Family-Based Association Studies

W. James Gauderman, John S. Witte, Duncan C. Thomas

We review case–control designs for studying gene associations in which relatives of case patients are used as control subjects. These designs have the advantage that they avoid the problem of population stratification that can lead to spurious associations with noncausal genes. We focus on designs that use sibling, cousin, or pseudosibling controls, the latter formed as the set of genotypes not transmitted to the case from his or her parents. We describe a common conditional likelihood framework for use in analyzing data from any of these designs and review what is known about the validity of the various design and analysis combinations for estimating the genetic relative risk. We also present comparisons of efficiency for each of the family-based designs relative to the standard population-control design in which unrelated controls are selected from the source population of cases. Because of overmatching on genotype, the use of sibling controls leads to estimates of genetic relative risk that are approximately half as efficient as those obtained with the use of population controls, while relative efficiency for cousin controls is approximately 90%. However, we find that, for a rare gene, the sibling-control design can lead to improved efficiency for estimating a $G \times E$ interaction effect. We also review some restricted designs that can substantially improve efficiency, e.g., restriction of the sample to case-sibling pairs with an affected parent. We conclude that family-based case–control studies are an attractive alternative to population-based case–control designs using unrelated control subjects. [Monogr Natl Cancer Inst 1999;26:31–7]

Association studies are routinely used by epidemiologists to investigate the relationship between an exposure and a disease. With the recent increase in the availability of genetic information, these exposures may now include genotypes at one or more susceptibility, candidate, or marker loci. The goals of genetic association studies will differ, depending on the state of knowledge about the given disease. For example, once a susceptibility locus has been cloned (e.g., BRCA1 for breast cancer), the goals include estimating the relative risk (RR) and penetrance associated with specific mutations and testing for interaction with environmental exposures or other genes *(1)*. If a candidate locus has been identified (e.g., the androgen receptor for prostate cancer), the primary goal is testing the null hypothesis of no association between the locus and the disease. Finally, if little is known about specific loci for the disease (e.g., multiple sclerosis), multiple tests of association with finely spaced markers may be used to screen the genome for candidate regions in the hopes of detecting linkage disequilibrium with markers close to one or more disease loci.

The case–control design is generally considered the design of choice for studying rare diseases, although suitably designed cohort studies, particularly family-based cohort studies *(2)*, are also useful in some circumstances. For results to be generalizable, the selection of case patients in a case–control study should be population based. This process is relatively straightforward for diseases like cancer, for which population-based registries are available, provided one can identify cases rapidly enough to enroll them and obtain blood samples. For effect estimates and hypothesis tests to be valid, control subjects should be selected from the same source population as the cases. In the situation of disorders with a genetic basis, this implies that cases and controls should derive from a similar genetic background.

One approach used to satisfy this requirement is to match cases and controls on their race or ethnicity. However, even within subgroups, strong variation can be found in allele frequencies at many genetic loci (e.g., the gradient in human leukocyte antigen allele frequencies from northern to southern Europeans). An additional complication is that, in many places, a given subject may represent a mixture of genetic backgrounds as a result of intermarriage between ancestors of varied ethnicities, and, as a practical issue, many subjects will not know with certainty their complete ancestral background. This uncertainty makes finding a suitable population-based control for such subjects very difficult. If the allele frequency at a particular genetic locus varies across ethnic groups and if ethnicity (or some unobserved factor that varies by ethnicity) is a risk factor for disease independent of that locus, then failure to adequately control for ethnicity can result in false associations between the gene and the disease *(3–5)*. This phenomenon is often referred to as population stratification by geneticists and as confounding by epidemiologists. The unobserved ethnic factor associated with disease can be either another gene or an environmental factor. An example of such confounding is the reported association between the Gm locus and non-insulin-dependent diabetes mellitus (NIDDM) in American Indians that disappeared when the analysis was restricted to full-heritage Pima-Papago Indians *(6)*. The likely explanation for this finding was that the Gm locus served as a surrogate for Caucasian heritage and that the risk of NIDDM varied with this level of ancestry.

Recently, much interest has been focused on the use of family-based controls to avoid the problem of ethnic confounding. One approach is to match each case with one or more unaffected siblings *(7,8)* or cousins *(8)* and to use analytic techniques for matched case–control studies *(9)* to estimate effects and to test hypotheses. A second approach is to match each case to a set of "pseudosiblings," formed as the set of genotypes that was not transmitted from the parents to the case. Several methods have been proposed for testing candidate gene associations and for estimating genetic RRs, including the transmission disequilibrium test (TDT), conditional logistic regression, and haplotype-sharing techniques *(4,10–19)*. Both the sib-control and pseudo-

Affiliations of authors: W. J. Gauderman, D. C. Thomas, Department of Preventive Medicine, University of Southern California, Los Angeles; J. S. Witte, Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH.

Correspondence to: W. James Gauderman, Ph.D., Department of Preventive Medicine, University of Southern California, 1540 Alcazar St., Suite 220, Los Angeles, CA 90089 (e-mail: jimg@rcf.usc.edu).

See "Note" following "References."

© Oxford University Press

Downloaded from https://academic.oup.com/jncimono/article/1999/26/31/908209 by guest on 21 August 2022

sib-control approaches have the advantage that they provide perfect matching on ethnicity.

We review the basic family-based case–control designs, describe a general approach to examine their validity and efficiency, and summarize what is known about their relative efficiency for estimating the genetic RR. We also describe variations on currently proposed designs that may be useful in some circumstances.

DESIGNS

For a disease like cancer with variable age of onset, we consider the genetic RR parameter of interest to be the ratio of age-specific incidence rates (i.e., the hazard rate ratio). With this choice, the odds ratio from any matched case–control design is a consistent estimator of the RR, provided controls are randomly selected from the "risk set" comprising those members of the population at risk who were disease free at the age at which the case was affected. Indeed, exclusion of subjects who later developed the disease of interest will bias odds ratio estimates away from the null *(20).* There is then no need for a rare disease assumption *(21).* Control subjects should also be matched to case patients on any potential confounders and generally should be matched on sex (particularly for sex-specific diseases).

Sib Controls

Instead of defining the source population as the entire population, one could consider only the immediate or more distant family members of the case as potential controls, leading to the various designs considered here. For example, in the sib-matched case–control design, the investigator matches each case patient to one or more unaffected sibling controls. The principles of risk-set sampling require that controls have attained the age of the case and still be disease free. If only recently incident cases are included, this criteria essentially restricts control selection to older siblings. Of course, a sibling who is younger than the case may achieve the case's age of diagnosis during the study period and then become eligible as a control. Use of siblings who have not yet attained the age of the case may lead to effect estimates that are biased away from the null, but this bias could theoretically be corrected with the use of knowledge of the population rates. Inclusion of such siblings would also pose problems if time-dependent covariates are involved.

Although genotypes do not change with age, a restriction to younger sibs could lead to confounding of the effects of environmental exposures that have secular trends (e.g., oral contraceptive use) and conceivably confounding of the effects of any genotypes with which such risk factors were associated. As in any case–control study of time-dependent factors, the exposure status of cases and controls should be compared at a common "reference age," such as the case's age at diagnosis (or some common interval prior to it to allow for latency). In addition to being perfectly matched on ethnicity, siblings will also be matched on many other potential confounding variables. Although this match offers protection from bias, siblings are likely to be overmatched on many factors (including genotype) that will result in less efficient parameter estimation. This situation will be explored quantitatively below.

From a practical standpoint, the use of sibling controls may offer several nonstatistical advantages over population controls. The occurrence of disease in the case may make his or her relatives much more willing to participate than an unrelated subject from the general population. In addition to reducing cost, this willingness may result in the control being more careful in filling out a risk-factor questionnaire. Because the case and sibling will share many exposures, researchers will be able to cross-validate questionnaire information that has been obtained from the case and control, such as the types of cancer in their ancestors, or to ask comparative questions, such as which of the two was more exposed to particular environmental factors. Many groups have, or are in the process of collecting, family-based cancer data resources. For example, the Cancer Surveillance Project for Orange and San Diego Counties routinely abstracts family history information on first- and second-degree relatives and first cousins of all cases; this resource is the basis of a population-based family study of breast and ovarian cancers involving a family-history stratified sample of cases *(22).* Once a resource such as this project has been established, selection of sibling controls can be much less expensive than finding controls from the general population. Conversely, not all cases will have an eligible and willing sibling available; in addition to the obvious loss of sample size, it is possible that selection bias could arise if availability of a sib control were related both to disease risk and to allele frequency.

Cousin Controls

Instead of a sibling, one could obtain another relative of the case as a control. If one is also studying risk factors of which distribution varies by generation, controls should probably be drawn from the same generation, such as first cousins. Compared with a sibling control, the advantage of a cousin is that one may be able to obtain closer matching on age and year of birth, with less loss in efficiency because the case and cousin are not as closely matched on genotype. The trade-off is that there is no longer the absolute protection from ethnic confounding because the case and cousin have only one side of their families in common and there is no guarantee that the two unrelated parents of the case and cousin derive from the same ethnic background. In this circumstance, one might want to select two cousin controls, one from each side of the family, but it remains to be shown that this will provide adequate protection from bias. From a practical standpoint, cousin controls have many of the same advantages as sibling controls, including increased willingness to participate and possible pre-identification through a family-based data resource. As in the sibling design, not all cases will have an eligible and willing cousin control, but there is generally a larger pool from which to choose.

Pseudosibling Controls

In this design, no actual controls are selected. Instead, genotypic data are obtained on the parents of the case, and the genotype transmitted to the case is then compared with the three genotypes (pseudosiblings) that were not transmitted to the case. Suppose we label the alleles of the two parents (a,b) and (c,d) and the case's alleles as (a,c) (recognizing that some of these alleles may be identical by state). Then the three pseudosibling genotypes are (a,d), (b,c), and (b,d) and the question that this design seeks to address is whether a specific allele or genotype occurs more commonly in cases than in their pseudosibs. Conditional logistic regression for 1:3 matched case–control studies is the appropriate analysis for such data *(11).* The TDT, which simply compares the case with his or her "antisib" (b,d), has been shown to be the score test from the conditional logistic

Downloaded from https://academic.oup.com/jncimono/article/1999/26/31/908209 by guest on 21 August 2022

likelihood under a multiplicative model for dominance, in which the homozygote RR is the square of the heterozygote RR *(4,16,19).*

Both the sib-matched case–control and the pseudosib (or TDT) designs test the same null hypothesis, i.e., that there is no association or no linkage between the candidate gene (or marker) and the underlying trait gene. Thus, neither design will detect association with a gene that is in disequilibrium with a causal gene (e.g., because of population stratification) but that is not linked to that causal gene. Essentially, sib controls can be thought of as a finite realization of genotypes from the theoretical distribution of pseudosib genotypes, with the only fundamental difference being that real sibs are required to have survived to the age of the case. The lack of this restriction in the pseudosib design produces an estimator of the genetic RR that is biased toward the null by an amount that disappears with increasing disease rarity, although the validity of the hypothesis test is not affected *(8).*

As with the sibling- and cousin-control designs, parents are more likely to be willing to participate than a population control, and the design will take advantage of existing information available in a family-based data resource. In practice, the utility of this design is limited to disorders that occur at young enough ages that parents of the case are still likely to be alive. This limitation excludes many cancers, unless the focus is on younger onset cases. It has been shown that, if the genotype is missing on one parent, there can be bias in the TDT *(23).* An alternative approach when parental data are missing is to use the sib-TDT *(24),* which involves a comparison of the genotype of an affected sibling to that in an unaffected sibling, and is similar to the sibling-control approach described above. One can also combine the TDT and sib-TDT, using parental genotypes if they are available and siblings if they are not *(24).* However, if there are multiple affected subjects (or multiple unaffected siblings for the sib-TDT), the TDT and sib-TDT provide only a valid test of linkage; the test of association will have an inflated type I error rate.

Restricted Designs

For diseases that are not too rare, one might consider any of the above designs with an additional restriction to subjects with a positive family history. The rationale would be to increase the allele frequency in the sample, thereby improving the statistical efficiency for detecting associations with rare genes. However, care must be taken that any restriction applied to cases is applied equally to controls. For example, if one required the case to have an affected first-degree relative, one would have to make the same requirement for controls. For a design with population controls, this requirement might entail some form of multistage sampling *(25,26),* in which one obtains family history information on an unrestricted series of potential cases and controls and then selects a subsample of those with a positive history. Sib controls are automatically matched on family history (among sibs, parents, and more distant relatives, but not their offspring), and such sibships might be easily identified from registries that contain family history data. Cousin controls with comparable family histories are not as easily identified, although case–cousin pairs that share an affected grandparent would be a valid comparison, as would those that each have an affected sibling. However, case–cousin pairs that each have an affected parent would be a valid comparison only if the two parents were related to each other or if neither parent was a relative of the other pair member. Such case–cousin pairs with two affected relatives would generally be quite uncommon.

Comparison of Designs

We now describe our basic approach to comparing the validity and relative efficiency for estimating the genetic RR for various family-control and population-control designs.

We assume that the data consist of diseased subjects (cases) and one or more matched controls (real or pseudosiblings) for each case. Let d_{ij} denote the disease status of subject j in matched set i, and let g_{ij} denote the genotype at some locus of interest. For simplicity, we assume that the alleles at the locus can be classified as either mutant (denoted by A) or normal (denoted by a), with population frequency q of the A allele, although the methods are easily extended to genes with more than two alleles. Let $G(g)$ denote a genetic covariate with values $G(g) = 0$ when $g = aa$, $G(g) = 1$ when $g = AA$, and $G(g) = \Delta$ when $g = Aa$. The parameter Δ is coded to reflect an assumed mode of inheritance, with $\Delta = 1$ corresponding to dominant inheritance, $\Delta = 0$ to recessive inheritance, and $\Delta = 0.5$ to multiplicative (or log-additive) inheritance; this parameter can also be estimated in a general codominant model.

For a binary trait, we assume a logistic model for penetrance, i.e.,

$$\text{logit}[Pr(d = 1|g)] = \alpha_i + \beta G(g), \qquad [1]$$

where α_i denotes the logit of the baseline risk for noncarriers in matched set i, and β is the log-RR for carriers of a mutation. For matched pairs data, the conditional likelihood is a function of only β, which we assume is common across matched pairs. If β were variable across the population, then a study would estimate some form of weighted average of the distribution of β values, the particular weighting being somewhat different for population-control versus family-control designs. In the family-control designs, we assume that the disease outcomes among relatives are conditionally independent, given their genotypes. Letting g_{i1} denote the genotype of the case in the i^{th} matched pair, the conditional logistic likelihood is

$$L(\beta) = \prod_i \frac{\exp(\beta G(g_{i1}))}{\sum_{j \in M_i} \exp(\beta G(g_{ij}))}, \qquad [2]$$

where M_i denotes the set of subjects in the i^{th} case–control set. For the case–pseudosib design, j ranges over the case and the three pseudosibs.

For a disease of variable age at onset, essentially the same likelihood can be derived from Cox's proportional hazards model,

$$\lambda(t,g) = \lambda_0(t) \exp(\beta G(g)),$$

where $\lambda(t,g)$ denotes the genotype-specific incidence rate at age t and $\lambda_0(t)$ denotes an unspecified set of baseline rates in noncarriers. Equation 2 then results when the controls are drawn at random from the risk set for the i^{th} case.

The models above can be expanded to include one or more environmental covariates (z) and gene–environment interaction terms. In this case, the logistic model becomes

$$\text{logit}[Pr(d = 1|g,z)] = \alpha_i + \beta G(g) + \gamma z + \eta G(g)z,$$

Downloaded from https://academic.oup.com/jncimono/article/1999/26/31/908209 by guest on 21 August 2022

with an analogous extension to Cox's proportional hazards model. For either model, the conditional likelihood is

$$L(\beta, \gamma, \eta) = \prod_i \frac{\exp[\beta G(g_{i1}) + \gamma z_{i1} + \eta G(g_{i1}) z_{i1}]}{\sum_{j \in M_i} \exp[\beta G(g_{ij}) + \gamma z_{ij} + \eta G(g_{ij}) z_{ij}]}.$$

In the pseudosibling design, z_{ij} is set equal to z_{i1} for all j, precluding estimation of the environmental main effect parameter (γ) and requiring an assumption of independence of the genetic and environmental factors conditional on parental genotypes for valid estimation of η.

To assess the validity of a design or analysis combination for estimation of β, we computed the expectation of the score statistic (the first derivative with respect to β of the log likelihood) under the true model. If this expectation is zero, then the estimator is said to be Fisher consistent, meaning that the maximum likelihood estimate will converge to the true value with increasing sample size. In this case, the asymptotic relative efficiency (ARE) for estimating β for one design compared with another is defined as the inverse of the ratio of their expected variances of $\hat{\beta}$ under the alternative hypothesis or equivalently as the ratio of the sample size required to attain the same precision and power. We compute the expected variance of $\hat{\beta}$ as the inverse of the Fisher information, evaluated at the true value of the parameters (α_0, β_0, q_0). For comparability across several parameter values, we fixed the population prevalence of the disease (K_p) and the attributable risk (AR) and then, for given values of the log-RR (β_0), solved the following two equations for α_0 and q_0:

$$AR = \frac{\sum_g (e^{\beta_0 G(g)} - 1)\, Pr(g|q_0)}{\sum_g (e^{\beta_0 G(g)})\, Pr(g|q_0)}$$

and

$$K_p = \sum_g Pr(d = 1|G(g), \alpha_0, \beta_0)\, Pr(g|q_0).$$

The factor $Pr(g|q_0)$ was computed assuming Hardy–Weinberg equilibrium, and the penetrance factor in the equation for K_p was computed as the anti-logit of the expression in equation 1. Letting *Rel* denote the relationship between the case and the control and assuming a 1 : 1 matched design, the Fisher information was computed as

$$E[I(\beta|Rel)] = \sum_{\boldsymbol{g}} I(\beta|\boldsymbol{g})\, Pr(\boldsymbol{g}|d_1 = 1, d_2 = 0, \alpha_0, \beta_0, q_0, Rel)$$
$$= \frac{\sum_{\boldsymbol{g}} I(\beta|\boldsymbol{g})\, Pr(d_1 = 1|g_1)\, Pr(d_2 = 0|g_2)\, Pr(\boldsymbol{g}|Rel, q_0)}{\sum_{\boldsymbol{g}} Pr(d_1 = 1|g_1)\, Pr(d_2 = 0|g_2)\, Pr(\boldsymbol{g}|Rel, q_0)}, \quad [3]$$

where $\boldsymbol{g} = (g_1, g_2)$ and $I(\beta)$ is the observed information, i.e. the negative of the matrix of second partial derivatives of the conditional log-likelihood. One can see from equation 3 that the joint distribution of the case and control genotypes is the factor that differentiates the informativeness of the various designs. If the case and control are unrelated, $Pr(\boldsymbol{g}|Rel, q_0) = Pr(g_1|q_0)Pr(g_2|q_0)$, and the weight is determined solely by the allele frequency. However, if the case and control are siblings, $Pr(\boldsymbol{g}|Rel, q_0) = \sum_{g_f, g_m} Pr(g_1|g_f, g_m)Pr(g_2|g_f, g_m)Pr(g_f|q_0)Pr(g_m|q_0)$ and the weight is a function of both the allele frequency and the genetic relationship between the pair. Note that, although computation of the expected information for the sib-matched design involves a summation over parental genotypes (g_f, g_m), the actual information depends only on the joint distribution of the genotypes of the case and sibling control.

RESULTS

In the presence of population stratification, the amount of bias in the estimate of genetic RR when using the population-based case–control design depends on the true RR and the ratio of allele frequencies in the strata *(8)*. The sib-control design is always consistent, the pseudosib design is approximately consistent for a rare disease but inconsistent for a common disease, and the consistency of the cousin design will depend on whether the unrelated parents of the cousins come from the same or different population strata. The bias in the pseudosib design for a common disease occurs even in the absence of population stratification, although a method has been proposed for correcting this bias *(8)*.

Fig. 1 provides a summary of the ARE of the three basic family designs for estimating β, relative to case–control studies using unrelated controls. The results are based on a disease with population prevalence K_p = 1% and AR = 5%, although the relative efficiencies are not substantially affected by these two parameters. Under a dominant model, the ARE is approximately 50% using sib controls, 88% using cousin controls, and 100% using pseudosib controls, regardless of the true underlying value of the genetic RR. Under a multiplicative model, these three AREs are nearly identical to those for the dominant model (data not shown). For the recessive model, the relative efficiencies are higher than for the dominant model in all three designs. As the genetic RR ranges from 2 to 20, the AREs range from 66% to 72% using sib controls, from 95% to 99% using cousin controls, and from 150% to 260% using pseudosib controls. Although on a per-case basis the pseudosib design is statistically more efficient than unrelated controls for a recessive gene, this design requires three genotypes per case rather than two and so may be less cost efficient if the cost of genotyping is high in relation to the cost of subject enrollment.

To provide some intuition to account for these efficiency comparisons, Table 1 provides the expected number of case–control pairs in the unrelated-control and sib-control designs for a dominant and a recessive gene and a particular choice of parameter values. No population stratification is assumed, and, in all the designs shown, the McNemar odds ratio (*c*/*b*) provides a good approximation to the assumed RR of 20. Compared with the unrelated-control design, use of the sibling-control design results in a larger proportion of genotype-concordant pairs (cells *a* and *d*) because of overmatching, and thus a smaller number of discordant pairs (*b* and *c*) on which the variance of $\hat{\beta}$ is determined. For relatively rare genes, it is evident that the primary determinant of this variance is the number *b* of case–noncarrier and control–carrier pairs.

Similar comparisons of relative efficiency for estimating the gene–environment interaction parameter η have also been carried out *(8)*. The efficiency for a particular design in this case depends on the distribution of the three types of discordant pairs: 1) genotype concordant and exposure discordant, 2) genotype discordant and exposure concordant, and 3) jointly discordant.

Downloaded from https://academic.oup.com/jncimono/article/1999/26/31/908209 by guest on 21 August 2022

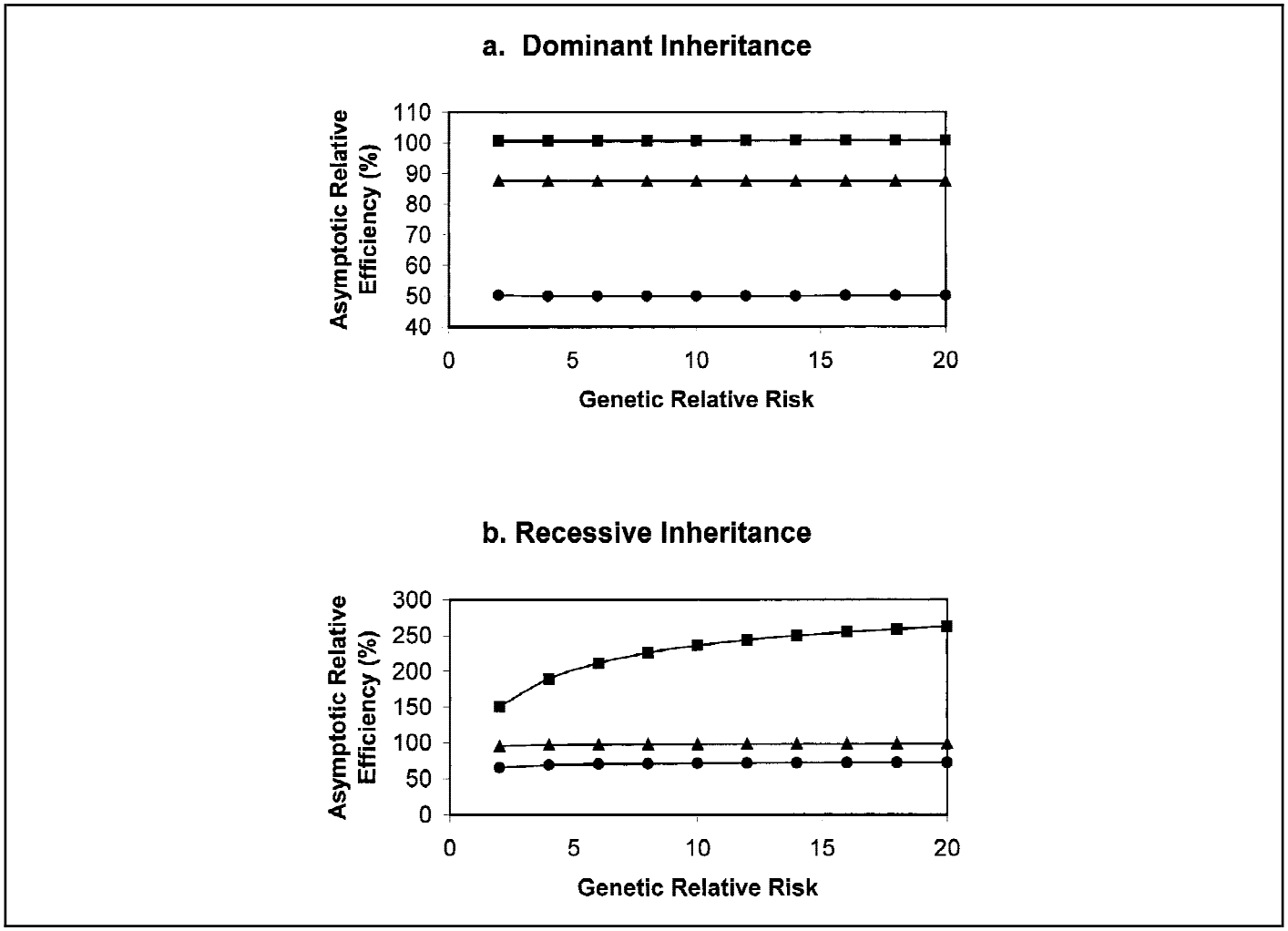

Fig. 1. Asymptotic relative efficiency versus genetic relative risk for the case–sibling (**circle**), case–cousin (**triangle**), and pseudosibling (**square**) designs, relative to the unrelated-control design, assuming disease prevalence of 1%, attributable risk of 5%, and either dominant (**a**) or recessive (**b**) inheritance.

Table 1. Illustrative example of the expected case–control genotype distributions for the unrelated-control and sib-control designs, under dominant and recessive models*

Model	Control type	Control genotype	Case genotype: Carrier	Case genotype: Noncarrier	$\mathrm{Var}(\hat{\beta}) = 1/b + 1/c$	ARE $= V_1/V_2$, %
Dominant	1) Unrelated	Carrier	$a = 46$	$b = 129$		
		Noncarrier	$c = 2576$	$d = 7249$	0.00814	
	2) Sibling	Carrier	$a = 1188$	$b = 66$		
		Noncarrier	$c = 1311$	$d = 7435$	0.01591	51
Recessive	1) Unrelated	Carrier	$a = 45$	$b = 127$		
		Noncarrier	$c = 2546$	$d = 7282$	0.00826	
	2) Sibling	Carrier	$a = 745$	$b = 88$		
		Noncarrier	$c = 1770$	$d = 7397$	0.01193	69

*Expected distributions were computed assuming the population disease prevalence = 1%, relative risk = 20, allele frequency = 0.14 (recessive) or 0.01 (dominant), and 10 000 case–control pairs in each design. ARE = asymptomatic relative efficiency.

For a rare gene, the use of sib controls can be substantially more efficient than the use of population controls for estimating a $G \times E$ effect. The reason is that, when the gene is rare, efficiency is determined primarily by the number of genotype concordant, exposure discordant pairs, although the other two types of discordant pairs also contribute (Table 2). Because of the partial matching on genotype, genotype concordant pairs are more common within sibships than within a case and unrelated control pair, leading to the improved efficiency. However, if sibs are also highly concordant for environmental exposure, this situation will tend to reduce their efficiency relative to unrelated controls.

In contrast to the basic designs, the relative efficiency of the restricted designs for estimating β depends strongly on the genetic RR and the AR but depends only weakly on the population prevalence of disease. Table 3 compares several restricted de-

Downloaded from https://academic.oup.com/jncimono/article/1999/26/31/908209 by guest on 21 August 2022

Table 2. Illustrative example of the expected case–control genotype × environment distributions under a dominant model*

Genotype†	Environment‡	Unrelated control subjects	Sibling control subjects
C–C	E–U	155	418
C–C	U–E	207	1578
C–N	E–E	44	34
N–C	E–E	883	688
C–N	E–U	883	1124
N–C	U–E	66	30

*Expected distributions were computed assuming genetic relative risk = 10, environmental relative risk = 2, interaction relative risk = 2, allele frequency = 0.1, exposure prevalence = 0.25, and sibling exposure concordance odds ratio = 2.

†Genotype of the case–genotype of the control (C = carrier, N = noncarrier).

‡Exposure status in the case–exposure status in the control (E = exposed; U = unexposed).

signs for $K_p = 1\%$, several modes of inheritance, and a range of relative and ARs. Generally, the efficiency gains are greatest for genes with low AR and large RR, i.e., for rare major susceptibility genes. Across inheritance modes, efficiency gains in the restricted designs are greatest for a dominant gene. In the sibling-control design, restriction to pairs with an affected parent (design SAP) substantially improves efficiency for a dominant gene, whereas, if the restriction is to pairs with an additional affected sibling (design SAS), one can expect substantial efficiency gains for either a recessive or a dominant gene.

Absolute power can be computed with the use of standard methods once the expected distribution of case–control genotype probabilities has been computed. For example, using the data in Table 1, for a recessive gene with $q = 0.14$ and RR = 20, we would expect $c = 17.7\%$ and $b = 0.88\%$ of sib-matched case–control pairs to be informative, leading to a McNemar test of $\chi^2 = (c - b)^2/(c + b) = 0.152N$. To obtain 90% power at a two-sided 5% significance level, one would therefore require $N = (1.96 + 1.28)^2/0.152 = 69$ matched pairs. Of course, for smaller RRs, the required sample size would be larger. One can also use a standard software program to compute sample size for a case and unrelated-control design and then use the values plotted in Fig. 1 to adjust the sample size to a family-controlled design. For example, if one assumed a dominant model and the required number of pairs for a study with unrelated controls was 100, the necessary number of pairs would be 200 (100/0.5) for a case–sib study, or 114 (100/0.88) for a case–cousin study.

Discussion

We have argued that family-based case–control studies offer an attractive alternative to population-based case–control designs using unrelated controls. Their primary advantage is that they overcome the problem of population stratification that can lead to spurious associations with noncausal genes that are not even linked with any causal genes. The sibling and pseudosib designs completely avoid this problem, whereas the cousin-control design avoids it only approximately to the extent that families tend to marry within ethnic groups. This protection from bias is arguably worth the penalty of reduced statistical efficiency resulting from overmatching on genotype. We have also shown that on a per-case basis, the pseudosib design can be more efficient than the unrelated-control design and that restriction to multiple case families can lead to even more efficient designs, if done appropriately. Finally, we have argued that family-based designs offer certain nonstatistical advantages, such as improved cooperation and reduced cost, that must be weighed against the potential loss in sample size from cases who do have a suitable family control and the potential selection bias if such losses are nondifferential.

A spin-off of these family-based designs, particularly those involving cousin controls or restriction to multiple-case families, is the availability of phenotype information on other family members not involved themselves as cases or controls, whose genotypes may not be known. We have considered here the analysis of only the measured genotypes for the selected cases and their matched controls. To take advantage of the entire vector $\boldsymbol{d}$ of phenotype data on family members, one could conduct a "modified segregation analysis" in which one forms a likelihood by summing over all possible genotypes of the untyped individuals $\boldsymbol{g}_u$ conditional on the observed genotypes $\boldsymbol{g}_o$. The ascertainment process (*Asc*) (e.g., that each family contains at least one case and at least one unaffected sibling) can be addressed either by forming a "retrospective" likelihood from terms of the form $Pr(\boldsymbol{g}_o|\boldsymbol{d})$ or by modeling the ascertainment process explicitly in a "prospective" likelihood $Pr(\boldsymbol{d}|\boldsymbol{g}_o,Asc)$ or "joint" likelihood $Pr(\boldsymbol{d},\boldsymbol{g}_o|Asc)$. For example the joint likelihood for a single family would be computed as

Table 3. Asymptotic relative efficiencies of various family-based case–control designs with restrictions on family history, relative to the unrestricted population-control design, assuming the population disease rate is 1%*

		Recessive		Multiplicative		Dominant	
Design	Relative risk	AR = 0.05	AR = 0.20	AR = 0.05	AR = 0.20	AR = 0.05	AR = 0.20
		ARE (%)		ARE (%)		ARE (%)	
SAP	2	71	58	58	53	70	61
	20	105	104	121	96	332	184
SAS	2	81	62	59	54	70	61
	20	313	197	124	103	332	184
CAG	2	102	96	94	90	102	93
	20	134	144	142	124	322	213
PAP	2	160	123	117	108	141	119
	20	318	273	248	204	654	351

*Design codes: SAP = siblings with affected parent; SAS = siblings with affected sibling; CAG = cousins with affected grandparent; PAP = pseudosiblings with affected parent; AR = attributable risk.

Downloaded from https://academic.oup.com/jncimono/article/1999/26/31/908209 by guest on 21 August 2022

$$L(\alpha,\beta,q) = Pr(\boldsymbol{d},\boldsymbol{g}_o|Asc) = \frac{\sum_{\boldsymbol{g}_u} Pr(Asc|\boldsymbol{d})\ Pr(\boldsymbol{d}|\boldsymbol{g}_o,\boldsymbol{g}_u,\alpha,\beta)\ Pr(\boldsymbol{g}_o,\boldsymbol{g}_u|q)}{\sum_{\boldsymbol{g}_u}\sum_{\boldsymbol{d}} Pr(Asc|\boldsymbol{d}) Pr(\boldsymbol{d}|\boldsymbol{g}_o,\boldsymbol{g}_u,\alpha,\beta)\ Pr(\boldsymbol{g}_o,\boldsymbol{g}_u|q)},$$

where the second sum in the denominator is taken over all possible vectors of disease status within the family. The likelihood for a set of families would be computed as the product of family-specific likelihood contributions.

An advantage of these segregation likelihoods is that they need not be restricted to families with at least one case and one unaffected relative. For example, if the initial ascertainment is based on selection of affected case patients from a population registry, all cases and their families can be included using the above likelihoods, while only those cases with an eligible unaffected sibling will be used in the conditional logistic likelihood for the case–sib design. However, whereas the conditional logistic likelihood depends only on the genetic RR parameter β, the segregation likelihoods also involve the baseline risk α and allele frequency *q* as nuisance parameters. Nevertheless, preliminary calculations indicate that incorporation of phenotypic data on relatives can lead to substantial gains in information compared with a case–control design, despite the need to estimate these additional parameters. Another disadvantage of the segregation likelihoods is the greater potential for bias if the form of the model is misspecified, e.g., if one were to assume the parameters were homogeneous across the population when in fact they were variable or if there was additional dependency within families that was not correctly modeled *(27)*.

An additional benefit of these family-based designs is that they can provide a resource for subsequent segregation and linkage analyses to test for and to localize additional genes, after accounting for any measured genes that may partially explain the observed familial aggregation *(28,29)*. To facilitate such studies, it would be helpful to have population-based disease registries with at least some family history data available, even if imperfect. This type of resource would more easily allow the ascertainment of cases with various types of family history, particularly the designs involving restriction to multiple-case families. In summary, family-based case–control designs have a number of attractive features that make them worth considering when designing a gene-association study for cancer or some other complex disease.

REFERENCES

(1) Goldstein AM, Andrieu N. Detection of interaction involving identified genes: available study designs. Monogr Natl Cancer Inst 1999;26:49–54.

(2) Gail MH, Pee D, Benichou J, Carroll R. Designing studies to estimate the penetrance of an identified autosomal dominant mutation: cohort, case–control, and genotype–proband designs. Genet Epidemiol 1999;16:15–39.

(3) Khoury MJ, Beaty TH. Applications of the case–control method in genetic epidemiology. Epidemiol Rev 1994;16:134–50.

(4) Schaid DJ, Sommer SS. Comparison of statistics for candidate-gene association studies. Am J Hum Genet 1994;55:402–9.

(5) Caporaso N, Rothman N, Wacholder S. Case–control studies of common alleles and environmental factors. Monogr Natl Cancer Inst 1999;26:25–30.

(6) Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. $Gm^{3,5,13,14}$ and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am J Hum Genet 1988;43:520–6.

(7) Curtis D. Use of siblings as controls in case–control association studies. Ann Hum Genet 1997;61:319–33.

(8) Witte JS, Gauderman WJ, Thomas DC. Asymptotic bias and efficiency in case–control studies of candidate genes and gene–environment interactions: basic family designs. Am J Epidemiol 1999;149:693–705.

(9) Breslow NE, Day NE. Statistical methods in cancer research: I. The analysis of case–control studies. Vol. 32. Lyon (France): IARC Scientific Publications; 1989.

(10) Rubinstein P, Walker M, Carpenter C, Carrier J, Krassner J, Falk C, et al. Genetics of HLA disease association: the use of the haplotype relative risk (HRR) and the "Haplo-Delta" (Dh) estimates in juvenile diabetes from three racial groups. Hum Immunol 1981;3:384.

(11) Self SG, Longton G, Kopecky KJ, Liang KY. On estimating HLA/disease association with application to a study of aplastic anemia. Biometrics 1991; 47:53–61.

(12) Falk CT, Rubinstein P. Haplotype relative risks: an easy reliable way to construct a proper control sample for disk calculations. Ann Hum Genet 1987;51:227–33.

(13) Ott J. Statistical properties of the haplotype relative risk. Genetic Epidemiol 1989;6:127–30.

(14) Terwilliger JD, Ott J. A haplotype based haplotype relative risk approach to detecting allelic associations. Hum Hered 1992;42:337–46.

(15) Knapp M, Seuchter SA, Baur MP. The haplotype-relative-risk (HRR) method for analysis of association in nuclear families. Am J Hum Genet 1993;52:1085–93.

(16) Schaid DJ, Sommer SS. Genotype relative risks: methods for design and analysis of candidate-gene association studies. Am J Hum Genet 1993;53: 1114–26.

(17) Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506–16.

(18) Tiret L, Nicaud V, Ehnholm C, Havekes L, Menzel HJ, Ducimetiere P, et al. Inference of the strength of genotype-disease association from studies comparing offspring with and without parental history of disease. Ann Hum Genet 1993;57:141–9.

(19) Schaid DJ. General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 1996;13:423–49.

(20) Lubin JH, Gail MH. Biased selection of controls for case–control analyses of cohort studies. Biometrics 1984;40:63–75.

(21) Greenland S, Thomas DC. On the need for the rare disease assumption in case–control studies. Am J Epidemiol 1982;116:547–53.

(22) Anton-Culver H, Kurosaki T, Taylor TH, Gildea M, Brunner D, Bringman D. Validation of family history of breast cancer and identification of the BRCA1 and other syndromes using a population-based cancer registry. Genet Epidemiol 1996;13:193–205.

(23) Curtis D, Sham PC. A note on the application of the transmission disequilibrium test when a parent is missing. Am J Hum Genet 1995;56:811–2.

(24) Spielman RS, Ewens WJ. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 1998;62:450–8.

(25) Whittemore AS, Halpern J. Multi-stage sampling in genetic epidemiology. Stat Med 1997;16:153–67.

(26) Siegmund KD, Whittemore AS, Thomas DC. Multistage sampling for disease family registries. Monogr Natl Cancer Inst 1999;26:43–8.

(27) Gail MH, Pee D, Carroll R. Kin-cohort designs for gene characterization. Monogr Natl Cancer Inst 1999;26:55–60.

(28) Zhao LP, Hsu L, Davidov O, Potter J, Elston R, Prentice RL. Population-based family study designs: an interdisciplinary research framework for genetic epidemiology. Genet Epidemiol 1997;14:365–88.

(29) Zhao LP, Aragaki C, Hsu L, Potter J, Elston R, Malone KE, et al. Integrated designs for gene discovery and characterization. Monogr Natl Cancer Inst 1999;26:71–80.

NOTE

Supported by Public Health Service grants CA58860 (National Cancer Institute) and 5P30 ES07048-03 (National Institute of Environmental Health Services) (W. J. Gauderman and D. C. Thomas), and CA73270 (National Cancer Institute) and RR03655 (National Center for Research Resources) (J. S. Witte), National Institutes of Health, Department of Health and Human Services.

Downloaded from https://academic.oup.com/jncimono/article/1999/26/31/908209 by guest on 21 August 2022