COMMENTARY

Choosing a Retrospective Design to Assess Joint Genetic and Environmental Contributions to Risk

C. R. Weinberg and David M. Umbach

The authors consider issues that should be weighed when designing a retrospective study in which a focus of interest is the joint role of genetic and environmental factors in causing a disease. In place of the classical case-control design, in which controls are sampled from the same population that gives rise to the cases, one could study cases only. The case-only approach can be usefully extended by genotyping the two biologic parents of each case and in effect letting the parental genotype data provide the genetic control. Alternatively, one could carry out a case-control study in which the controls are siblings or cousins of the cases and inference is based on within-family parameters. The authors compare and contrast the parameters that can be estimated and the assumptions that must be made when each of these designs is used. The investigator must also consider certain practical issues, such as the availability of parents or sibling controls.

With the advent of the polymerase chain reaction and the subsequent burgeoning knowledge of molecular genetics, epidemiologists have powerful new tools to piece together the roles of environmental agents and genetic susceptibility in causing disease. As epidemiologists increasingly focus on genetic contributions to etiology, some geneticists are becoming more interested in the role of exposures, particularly in relation to genes involved in metabolic and detoxification processes. The two cultures, epidemiology and genetics, can profitably exchange technologies and learn each other’s beliefs and languages in an inevitable cross-fertilization that should prove fruitful to both. But, in this dynamic context of evolving approaches, the best design for jointly studying genetic and environmental factors is not always obvious.

Geneticists have tended to mistrust case-control “associational” studies because of their potential vulnerability to effects of genetic “admixture.” Admixture, or “stratification,” arises when genetically diverse subpopulations are incompletely mixed. In an admixed population, some subgroups may carry both a higher prevalence of an allele of interest (and/or an exposure under study) and a higher risk of disease. In such a population, an association between a variant allele and risk could be noncausal, reflecting only confounding with subpopulation. Although some have argued that admixture causes little bias in a diverse population such as the United States (1), family-based approaches offer robustness against this problem. For this and other (perhaps better) reasons, for example, convenience of implementation, recruitability of suitable controls, or statistical efficiency, family-based designs are worth considering.

The purpose of this commentary is to lay out some of the issues that should be considered in choosing a retrospective design for studying joint effects of a candidate gene (or a marker gene) and an exposure. The issues related to using a prospective cohort study are important but lie beyond our present scope. To the extent that nested retrospective designs might also be incorporated in that context, however, some of our comments will be relevant there as well. We

Received for publication July 23, 1999, and accepted for publication March 20, 2000.

Abbreviations: E, exposure; G, genetic factor.
From the Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC.
Reprint requests to Dr. C. R. Weinberg, Biostatistics Branch, MD A3-03, National Institute of Environmental Health Sciences, PO Box 12233, Research Triangle Park, NC 27709 (e-mail: weinberg@niehs.nih.gov).
begin by describing a range of retrospective designs that could be used, discussing available methods of analysis for these designs, and then contrast the assumptions needed by the various strategies.

RETROSPECTIVE DESIGNS

The classical workhorse of retrospective designs is the population-based case-control study, in which a random sample of cases is studied and compared with a set of controls from the same population that gave rise to the cases. Sampling of controls can be (and usually is) constrained by matching factors, such as age and sex. Joint effects of genetic and environmental factors can be studied by using such a design, although the sample size required for assessing interactions tends to be very large (2). Much has been written about the case-control design, and our purpose here is not to add to that discussion but to consider certain variations and alternatives that have been proposed recently.

The most economical alternative is the case-only design, first set forth as a method for studying gene-by-environment interaction for susceptibility genes in 1994 (3). If an inherited dichotomous genetic factor, G, and a dichotomous exposure, E, can be assumed to occur independently in the population, then one can show algebraically that the odds ratio for G in relation to E among cases can be treated as an estimate of the multiplicative interaction, the ratio of relative risks (4), between G and E in causing the disease. Consequently, one can assess such interactions by carrying out logistic regression analysis using only cases, treating G as the outcome. In fact, under the required independence assumption, the precision of the estimate based only on cases is superior to that based on a classical logistic regression analysis of data that has included any number of controls for the same cases. A recent application of the case-only approach demonstrated an interaction between inheritance of the null-null genotype for glutathione S-transferase M1 and exposure to environmental tobacco smoke in causing lung cancer in women who had never smoked (5). This study made good use of existing archived tissue specimens for cases, but, as was pointed out in an accompanying editorial (6), the investigators did not fully use their existing data: Even though genotype data for controls were lacking, what data were available could have been used to estimate exposure effects within each specific genotype category (7). Such a set of estimates is epidemiologically more informative than is a single interaction parameter.

Independence of G and E is often a natural assumption, but it can certainly fail to hold. The prevalences of E and G can both vary across subpopulations, inducing a correlation. Some genetic factors, for example, those governing metabolic processes, can actively influence certain exposures. In addition, genetic factors strongly related to risk, for example, BRCA1 or BRCA2 with breast cancer, could become increasingly related to other risk factors across age-time, through selective attrition. While the required assumption of the independence of genotype and exposure cannot be verified without including a sample of controls, the case-only design nevertheless can be useful for inexpensively exploring possible gene-by-exposure interactions by using archived tissue from cases. But once an E-by-G association has been noted in a case-only study, methods that are epidemiologically more rigorous, such as a population-based case-control study, should be used for follow-up confirmation.

Both approaches, however, are theoretically subject to bias due to genetic admixture, as noted above. With the case-only approach, bias could arise because of incomplete mixing of subpopulations that differ by exposure prevalence and genotype prevalence, even if their baseline risks of disease do not differ. Hence, family-based approaches might offer some advantages over both the case-only and the case-control designs.

In contrast with the population-based case-control design, family-based designs enroll related persons to serve as controls, a strategy that confers robustness against the potentially biasing effects of genetic admixture. Such controls could include the sibling(s) of affected persons or perhaps, as suggested recently by Witte et al. (8), cousins of affected persons. Of course, use of siblings (or cousins) raises the troublesome issue that some cases may have to be excluded only because they do not have a suitable relative willing to participate. Another approach, particularly well adapted to studying a disease with early onset, such as a birth defect, insulin-dependent diabetes, or schizophrenia, is instead to genotype the parents of affected persons and study the apparent transmission of particular alleles from such parents to their affected offspring.

Studies in which siblings or cousins are used as controls are useful for diseases with early or later onset and can be analyzed by using conditional logistic regression (8), with fine stratification on family. Under the usual assumptions required for logistic modeling, this analysis permits estimation of the within-family odds ratios and the within-family interaction parameter. However, the investigator should recognize that, even without bias from genetic admixture, the within-family odds ratio generally will not be the same as the corresponding odds ratio based on a population-based case-control study. This difference arises because the within-family parameter is subject specific, the “subject” being the family, while the population-based odds ratio is averaged across the population. (The within-family odds ratio has implicitly adjusted out the unknowable baseline risks that vary across families.) Neuhaus et al. have shown that the population-based odds ratio will generally be closer than the former to 1 (9). For similar reasons, a study that uses siblings will be estimating a parameter that is further from the null value than the same parameter estimated by a study in which cousins are used as controls. Meta-analysts should take notice.

Designs in which the parents provide the genetic control offer some practical and analytical advantages, particularly for studying diseases with early onset, such as birth defects. In this design, a sample of unrelated cases and their parents are enrolled, and each threesome of persons is genotyped. The resulting triads (of genotypes) can be assumed to be mutually statistically independent.

Several methods of analysis have also been proposed for case-parent data. Consider a diallelic gene producing three
genetic categories of risk corresponding to whether a person carries zero, one, or two copies of the variant allele. The data can be analyzed by considering the “pseudosiblings” of the affected proband, that is, the three other equally likely potential offspring (with respect to the gene under study) that those two parents could have produced. In the pseudosib method of analysis, first proposed by Self et al. (10), one can use conditional logistic regression; each case is considered matched to the pseudosibs in a 1:3 match. Interaction between genotype and an exposure is assessed multiplicatively by comparing the estimated genetic relative risks among exposed probands with the corresponding genetic relative risks among unexposed probands. Because of the associated computational trick, the case-parent design is sometimes referred to as the pseudosib design (8), although “pseudosib” really refers to a method of analysis and not the design (11). One feature of this approach that has caused some confusion (8, 11) is that what can be estimated is the within-family relative risk, not the within-family odds ratio, despite the fact that the computational gimmick applies conditional logistic software to the pseudosib data.

An alternative approach to the same data explicitly treats the family, that is, the triple of genotypes, as the unit of analysis. One first counts the triads, categorized on the basis of the three genotypes, and then applies Poisson regression. The log-linear model must stratify on the parental genotypes (12, 13), with further possible stratification on the exposure status of the proband (14). The use of parents in this way confers robustness against confounding due to genetic admixture. For effects due to the inherited genes and their multiplicative interactions with exposures, maximum likelihood analysis based on the log-linear Poisson model gives results identical to those based on conditional logistic regression (10) of cases and their pseudosibs. The estimated genetic relative risks and their standard errors produced by using the two methods will be the same (regardless of whether the disease is rare). Moreover, both are also equivalent to the “conditional on parental genotype” likelihood approach advanced by Schaid and Sommer (15, 16). These are not three different models but, in this limited context, should be thought of as three different ways to do the same thing.

The log-linear approach has some important advantages, however, over the explicitly conditional alternatives. Standard Poisson regression software can be used. The log-linear model can be extended easily to allow for effects of the maternal genotype (13), which can be important, for example, in the study of birth defects; in addition, it can be modified to assess imprinting (17). It can also efficiently use incomplete triads by allowing for missing parents (18). (Although their simulations suggest that their approach is less efficient than others, Sun et al. (19, 20) provided closed-form estimators for the genetic relative risks that remain valid when some parents are missing.)

Table 1 summarizes the primary analytical options available for data from a case-parents design. \( R_1 \) and \( R_2 \) denote the relative risks of disease for heterozygous and homozygous carriers of the variant allele, respectively, relative to those with no copies. While seven methods are listed, the first two are not useful for estimation, as has been discussed (14), and the third and fourth would be expected to be less efficient than full maximum likelihood because they discard information from families in which all three persons are heterozygous. The remaining three methods are all equivalent in a setting in which interest centers on the effects of the inherited gene and its multiplicative interactions with exposure. All seven methods shown in table 1 share the same set of assumptions required for valid analysis. It is important to understand what this design can and cannot do in relation to other designs.

ASSUMPTIONS COMPARED

Table 2 compares the case-parents design with sibling-control (or cousin-control) designs and with the traditional population-based case-control design and the case-only design. The major assumption required for validity of the case-parents design is that, conditional on the parental genotypes, the proportions of offspring with the three genotypes follow Mendelian probabilities in the population. In short, the assumption is that the process by which one of the two alleles is selected from each parent for transmission to a particular offspring is stochastically similar to the toss of a coin. This assumption could fail if, for example, an allele that increases risk of a disease were to cause early embryonic death in all embryos that inherit two copies. Then, a case-parents design would always estimate \( R_2 = 0 \), because no homozygous probands would be seen. If the Mendelian assumption fails, all seven methods shown in table 1 are invalid. However, the assumption may often be biologically plausible, and geneticists routinely rely on it.

Independence of the genotype and the exposure, conditional on the parents, is required by the case-parents design for valid estimation of the multiplicative interaction parameters and the exposure-specific genetic relative risks. (The “main effect” of the exposure cannot be estimated by using the case-parents design.) The required independence assumption could fail if, for example, a gene under study influences metabolism of alcohol, and alcohol consumption is the exposure under study. We provide a more detailed discussion of this issue elsewhere (14). Although independence is not required for the two case-control designs, if genotype influences exposure, then interpretation of parameters from the logistic regression analysis can be problematic (21).

The two key additional limitations imposed by the case-parents design are its inability to assess nonmultiplicative models and its inability to estimate exposure effects. Another practical limitation is that the case-parents design is most useful when many cases’ parents will still be living, such as for studies of diseases such as juvenile-onset diabetes and birth defects.

The inability of the case-parents design to estimate exposure effects is perhaps its most important limitation. With such a study, one can only estimate effects of genotype in the absence and presence of an exposure (or at multiple levels of exposure), but such estimates can be difficult to interpret without an estimate of the main effect of the exposure. Thus, for example, one would not be able to distinguish a scenario in which an exposure confers excess risk but only
in the presence of a mutant allele from a scenario in which the exposure confers protection but only in the absence of the allele. Such distinctions can be key to understanding the biology of the joint effects of the allele and the exposure.

The case-sibling design has a different set of limitations, many of which are listed in table 2. One assumption not listed in table 2 is required when case-sibling data are analyzed with conditional logistic regression (and some families contribute more than two persons). In that setting, the model requires that the outcomes be conditionally independent within families. If the locus under study is in linkage with a locus that influences risk, and if some probands have been matched to more than one related control, then the usual conditional logistic model will not fully capture the dependency structure in sibship- or cousin-based data; hence, conditional logistic regression is not strictly statistically valid. Other methods have been proposed that avoid the dependency issue, for example, the sign-test-inspired testing method of Horvath and Laird (22), which does not fully use the data nor permit estimation of relative risk parameters. Another method for avoiding the problem of within-family dependency repeatedly samples case-control pairs from each informative sibship then carries out a matched-pairs logistic analysis, which relies only on the statistical independence (across clusters) of the paired differences. The estimates resulting from a large number of such resamplings are then averaged, and a corresponding adjusted variance implicitly handles the dependency among the resampled data sets (23). The dependency-structure problem is not listed in table 2 because it is not a feature of the design but rather arises when conditional logistic regression is the selected analytical method.

Some important, more practical concerns also arise when siblings are used as controls, some of which were mentioned by Witte et al. (8). One problem with use of unaffected siblings is that some may really be “affected” but just not have lived long enough to express the disease phenotype. Witte et al. handle this issue, which they describe in terms of risk-set sampling, by requiring the unaffected sibling to be older than the proband. (This approach presumes that incident

<table>
<thead>
<tr>
<th>Method (reference no.)</th>
<th>Strategy</th>
<th>Estimable parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transmission disequilibrium test (28)</td>
<td>Test whether the observed rate at which heterozygous parents transmitted the variant allele to affected offspring is statistically consistent with 0.5</td>
<td>None, unless ( R_1 = R_2 ), in which event ( R_1 ) and ( R_2 ) can be estimated*</td>
</tr>
<tr>
<td>Harley et al. (29) and Maestri et al. (30)</td>
<td>Logistic model for transmission of variant allele to affected offspring from heterozygous parents</td>
<td>Not valid, unless ( R_1 = R_2 ) (13), in which case ( R_1 ) and ( R_2 ) can be estimated</td>
</tr>
<tr>
<td>Sun et al. (19, 20)</td>
<td>Closed-form estimators for ( R_1 ) and ( R_2 ) and hypothesis testing if only 1 parent, under assumptions</td>
<td>( R_1 ) and ( R_2 )</td>
</tr>
<tr>
<td>Pseudo-complementary: sib made up of the 2 non-transmitted parental alleles (31, 32)</td>
<td>Conditional logistic regression based on 1:1 matching; triads where all 3 are heterozygous are treated as not informative†</td>
<td>( R_1 ), ( R_2 ), and gene-by-exposure interactions</td>
</tr>
<tr>
<td>Self et al. (10), “pseudo-sib” method</td>
<td>Conditional logistic regression based on comparing the proband with the 3 other hypothetical offspring their parents could equally likely have had</td>
<td>( R_1 ), ( R_2 ), and gene-by-exposure interactions</td>
</tr>
<tr>
<td>Maximum likelihood conditional on parental genotypes (15, 16)</td>
<td>Maximize a conditional likelihood or test via a score statistic</td>
<td>( R_1 ), ( R_2 ), and gene-by-exposure interactions</td>
</tr>
<tr>
<td>Maximum likelihood via log-linear model (12–14)</td>
<td>Poisson regression of counts of triads with stratification on parental genotypes (some parents can be missing)</td>
<td>( R_1 ), ( R_2 ), maternal effects, imprinting effects, and gene-by-exposure interactions</td>
</tr>
</tbody>
</table>

* \( R_1 \) and \( R_2 \) denote the relative risk for persons who carry 1 or 2 copies of the variant allele, respectively, relative to a person who carries no copies.

† The conditional logistic regression approach mentioned here is not in the literature as far as we know, but we propose it as an efficient and robust (against admixture) approach among those based on such pairs.
cases are studied immediately. A slightly less-restrictive design might require the unaffected sibling to be at least as old as the case was at the time of diagnosis.) The requirement that a sibling of a suitable age exists and can be studied can substantially reduce the number of eligible cases, as occurred in the California breast cancer study cited (8), in which only a quarter of cases had an eligible sibling control. Witte et al. propose to fix this problem by advising the investigator to enroll a cousin when a suitable sibling is not available. As they acknowledge, this strategy may improve sample size but weakens the design’s protection against genetic admixture. Moreover, use of either siblings or siblings supplemented by cousins could result in sampling in which the cases and controls are not comparable with respect to factors such as region of residence.

Correlation in exposure status can cause overmatching and consequent loss of efficiency when siblings are used (8), but bias also may be produced by asymmetric structure in the exposures of older versus younger offspring. If birth-order effects exist, then any asymmetry in eligibility criteria for cases versus controls could produce bias. For example, birth order is evidently related to the likelihood that a person becomes a smoker, with the younger siblings being at increased risk (24). Such asymmetry could result in an apparent effect of smoking on risk, whether one exists or not.

**TABLE 2. Sets of related assumptions required by retrospective designs for studying genetic and environmental effects**

<table>
<thead>
<tr>
<th>Case-only design</th>
<th>Case-parents design</th>
<th>Case-sibling (or-cousin) design</th>
<th>Population-based case-control design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recruitability can be differential by exposure or by genotype but not by the E/G* combination in a nonmultiplicative way.</td>
<td>Recruitability is independent of genotype, given parents. (Recruitability can be differential by exposure.)</td>
<td>Recruitability is not differentially related to genotype/exposure for cases versus controls, given parents.</td>
<td>Recruitability is not differentially related to genotype/exposure for cases versus controls.</td>
</tr>
<tr>
<td>Interactions of interest are ratios of relative risks.</td>
<td>Relative risks (within-family) are of interest; odds ratios are not needed; interactions of interest are ratios of relative risks.</td>
<td>Odds ratios (within-family) are of interest; relative risks are not needed.</td>
<td>Odds ratios (population-based) are of interest; relative risks are not needed.</td>
</tr>
<tr>
<td>Gene and exposure occur independently in the population under study, at the ages of cases.</td>
<td>Genotype and exposure are independent given parents (gene does not influence exposure).</td>
<td>Bias from genetic admixture is negligible.</td>
<td></td>
</tr>
<tr>
<td>Genetic admixture has not induced a correlation between the exposure and the gene under study (or the independence assumption given above will be violated).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No maternally mediated interaction effects on offspring risk.</td>
<td>Mendelian assortment; also; survival from fertilization to age at study does not depend on the gene, given the parents.</td>
<td>No maternally mediated interaction or genetic effects on offspring risk.</td>
<td></td>
</tr>
<tr>
<td>Survival from fertilization to age at study does not depend jointly on exposure and genotype (or the independence assumption given above will be violated).</td>
<td>Biologic parents are available for study.</td>
<td>Cases randomly sampled from eligibles; controls sampled from the population producing the cases.</td>
<td></td>
</tr>
<tr>
<td>Availability of genotyped cases is not differential by genotype and exposure jointly.</td>
<td>Proband and sibling control(s) do not systematically differ in ways that could be related to genotype and/or exposure (e.g., via birth-order effects on exposure).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No need to estimate effects of exposure or of genotype per se.</td>
<td>Biologic parents are available for study.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiplicative model for interaction.</td>
<td>&quot;Siblings&quot; have the same mother and father (can be verified).</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
* E, exposure; G, genetic factor.
Other temporal trends can also introduce bias. For example, suppose a birth defect is under study and the exposure of interest is maternal smoking. If two babies born to the same woman are discordant with respect to maternal smoking, it may be more likely that the more recent gestation was the smoke-free one. The point is that women may be more likely to quit smoking before to take up smoking between pregnancies. Such a temporal pattern can produce a spurious protective effect of maternal smoking if the case is matched more often to an older sibling. Of course, with birth defects (being prevalent events at birth), there is no good reason to require the control sibling to be older. (On the other hand, if incident birth defects are studied, there will be no younger siblings to sample!) Even for diseases with later onset, such as leukemia, the same kind of bias can arise if one wishes to consider early life exposures, such as parental smoking. In general, young couples may experience systematic trends over time that differentially affect their children, such as increasing affluence and increasing exposure to infectious agents with increasing numbers of children. All such temporal trends can produce bias if one requires the sibling control to have been born before the proband.

Clearly, despite having certain practical and theoretical advantages, both the case-parents design and the sibling-control design suffer from important limitations. On the other hand, family members are often well motivated, and the sampling assumptions required for a family-based design are weaker than those required for the corresponding population-based case-control study.

In practice, the optimal family-based design may be a hybrid that genotypes all available first-degree relatives; then the analysis takes all available information into account. Procedures have been proposed for testing for genetic effects based on such hybrid data (18, 20, 25), as have methods for estimating genetic relative risks (18). However, to our knowledge, a comprehensive analytical approach that would allow full exploitation of data from family members in such a setting has not been set forth.

CONCLUSIONS

We have focused attention on conditions and assumptions that must be fulfilled so that particular retrospective designs can be epidemiologically informative. In practice, the investigator will also want to consider the relative costs of finding, recruiting, genotyping, and interviewing study subjects and weigh these costs against the statistical efficiency and protection from bias (the information benefit) offered by the various approaches. Consequently, we cannot recommend one of the four designs as universally superior to the others. If the investigator already has access to archived tissue specimens from cases, then a case-only study becomes cost effective. If parents are readily available and highly motivated, as might be expected when a birth defect is being studied, the case-parents design becomes advantageous. If one is studying a late-onset disease in a population that is genetically admixed but residentially stable, then a sibling-control design can make good sense. If the population under study is either genetically homogeneous or genetically diverse enough, and healthy people are approachable and willing to be genotyped, then a case-control design works well.

Our assessment of problems associated with family-based designs should be tempered by the recognition that the population-based case-control design is also not problem free. There is, again, the issue of possible confounding due to incomplete genetic mixing of sectors of the population under study, which can produce bias in assessing both genetic effects and gene-by-environment interactions. Family-based studies avoid this potential source of bias by making within-family inferences. Another issue not always appreciated is that the maternal genotype can serve as a confounder. Genetic effects can take two forms: the inherited genotype can affect the offspring’s risk (the usual mechanism assumed), or the mother’s genotype can influence the offspring’s prenatal environment (13). If the genetic effect is fully or partly mediated by the mother’s genotype, a case-control study could to some extent misattribute such effects to the inherited genotype and also estimate with bias the interaction between genotype and exposure. Hence, when using a case-control design, one needs to assume that no maternally mediated genetic effects exist (table 2) or recognize that the two cannot be fully separated.

Another problem that can potentially impact a population-based case-control study is that random sampling of suitable controls from a defined population may become increasingly difficult for both practical and cultural reasons (26). Although data are hard to find on trends over time in recruitment rates for controls, a trend toward decreasing participation among healthy potential controls could be anticipated. Many people use telephone answering machines to avoid unwanted solicitations. Although leaving a message on the answering machine has been shown to improve response rates (27), such machines do present a barrier that could become increasingly difficult to surmount, particularly if telemarketing continues to expand. Moreover, with recent press attention focused on privacy and the abuse of genetic and other types of information, aversion to genetic testing may grow. In the face of these obstacles, the key assumption that recruitability in relation to genotype and exposure is not differential by case status may become less tenable. If healthy people increasingly resist recruitment as controls into epidemiologic studies, family-based designs may become increasingly important tools in the 21st century for the epidemiologist who wishes to study joint effects of genes and exposures.

ACKNOWLEDGMENTS

The authors thank Drs. Norman Kaplan, Stephanie London, Allen Wilcox, and Glinda Cooper for their helpful suggestions.

REFERENCES

1. Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and