of the bias is what one would expect from the usual principles of confounding in epidemiology: if the ratio of allele frequencies and baseline risks are in the same direction, the bias is positive; if different, the bias is negative. In figure 1, we restrict our comparisons to situations when the stratum RRs ($e^{a_2}/e^{a_1}$) are greater than 1; using stratum RRs less than 1 would give a mirror image as that presented in figure 1. For all models evaluated here, in general, the larger either ratio, $q_f/q_2$ or $e^{a_2}/e^{a_1}$, the larger the bias. The bias is slightly less pronounced for a higher average population rate, and for dominant or recessive modes of inheritance (not shown). For some parameter values considered, the bias can be very severe—as much as tenfold. Regardless of whether or not there is population stratification of the type modeled here, using siblings gives unbiased estimates of the genetic RR. Using cousins is also unbiased, provided that the cousins derive from similar gene pools. If not, cousins remain susceptible to population stratification, although generally less so than population-based controls. For example, assume that there is population stratification, whereby $q_f/q_2 = 5$ and $e^{a_2}/e^{a_1} = 5$ in two strata. Also assume that cousins randomly arise from both strata. Then, under an additive model with an average genetic RR of 20, this population stratification results in population controls overestimating the true RR by 301 percent, while cousin controls only overestimate the RR by 11 percent.

While population stratification is not an issue for the pseudosib design, table 1 shows that, when using a conventional conditional logistic regression framework, this design can produce expected RR estimates that underestimate the true RRs. This asymptotic bias increases with increasing absolute penetrance, is proportionally larger for large relative risks, and is highest for the dominant and recessive models. (Note, however, that the corresponding score test—as well as the TDT—is valid.) The asymptotic bias in the RR estimate arises because this analysis assumes that the non-
transmitted genotypes would not have led to disease (i.e., the pseudosibs are unaffected). Hence, one should base the likelihood for the case-pseudosib design on the risk of disease, not the odds of disease. To deal with this, one can reformulate the conditional likelihood using only the logistic probabilities for the case (Appendix 1). The last three columns of table 1 compare the variance estimates from the conventional and reformulated likelihoods (both evaluated at the true RR). In some instances, the reformulation results in loss of efficiency, while in others it results in much improved efficiency. The loss of efficiency occurs because we must estimate additional baseline risk parameters when using the reformulated likelihood (see Appendix 1). For comparison's sake, the following investigation of efficiency focuses on situations when the asymptotic bias is not an issue ($R = 0.001$), so we can use the conventional conditional likelihood.

**FIGURE 1.** Population stratification in population-based case-control study for additive model. The ratio of allele frequencies in two strata ($q_1/q_2$) are plotted against the asymptotic bias (the ratio of expected, i.e., ignoring population stratification, and true relative risks, where the true RR = 1). Each line corresponds to a ratio of stratum-specific risks (i.e., $e^\beta_1/e^\beta_2$). The average allele frequency ($q_1 + q_2)/2 = 0.19$.

**TABLE 1.** Bias in pseudosib design: expected relative risk (RR) estimates from the pseudosib design without bias correction, and relative efficiency after bias correction (see Appendix 1)

<table>
<thead>
<tr>
<th>Inheritance</th>
<th>RR frequency</th>
<th>$\bar{p}^*$</th>
<th>Expected RR estimate</th>
<th>Relative efficiency†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.1</td>
<td>0.01</td>
</tr>
<tr>
<td>Recessive</td>
<td>20</td>
<td>0.14</td>
<td>7.4</td>
<td>17.5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.44</td>
<td>1.8</td>
<td>2.0</td>
</tr>
<tr>
<td>Multiplicative</td>
<td>20</td>
<td>0.02</td>
<td>11.0</td>
<td>18.6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.19</td>
<td>1.9</td>
<td>2.0</td>
</tr>
<tr>
<td>Dominant</td>
<td>20</td>
<td>0.01</td>
<td>7.5</td>
<td>17.5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.10</td>
<td>1.8</td>
<td>2.0</td>
</tr>
</tbody>
</table>

*R, population average disease risk.†Efficiency of bias corrected likelihood, relative to the uncorrected likelihood, both evaluated at the true RR (in percent).

Am J Epidemiol Vol. 149, No. 8, 1999
Table 2 presents the asymptotic relative efficiency of the case-control designs for estimating the main effect of a candidate gene. We give values relative to the efficiency of the population-control design, assuming that there is no confounding (e.g., due to population stratification). Here, use of siblings is about one-half as efficient under dominant and additive models, and is about two-thirds as efficient under the recessive model. Using cousins is slightly less efficient than using population controls, although this difference is limited under the recessive model. Varying the genetic RR, allele frequency, and population average disease rate leaves the sibling and cousin results essentially unchanged. In contrast, use of pseudosib controls provides equal or higher efficiency than using population controls; again, a recessive mode of inheritance provides the highest relative efficiency, and here the pseudosib design has increasing efficiency with increasing genetic RR or decreasing allele frequency (not shown).

For estimating interactions between a major gene and a dichotomous environmental exposure, use of siblings or cousins can actually be more efficient than using population controls. Specifically, sibling and cousin controls are more efficient when estimating an interaction comprising a rare major susceptibility gene (figure 2). In contrast, for a common gene, sibling and cousin controls remain somewhat less efficient than population controls, though less so than for the main genetic effect (not shown). The magnitude of the relative efficiency for sibling controls depends inversely on the degree of environmental sharing, whereas this has little effect on the efficiency of the case-cousin design (figure 2). Pseudosib controls are generally slightly less efficient than population controls for estimating $G \times E$ interactions, except under a recessive model (figure 2). Of course, with pseudosibs there is no dependence on the amount of environmental sharing for $G \times E$ efficiency considerations, although there is the assumption of conditional independence between genotype and environment (discussed below).

**DISCUSSION**

We have evaluated the bias and efficiency in estimating the effects of candidate genes and $G \times E$ interactions with case-control designs that use population and familial controls. Our results indicate the following:

1. Population controls can give asymptotically biased estimates of candidate gene effects, where the magnitude of bias depends on the extent of population stratification. Sibling and pseudosib controls eliminate the problem of population stratification. Cousin controls help deal with this potential problem, but are still susceptible to varying levels of population stratification.
2. Pseudosib controls can give slightly underestimated genetic RR estimates if one uses a conventional odds ratio formulation of the likelihood, although this bias disappears with disease rarity, and a reformulation of this likelihood corrects the bias.
3. When looking at the main effect of candidate genes: sibling controls can be considerably less efficient than using population controls; cousin controls are almost as efficient as using population controls; and pseudosib controls provide the most efficient family-based design. Furthermore, under a recessive mode of inheritance, the pseud-
dosing design is often substantially more efficient than the conventional population-control design.

4. When looking at $G \times E$ interactions, family-based controls can be more efficient than population controls.

In light of our findings, the issue of bias due to population stratification when using population-based controls merits additional consideration. For ethnicity to be a confounder, it must be related to the population allele frequency of the candidate gene under study and be an independent risk factor for the disease of interest. There are a few ways this might come about. Suppose first that there is another unmeasured gene which is causally related to the candidate gene under study, and which is the reason ethnic groups differ in their disease rates. Here, confounding can only occur if the two genes are associated in the population, i.e., are in disequilibrium with each other. If the two genes are also in linkage with each other, then an apparent association between the candidate gene and disease—even if not directly causal—might still be of interest as a means of localizing the truly causal gene. Furthermore, if an association with a particular locus is consistently replicated in a range of ethnic populations, but different alleles appear associated with the disease, the association is likely noncausal, but merely a sign of linkage disequilibrium. Disequilibrium can also occur between two genes that are not in linkage with each other. The simplest mechanism that could produce such an association is through admixture of populations with differing allele frequencies of the candidate and causal genes. In this case, an apparent association with the candidate gene could occur which is simply due to the causal gene, but the association would tell us nothing about the location of the true causal gene and would be of no interest. Repetition of the study in ethnically diverse populations would be unlikely to confirm the original association, but simple repetition in similar populations (e.g., different regions of the United States) might “confirm” the spurious association, or at least lead to conflicting results without any obvious explanation. Similarly, associations between a candidate gene and some unidentified “environmental” risk factor can occur through admixture of subpopulations with different prevalences of the two factors, even if the gene is not associated with the environmental factor within strata. Thus, the bias that is due to “population stratification” in case-control studies using population controls is not limited to genetic effects.
To deal with potential population stratification, one might match on self-reported ethnicity (or adjust for ethnicity in the analysis). This may be problematic, however, due to the difficulty of quantitatively characterizing ethnicity or other determinants of gene frequency. Instead, one could genotype individuals at a panel of highly polymorphic loci that distinguish ethnicity, and use this information as the basis for stratification (28-30). But one should not stratify on a locus that is in strong disequilibrium with the putative disease locus. If the panel is inexpensive to run, one could “pre-genotype” and match cases and controls on the markers. Otherwise, one could stratify on such markers during the analysis. For practical purposes, however, at the present time we contend that the issue of population stratification is best addressed by using family members as controls. As noted above, the literature contains examples of candidate gene associations that have not been verified by subsequent linkage. Some of these nonreplications are due to selection bias (discussed below), rather than to confounding bias per se; regardless, the use of family controls helps overcome these sources of bias, albeit possibly at the price of some loss of efficiency.

Another example illustrates a somewhat different perspective on population stratification. Bell et al. (31) reported an association of insulin-dependent diabetes mellitus (IDDM) with the 5'FP RFLP adjacent to the insulin gene on chromosome 11p. Whereas this association had been consistently replicated across several populations, no evidence of linkage was initially found using standard affected sib pair methods (32). It was this puzzling situation, which suggested that the association results might be due to population stratification, that led Spielman et al. (33) to develop the TDT, which showed highly significant evidence of linkage. Although the TDT result might have been spurious, due to subjects with the high risk alleles being eliminated from the unaffected population, a comparison with transmissions of the high-risk allele to unaffected offspring of heterozygous parents ruled out this interpretation. In this case, it appears that the association may be real and the initial failure of the affected sib pair linkage tests was due to their low power for detecting genes with modest relative risks. Here, the use of family member (pseudosib) controls helped confirm a population association that could otherwise have been dismissed as due to population stratification.

While sibling and pseudosib controls completely avoid the bias due to population stratification, cousin controls might provide only partial protection from this phenomenon, to the extent that pedigrees tend to derive from a common gene pool. One might attempt to deal with the potential for population stratification in the case-cousin design by selecting two cousins for each control—one from the mother’s side and one from the father’s side of the family. Thus, the two controls would, between them, represent both gene pools of the case. However, how well this 1:2 case-cousin approach deals with population stratification depends on where in the pedigree any intermarriage occurs. For example, if the case’s parents have different ethnic backgrounds, then this approach will generally address issues of population stratification. In contrast, if one of the potential cousin’s parents have different ethnic backgrounds, then this approach will not necessarily deal with the bias, and one might be best off excluding such cousins. Furthermore, this approach results in effect estimates that can be slightly biased toward the null, even in the absence of population stratification. This can occur because one should not be able to identify who is the case merely by examining the structure of the matched set. For this design, the case is the only member of the matched trio who is a blood relative of both of the other two members, and is thus immediately identifiable as the case. Therefore, when considering this 1:2 case-cousin design, one must balance the tradeoff between potential improvements over the 1:1 case-cousin design (i.e., with regard to population stratification) and the latter “identification” bias.

Although, the pseudosib design is the most efficient for investigating a candidate gene on a per-case basis, this approach requires genotyping three individuals per matched set whereas the other designs require genotyping only two. To further explore this difference, we evaluated the efficiency of the other designs considered here, expanded to include three individuals. Specifically, we calculated the efficiency of the sibling and cousin designs relative to the population design, where each design included two controls for each case. In addition, we calculated the efficiency of the pseudosib design relative to the 1:2 population design. The bottom panel of table 2 gives the relative efficiencies. Here, the pseudosib design was generally no longer more efficient than the population design, and was in fact less efficient than the cousin design—which approximately retained its relative efficiency—in all but one situation (recessive, high risk allele). The sibling design had even lower relative efficiency than observed for the 1:1 design. Of course, as with the other comparisons made here, there are other issues that must be considered beyond how many individuals need to be genotyped. In particular, one must also consider the expense and feasibility of recruiting cases, parents, and relative controls (discussed further below).

The reduced efficiency from using siblings as controls to estimate a candidate gene effect reflects their greater concordance on genotype. This concordance,
however, helps drive the increased efficiency observed under certain conditions when estimating $G \times E$ interactions. As an example, consider a dominant model when the $G$ allele is rare and is a strong independent risk factor. Cases will be more likely than population controls to carry the allele, and one expects a higher allele frequency among sibling controls and cousin controls than in population controls. Furthermore, the primary determinant of the efficiency of an interaction estimate is the number of study subjects in the smallest cell. Here, the smallest cell will contain controls who have the $G$ allele and who are exposed to the environmental factor $E$ (if exposure is uncommon), and the frequency of that cell will be higher in sibling or cousin controls than in population controls. Therefore, the estimate of the $E$ effect in the carrier stratum, and hence the interaction, will be estimated more precisely by sibling or cousin controls than by population controls. As the allele becomes more common, the increased efficiency observed for sibling or cousin controls diminishes. We provide a more technical explanation of this phenomenon in Gauderman et al. (34). In addition, Thomas and Greenland (35) give an extensive investigation of efficiency in matched case-control studies looking at interactions.

Of course, one does not need controls to investigate $G \times E$ interactions (36-38). Under the assumption that genes and environmental exposures are independently distributed in the source population, an efficient estimator of their interaction is simply the log odds ratios from an unmatched association between genes and environment among cases. For the situations considered here, this case-only approach is 2-10 times more efficient than using population controls (not shown). This greater efficiency derives from replacing the information on the joint distribution of $G$ and $E$ in the source population by the assumption that the two are independently distributed. If this assumption is violated, the odds ratio estimates the $G \times E$ interaction effect times the odds ratio for that association in the source population. Therefore, this approach should only be used for genetic and environmental factors where such an assumption is tenable, at least within definable strata. For example, in a recent breast cancer case-only investigation of the interaction between BRCA1 mutations and oral contraceptive (OC) use (39), the observed association could theoretically reflect a lack of independence between OC use and family history. In particular, the $G \times E$ independence assumption might be violated if women with a strong family history of breast cancer are more or less likely to take OCs. The assumption of BRCA1 genotype and OC use independence is arguably more plausible within strata defined by family history of breast cancer.

The case-pseudosib design also appears to offer a simple way to test $G \times E$ interactions: all one need do is to stratify the cases on $E$ and compare the genetic RRs between exposed and unexposed. However, now one must make an assumption of conditional $G \times E$ independence given parental genotypes (i.e., within families) (Clarice Weinberg, personal communication, 1998). This assumption is somewhat weaker than required for the case-only study, where the assumption of $G \times E$ independence must hold for the population at large. For example, population stratification could violate the case-only assumption of $G \times E$ independence, while not affecting the case-pseudosib assumption of conditional $G \times E$ independence.

For all of the designs considered here, one must consider the potential for selection bias, specifically whether the probability of selection depends on genotype. For example, if controls with a positive family history are more willing to participate, then the set of cases with available controls would tend to overrepresent the gene frequency in all cases. Alternatively, the genotype might affect the existence of a sibling, for example, if having an affected child affects parents’ willingness to have additional offspring or if affected parents have lower fertility (due to premature truncation of their reproduction by death or disability). In these situations, using cousins instead of siblings could result in a more representative set of subjects. However, it is worth noting that all of these examples would lead to unrepresentativeness of the case series but not bias in the estimated genetic relative risk, since cases and controls would be excluded equally. As another example, if parental missingness is associated with the genotype of interest, excluding cases from the pseudosib design without parental information could result in a non-representative set of cases (e.g., if mortality is associated with the gene under study). Nevertheless, when dealing with diallelic markers, one should exclude cases from the pseudosib design when a parent is missing; when there are more than two alleles, no bias will occur if one includes cases who are heterozygous and have a different genotype than the observed parent (40). The pseudosib and related “parental-allele" designs are considered further elsewhere (9, 19-26, 32, 33).

Regardless of the design used, associations detected in case-control studies can be due to the direct effect of a causal gene or the induced RR of a noncausal locus that is in linkage disequilibrium with the causal gene. Although these two situations are essentially indistinguishable with case-control data, they are both interesting, the first as a causal hypothesis, and the second as a means of localizing the causal gene. Therefore, here we have not distinguished between these situations, taking as the parameter of interest the relative risk corresponding to the candidate gene or gene-

Am J Epidemiol Vol. 149, No. 8, 1999
environment interaction (whether direct or induced). Throughout, we assume that the genetic and $G \times E$ relative risk estimates are homogeneous quantities across the population at risk. If they vary across subgroups, then any genetic epidemiologic study will estimate some form of weighted average of the distribution of relative risks. The particular form of weighted average will vary from one design to another. We do not specifically consider this situation further here, on the grounds that if there really is heterogeneity in the RR's, what is needed is a description of the distribution and determinants of the variability in RR's rather than a single summary of its central tendency.

We limited consideration to a single, dichotomous environmental factor, although one could easily include categorical, continuous, and time-dependent covariates in the standard logistic model. We also assumed that the disease outcomes are conditionally independent between family members given their genotypes. Two separate mechanisms could lead to the dependence of family members' outcomes given their genotypes: shared dependence on other unmeasured risk factors and direct causal influences of one member's phenotype on another's. This does not induce any bias, however, since all the comparisons are made within family (i.e., using the conditional likelihood) and are thus not affected by residual familiality (see Appendix 2). In addition, since the same set of cases could be used for any of the designs discussed here, one might consider a hybrid approach that selects some combination of relatives as controls. Further work by the present investigators is exploring the efficiency of case-control designs when restricting subjects to those coming from families that have multiple individuals with disease.

In any case-control study of candidate genes and $G \times E$ interactions, one should determine the genotypes of cases and controls in the same manner (or at least subject to the same measurement errors). However obvious this may seem, there are subtleties involved. For example, in a study of rare mutations (such as in the BRCA1 gene), one might be tempted to sequence the cases first and then genotype controls for the same mutant allele as the case. However, the only discordant case-control pairs this approach could detect would be case-carrier/control-noncarrier, leading to an infinite relative risk estimate even if there was no association. Hence, one must either genotype all subjects "blind" to case status by direct sequencing even though one might anticipate a small yield of case-noncarrier/control-carrier pairs or by using only the results of a cheaper screening test (such as allele-specific oligonucleotides for a fixed panel of mutations) for cases and controls equally (despite the possibility of missing some mutations, which would be nondifferential).

The asymptotic bias and efficiency comparisons presented here must be integrated with practical considerations surrounding the use of family-based controls. Whereas one might anticipate increased motivation to participate among family members of cases, the familial control requirement could substantially restrict the number of eligible matched sets. For the sibling and cousin designs, matching with respect to sex or age will limit this number even further, particularly in countries with low fertility rates (e.g., China). As an example, in the University of California Irvine Breast Cancer Registry, only about one-quarter of 1,600 breast cancer probands could be matched to older, non-diseased sisters (H. Anton-Culver, personal communication, 1998). Furthermore, while the pseudosib design does not require that a case have a nondiseased sibling or cousin, it does require genotypic information from both the case's parents (40); this information may be difficult to obtain, especially if one is studying a disease that primarily affects older people.

An attractive approach for dealing with the potential bias due to population stratification entails using siblings, cousins, or pseudosibs as controls. When designing genetic epidemiologic case-control studies, however, numerous other issues also require consideration. In particular, one must also consider the efficiency, feasibility, expense, and other sources of bias that arise from using a particular set of controls. The relevance of these issues will vary from study to study, and thus require thoughtful examination during the design phase of each investigation.

ACKNOWLEDGMENTS

This work supported by National Institutes of Health grants CA73270 and RR03655 (J. S. Witte), and CA52862 and ES07048 (D. C. Thomas and W. J. Gauderman).

The authors thank Robert Elston, Robert Haile, Nan Laird, Bryan Langholz, Stephanie London, Nat Rothman, Sholom Wacholder, and Clare Weinberg for helpful comments.

REFERENCES

4. Altshuler D, Kruglyak L, Lander ES. Genetic polymorphism

**APPENDIX 1**

**Correcting the bias in the pseudosib design**

Equation 2 gives the likelihood for the pseudosib design, where \( M_i \) denotes the set of four possible genotypes \( g_i \) that case \( i \) could have inherited given the parents’ genotypes. Substituting this likelihood into the expression for the expected score, equation 3, produces

\[
E[U_{\beta}(\theta_0)] = \sum_{g_i} \Pr(g_i|q) \Pr(g_{m}|q) \sum_{g_m} \left( \frac{e^{\alpha g_i + \beta g_m} G_{g_i}}{1 + e^{\alpha q + \beta q} G_{q}} \right) \left( G_{g_m} - \frac{\sum_{g_{m}} G_{g_m} e^{\beta g_m}}{\sum_{g_{m}} e^{\beta g_m}} \right)
\]
This expression will not usually equal zero, unless $\beta_0 = 0$ or in the limit as $\alpha_0 \to -\infty$. The problem here is that the score contribution is based on an odds ratio approximation to the relative risks, not the penetrances themselves that form the weights for each possible proband contribution. This problem does not arise when using real controls (provided they are appropriately selected from their age-matched risk sets $M_i$, following the principles described by Lubin and Gail (17)), because the weights are replaced by

$$\Pr(d_1 = 1, d_2 = 0 | g_1, g_2, d_1 + d_2 = 1) = \frac{e^{BG_i}}{e^{BG_i} + e^{BG_n}}. \quad (5)$$

In order to correct the bias in the pseudosib design, one can reformulate the conditional likelihood in terms of the penetrances themselves rather than the penetrance odds ratios. The appropriate conditional likelihood then becomes

$$f(\alpha, \beta) = \Pr(g_i | g_f, g_m, d_i = 1) = \frac{\Pr(d_i = 1 | g_i) \Pr(g_i | g_f, g_m)}{\sum g \Pr(d_i = 1 | g) \Pr(g_i | g_f, g_m)}$$

$$= \frac{e^{\alpha + BG_i}/(1 + e^{\alpha + BG_i})}{\sum g d_i e^{\alpha + BG_i}/(1 + e^{\alpha + BG_i})}.$$ \quad (6)

which is now an explicit function of the baseline rate $\alpha$. One cannot estimate this parameter directly from the case-pseudosib data alone. However, if the population average rate $\bar{R}$ is known, $\alpha$ can be expressed as a function of $\beta$ and $\bar{R}$. Then the profile likelihood $f(\alpha(\beta, \bar{R}), \beta)$ provides a consistent estimator of $\beta$, and the asymptotic variance of $\beta$ from this profile likelihood is obtained from the Fisher information.

**APPENDIX 2**

Dependent outcomes in familial case-control designs

There are two ways in which family members can have dependent outcomes given their genotypes. First, if they have a shared dependence on other unmeasured risk factors. Second, if there is a direct causal influence of one member’s phenotype on another’s. Both types of residual dependency are controlled for by conditioning on family membership using the conditional logistic regression formulation we have adopted.

In the first situation, we assume that the matched sets (sibships or case-cousin sets) differ in their baseline risks, but all members share a common dependence on unmeasured factors, subsumed in the baseline risk parameter $\alpha$. Then the conditional argument completely eliminates any such dependency of the likelihood for the RR parameters $\beta$ on these $\alpha$s. In particular, if the matched sets consist of only two individuals, the conditional argument is sufficient. If there are more than two members of a matched set and members may differ in the dependence of their baseline risk on common factors, the necessary assumption is that the members are “exchangeable” with respect to such dependencies, i.e., that such dependencies are independent of the measured variables.

The second situation is subtler. Suppose, for example, we postulate a regressive model for sibships, in which the phenotype of younger sibs depends on their own genotypes as well as on the phenotypes of their older sibs. Then, for a matched pair with the older sib being the case, the conditional logistic regression can be written as:

$$\Pr(d_1 = 1 | g_1, g_2, d_1 + d_2 = 1) = \frac{\Pr(d_1 = 1 | g_1) \Pr(d_2 = 0 | g_2, d_1 = 1)}{\Pr(d_1 = 1 | g_1) \Pr(d_2 = 0 | g_2, d_1 = 1) + \Pr(d_1 = 0 | g_1) \Pr(d_2 = 1 | g_2, d_1 = 0)}.$$

Substituting a logistic function for $\Pr(d_2 | g_2, d_1)$ into this expression and averaging over the two possible outcomes that could be observed, it becomes clear that the “naive” logistic regression that omits the dependence of $d_2$ on $d_1$ provides a consistent estimator of the RR parameter for $g$ that is of real interest.