COMBINING INFORMATION WITHIN AND BETWEEN PEDIGREES IN ALLELE SHARING METHODS OF LINKAGE ANALYSIS

by

Jun Teng and David Siegmund

TECHNICAL REPORT NO. 6
August 1996

Prepared Under Grant NIH 1-R01-HG00848
For The National Institutes of Health

Approved for public release; distribution unlimited

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Combining Information Within and Between Pedigrees in Allele Sharing Methods of Linkage Analysis

by

Jun Teng and David Siegmund

TECHNICAL REPORT NO. 6
August 1996

Prepared Under Grant NIH 1-R01-HG00848
For The National Institutes of Health

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Combining Information Within and Between Pedigrees
in Allele Sharing Methods of Linkage Analysis

by

Jun Teng
Department of Statistics
Stanford University

D. Siegmund
Department of Statistics
Margaret Jacks Hall
Stanford University
Stanford, CA 94305

Running Head: Combining Pedigrees in Linkage Analysis
Summary.

This paper is concerned with efficient strategies for linkage analysis using pedigrees containing small numbers of affecteds and identity by descent data from closely spaced markers throughout the genome. Particular attention is paid to additive traits involving phenocopies and/or locus heterogeneity. For a sample of pedigrees containing one particular configuration of affecteds, e.g., pairs of siblings or pairs of siblings together with a first cousin, the power to detect linkage depends primarily on a single parameter, which is a specific function of the configuration and the underlying genetic model, and which can be used to evaluate the efficiency of different strategies. We identify configurations of affecteds that are particularly powerful for detecting linkage, e.g., affected siblings together with a more distant relative, say an affected cousin; and we show how pedigrees containing different numbers and configurations of affecteds can be efficiently combined in an overall test statistic.
Introduction

For parametric linkage analysis of traits having a simple Mendelian mode of inheritance the LOD score is a function of a single parameter, the recombination fraction. The computation of LOD scores within pedigrees can be extremely complicated, especially when the pedigree is large, and almost invariably involves numerical methods; but among pedigrees LOD scores are combined by simple addition. One can introduce a small number of additional parameters to account for reduced penetrance or heterogeneity (Ott 1991), and while these can complicate the numerical problems considerably the principles do not change.

For complex traits methods of linkage analysis based on identity by descent of affected pedigree members appear to offer advantages over parametric methods because they do not require complete specification of the mode of inheritance. Most theoretical research has been restricted to the simple case of affected pairs of a given relation, usually siblings; and reasonable rules for combining information within and between different kinds of pedigrees are not so clearly established. Recent discussions attempting to move beyond consideration of pairs of affected relatives are Feingold, Brown and Siegmund (1993), who discuss pedigrees consisting of trios and suggest some principles for combining different pedigrees, Thomas, Skolnick and Lewis (1994), whose discussion focuses on the extreme of a single large pedigree, Sribney and Swift (1992), who consider sib trios along with the issues of assortative mating and use of unaffected sibs, and Whittemore and Halpern (1994a,b), who propose a statistic for pedigrees containing arbitrary numbers of affecteds, but do not study its power. Kruglyak et al. (1996) have studied the "all pairwise comparisons" statistic and the Whittemore-Halpern statistic within pedigrees, but do not consider the problem of efficiently combining different kinds of pedigrees.

The affected pedigree member method (Weeks and Lange 1988, 1992) is concerned with the linkage analysis of complex traits using data on identity by state together with marker allele frequencies, which are assumed to be known. Different pedigrees are weighted according to the number of affecteds in each pedigree; pairs within pedigrees are weighted without regard to the different amounts of information implied by different relationships among the affecteds. The degree to which such weighting schemes make
efficient use of the data appears not to have been studied.

In this paper we concentrate on a trait having penetrances that are additive within and between loci, i.e., the trait may be heterogeneous, but it is essentially dominant at each individual locus (Risch 1990a,b). We envision a sample consisting primarily of sibling pairs with smaller numbers of other relative pairs, e.g., second degree relatives, first cousins, and also a comparatively small number of moderately larger pedigrees, and we ask how identity by descent data from such a collection of pedigrees can be efficiently utilized. We assume a genome wide search is planned, and that data are available from a dense set of completely informative markers. We use the Haldane mapping function (no interference) to model the recombination process. Throughout we assume that our sample sizes are sufficiently large that it is reasonable to use normal approximations, which simplify an otherwise more complicated situation.

The effect of departures from our basic assumptions is discussed below.

Results

We begin by recalling the calculations used to compare the power of different relative pairs, and then give some generalities about combining pedigrees. Next we consider larger pedigrees, specifically triples and quadruples, where the first issue is one of combining information within pedigrees. Finally we discuss the problem of combining pedigrees of different sizes and configurations of relatives. Broadly speaking our conclusions are (i) for combining affected pairs it is possible to use a rough estimate of the relative risk for offspring to determine a statistic that is quite efficient over a broad range of conditions; (ii) to incorporate moderately larger pedigrees one can consider each pedigree to be worth a number of affected pairs of a particular kind and then group the appropriately weighted pedigrees with the affected pairs; and (iii) pedigrees containing more than two affecteds can under certain conditions be especially powerful.

The situation is intrinsically complex, so in some respects the tools for conceptualizing and studying the problem may be more important than specific conclusions.
1. Relative power of different affected relative pairs

Given a number $N$ of pedigrees consisting of pairs of affected relatives of one kind, e.g., sibs, half sibs, or first cousins, we can scan the genome for linkage using the approximately Gaussian process $Z_t$ described by Feingold, Brown and Siegmund (1993). For unilineal pairs

$$Z_t = (X_t - NP_0)/(NP_0(1 - p_0))^{1/2},$$  \hspace{1cm} (1.1)

where $X_t$ is the number of pairs having identity by descent at locus $t$, which measures genetic distance from one end of the chromosome (the $i$th chromosome, say, although we suppress the chromosomal number to simplify the notation), and $p_0$ is the probability of identity by descent for unaffected relatives of the relation composing our sample. For example, $p_0 = 1/4$ for first cousins and $= 1/2$ for second degree relatives. The case of bilineal relatives is more complicated. For sib pairs and a trait whose penetrances are additive within loci the score statistic is

$$Z_t = (X_{2,t} + X_{1,t}/2 - N/2)/(N/8)^{1/2},$$  \hspace{1cm} (1.2)

where $X_{j,t}$ is the number of sib pairs having identity by descent on $j$ chromosomes at the locus $t$ ($j = 0, 1, 2$) (cf. Blackwelder and Elston 1985, Dupuis, Brown and Siegmund 1995).

As $t$ varies along an unlinked chromosome the statistics (1.1) and (1.2) are approximately stationary Gaussian processes having mean value 0, variance 1, and a covariance function $R(t) = \text{Cov}(Z_s, Z_{s+t})$ involving a parameter $\beta$ defined by $\beta = -R'(0)$, which measures the rate of recombination for the relative pairs involved. For example, when genetic distance $t$ is measured in centimorgans (cM), for sibs and half sibs $R(t) = \exp(-\beta|t|)$ with $\beta = 0.04$. Generally speaking more distant relatives involve larger values of $\beta$, which reflects the greater amount of recombination involved in the larger number of meioses separating the individuals. See Feingold, Brown and Siegmund (1993) for other examples. Linkage is detected if the maximum value of the process $Z_t$ exceeds an appropriate threshold $b$ at a locus $t$ on the genome. The threshold $b$ can be determined by the condition that the false positive rate should be less than a given small number, usually taken
to be 0.05, and the approximation

\[
P\{\max_t Z_t \geq b\} \approx 1 - \exp\{-C[1 - \Phi(b)] - \beta L b \phi(b)\}.
\] (1.3)

Here \( L \) is the genetic length of the genome, \( C \) is the total number of chromosomes, \( \phi \) is the standard normal density function and \( \Phi \) is the cumulative normal distribution function. For a human genome consisting of 23 (pairs of) chromosomes having total genetic length of about 3500 cM, typical values of \( b \) are about 4.1, which corresponds to about 3.65 on the LOD scale. See Feingold, Brown and Siegmund (1993), Kruglyak and Lander (1995) or Lander and Schork (1994) for additional information.

On a linked chromosome containing a single trait susceptibility locus at \( \tau \) the expected value of \( Z_t \) is (Feingold, Brown and Siegmund 1993)

\[
E(Z_t) = \xi R(t - \tau).
\] (1.4)

The noncentrality parameter \( \xi \) can be written \( \xi = N^{1/2} c \alpha \), where \( \alpha \) depends on the mode of inheritance of the trait and the relationship of the affected pair, and \( c \) is a numerical constant depending only on the relationship. For siblings \( c = 2^{-1/2} \), while for unilineal relatives \( c^2 = (2^{d-1} - 1) \), where \( d \) is the degree of the relationship, i.e., \( d = 2 \) for half sibs, 3 for first cousins, etc. Except for a few remarks given below, we shall assume that penetrances are additive within loci. This means that the mode of inheritance is essentially dominant; and for a monogenic trait the parameter \( \alpha \) is a function of \( \lambda_O \), the relative risk of the offspring of an affected individual to be affected (Risch 1990b). For sibs \( \alpha = (\lambda_O - 1)/\lambda_O \). For unilineal relatives \( \alpha = (\lambda_O - 1)/(\lambda_O + c^2) \). Feingold, Brown and Siegmund (1993) give an approximation for the power to detect linkage, which is an increasing function of the noncentrality \( \xi \). In particular a value of \( \xi \) about equal to 5 yields power of approximately 90%. Simple calculations show that for a monogenic trait close relatives are more powerful for small values of \( \lambda_O \) and distant relatives for large \( \lambda_O \). (See also Risch 1990b.) Since the parameters \( \xi \) are directly proportional to the square root of \( N \), for two different groups of relatives of sizes \( N \) and \( N' \), we have equal values \( \xi = \xi' \), hence roughly the same power to detect linkage when \( N/N' = (c'\alpha'/ca)^2 \). For example, for siblings and first cousins this ratio is \( 6\lambda_O^2/(\lambda_O + 3)^2 \), so for a monogenic
trait having $\lambda_O = 4$, one would need only $49/96$ or about half as many first cousin as sib pairs to have a given power to detect linkage. We describe this situation by saying that sib pairs are only about $(49/96) \times 100\% = 51\%$ as efficient as first cousin pairs. See Feingold, Brown and Siegmund (1993) for a systematic comparison of siblings, second degree relatives and first cousins as a function of the relative risk parameter $\lambda_O$.

1.1. Heterogeneous traits

Essentially identical conclusions apply to heterogeneous traits modeled by assuming additivity of penetrances at different (unlinked) loci. It may be shown (Risch 1990a) that $\lambda_O = 1 + \Sigma V_{Ai}/2K^2$, where $K$ is the incidence of the trait and $V_{Ai}$ is the additive variance of the penetrances at the $i$th locus. The parameter $\alpha_i$ associated with the $i$th locus is $V_{Ai}/2K^2/\lambda_O$ for sibs and $V_{Ai}/2K^2/(\lambda_O + c^2)$ for unilineal relatives, so $\Sigma \alpha_i = (\lambda_O - 1)/\lambda_O$ for siblings and $(\lambda_O - 1)/(\lambda_O + c^2)$ for unilineal relatives (Dupuis, Brown and Siegmund 1995). These relations have two important consequences. (i) The ratio of noncentrality parameters $\xi_i$ at the $i$th locus is for different relative pairs exactly the same as for a monogenic trait having the same value of $\lambda_O$. Hence the comparative mapping power of different pairs is the same; and as we shall see below, they should get the same relative weights in an overall statistic as in the monogenic case. (ii) While for a monogenic trait a large value of $\lambda_O$ implies a large noncentrality, so only a comparatively small sample size is required to detect linkage, with a heterogeneous trait large values of $\lambda_O$ are compatible with small noncentrality parameters, which in turn require large samples to detect linkage.

1.2. Some qualifying remarks

The picture sketched above is more complicated and less easily described if some of the assumptions are changed. (i) The assumption that markers are completely informative is not realized in practice. However, it allows us to achieve maximum insight with a minimum of technical complications. The complete multipoint algorithm of Kruglyak et al. (1996) provides a useful statistical technique for approximating the ideal of unambiguous identity by descent at each marker, while improvements in laboratory methods may also alleviate this problem in the future. We will report elsewhere on the effects of
marker density and polymorphism on the power to detect linkage when using a complete multipoint method. (ii) If the penetrances are not additive within loci, so the dominance variance of the penetrances, \( V_D \), is not zero, the relative power of bilineal relatives increases. For example, for siblings the noncentrality parameter \( \xi \) associated with (1.2) changes from \((N/2)^{1/2}\alpha\) to \((N/2)^{1/2}(\alpha + \delta)\). For a monogenic trait we can express \( \alpha \) and \( \delta \) in terms of the recurrence risk for offspring, \( \lambda_O = 1 + V_A/2K^2 \), and the recurrence risk for sibs, \( \lambda_S = 1 + V_A/2K^2 + V_D/4K^2 \) by the relations \( \alpha = (\lambda_S - 1)/\lambda_S \), \( \delta = (\lambda_S - \lambda_O)/\lambda_S \). For unilineal relatives the value of \( \xi \) does not change. See Feingold and Siegmund (1995) for a complete discussion and examples. (iii) If we relax the assumption of a dense set of markers, then usually the trait locus \( \tau \) lies between flanking markers, and the power to detect linkage depends not only on \( \xi \), but also on the position of \( \tau \) in relation to the markers and the recombination rate for those relative pairs. Distant relatives involve larger recombination rates, which means that \( R(t) \) decreases more quickly as \( t \) moves away from 0, so \( E(Z_t) \) given in equation (1.4) decays faster as \( t \) moves away from \( \tau \). The result is a slightly greater loss of power for more distant relatives, although the difference is not substantial at small to moderate intermarker distances (cf. Dupuis 1994). (iv) If several loci interact epistatically in a way that can be modeled by a multiplicative model (Risch, 1990a) or the generalization that allows phenocopies and other additive components (Dupuis, Brown and Siegmund 1995), the power of closer relatives is comparatively greater, sometimes substantially. However, in these cases detection of linkage is usually somewhat easier, so efficient use of the available data may not be as critical. (v) The accuracy of our Gaussian approximations depends on a variety of factors and ranges from excellent to fair. They often exaggerate the relative efficiencies among different classes of relatives, especially when distant relatives are involved. The methods of Feingold (1993), Feingold and Siegmund (1995) and Dupuis, Brown and Siegmund (1995) can be useful in providing more accurate approximations, although for only a subset of the problems addressed below. (For Figure 2 we have used these more accurate approximations.)

In any particular case we can try to take these complications into account, and some of them deserve systematic study. In this paper we have sought to make the analysis as simple as possible. This approach oversimplifies an intrinsically complex situation,
so one should keep in mind that these qualifications decrease somewhat the size of the advantage indicated by our calculations of using more distant relatives.

2. General considerations in Combining Pedigrees

Suppose now that we have \( k > 1 \) groups of relative pairs, say sib pairs, first cousin pairs, etc. Since some of these groups are comparatively more powerful for detecting linkage, it seems plausible that they should get greater weight in an overall statistic. To simplify the discussion, we first consider the case \( k = 2 \) and to be specific assume that our sample consists of \( N \) sibling pairs and \( N' \) first cousin pairs. We assume also that these groups of relative pairs do not contain anyone in common. Let \( Z_t \) denote the sibling process in (1.2) and \( Z'_t \) denote the cousin process in (1.1). Regarded as independent Gaussian processes, the log likelihood function is (Feingold, Brown and Siegmund 1993)

\[
\xi Z_t + \xi' Z'_t - (\xi^2 + (\xi')^2)/2.
\]

This implies that if \( \xi \) and \( \xi' \) were both known, our test statistic would be the maximum as we search over all loci \( t \) of the linear combination

\[
(\xi Z_t + \xi' Z'_t)/[\xi^2 + (\xi')^2]^{1/2},
\]

which has been standardized to have mean 0 and variance 1 on unlinked chromosomes. At a trait locus \( \tau \) it has noncentrality

\[
[\xi^2 + (\xi')^2]^{1/2}.
\]

Since the values \( \xi \) and \( \xi' \) are in general unknown, we cannot routinely use this statistic. However, in order to give relative weights to \( Z_t \) and \( Z'_t \), (i) we need not know the actual values \( \xi, \xi' \), but only the ratio \( \xi/\xi' \) and (ii) for the very important cases of a monogenic or heterogeneous trait whose penetrances are additive between loci, this ratio depends only on the parameter \( \lambda_o \), which in principle can be estimated from pedigree data. For the example under discussion \( \xi/\xi' = (N/N')^{1/2}(\lambda_o + 3)/6^{1/2}\lambda_o \). Hence it seems reasonable to assume that often the ratio \( \xi/\xi' \) will be known to some extent, so we will have some guidance in choosing relative weights for \( Z_t \) and \( Z'_t \). We also note
that the various common violations of our idealized assumptions noted in the remarks at
the end of the preceding section by and large tend to make close relatives, in particular
siblings, comparatively more powerful. Hence to be conservative it seems reasonable that
we should prefer to err on the side of giving siblings somewhat more weight than our
idealized theory would indicate.

Now suppose that we use values \( \hat{\xi} \) and \( \hat{\xi}' \), perhaps guided by pedigree studies, in lieu
of the optimal but unknown weights \( \xi \) and \( \xi' \), so our statistic becomes the maximum over
t of

\[
\hat{Z}_t = (\hat{\xi} Z_t + \hat{\xi}' Z'_t)/[\hat{\xi}^2 + (\hat{\xi}')^2]^{1/2}.
\]

(2.3)

The parameters of the process \( \hat{Z}_t \) are easily derived from the constituent processes.
The parameter \( \hat{\beta} \) of the covariance function is

\[
\hat{\beta} = (\hat{\xi}^2 \beta + (\hat{\xi}')^2 \beta')/[\hat{\xi}^2 + (\hat{\xi}')^2],
\]

which allows us to determine the detection threshold \( b \) by using (1.3), although as noted above
in most cases the threshold is about the same as for sibling pairs. The expected value is

\[
E(\hat{Z}_t) = [\hat{\xi} \xi R(t - \tau) + \hat{\xi}' \xi' R'(t - \tau)]/[\hat{\xi}^2 + (\hat{\xi}')^2]^{1/2},
\]

so the noncentrality parameter of (2.3) equals

\[
E(\hat{Z}_t) = [\hat{\xi} \xi + \hat{\xi}' \xi']/[\hat{\xi}^2 + (\hat{\xi}')^2]^{1/2}.
\]

(2.4)

The noncentrality (2.4) is smaller than the optimal noncentrality (2.2), which, however,
would require knowledge of \( \xi/\xi' \). We can reason as follows to define the efficiency of
the statistic (2.3). Assume that the ratio of the number of sibling pairs to the number of
cousin pairs is fixed. We then ask: if we were to use (2.3) with noncentrality parameter
(2.4), how much larger sample would be required to obtain the optimal noncentrality
(2.2). Since the noncentrality parameters are proportional to the square root of the sam-
ple sizes, increasing the sample size by a factor of \( k \) would increase (2.4) by the factor
\( k^{1/2} \). This would make the increased value of (2.4) equal to (2.2) provided \( k^{1/2} \) equals
(2.2) divided by (2.4), or \( k \) equals the square of this ratio. For example, if \( k = 1.1 \), then
we would need 1.1 times as large a sample to have the same noncentrality using (2.4) as
we would need using the optimal statistic with the original sample. We summarize this
by saying that (2.3) is \((1/k) \times 100\%\) efficient, i.e., we define the efficiency of (2.3) as the square of the ratio of (2.4) to (2.2). Geometrically this is the square of the cosine of the angle between the vectors \((\xi, \xi')\) and \((\hat{\xi}, \hat{\xi}')\). Cox and Hinkley (1974) give an extensive discussion of statistical concepts of efficiency.

Our ability to choose a reasonable value of \(\lambda_O\) and hence reasonable values for \(\hat{\xi}, \hat{\xi}'\) to use in (2.3) will vary from case to case. Although the preceding calculations indicate that when \(\lambda_O\) is large, we should ideally give distant relatives substantially more weight than siblings, we might ask whether it is possible to choose a value that maintains reasonable efficiency over a wide range of values of \(\lambda_O\), no matter what this value is.

Let \(c = \xi'N'^{-1/2}/\xi N^{-1/2}, \hat{c} = \hat{\xi}'N'^{-1/2}/\hat{\xi} N^{-1/2}, r = N/N'\). For siblings and first cousins \(c = 6^{1/2}\lambda_O/(\lambda_O + 3)\), and a similar formula holds for \(\hat{c}\). The loss of efficiency in using \(\hat{\xi}, \hat{\xi}'\), namely the difference between one and the square of the ratio of (2.4) to (2.2), is a function of \(c, \hat{c}\) and \(r\), which for given values of \(c\) and \(\hat{c}\) is maximized as a function of \(r\) at \(r = cc\hat{c}\), where it equals \((c - \hat{c})^2/(c + \hat{c})^2\). From this expression it is easy to see that for the choice \(\lambda_O = 4\) the statistic (2.3) is at least 93% efficient for all \(\lambda_O\) in the range 1.5 to 100. For a true value \(\lambda_O = 10\), it is more than 97% efficient. For the choice \(\lambda_O = 3\), it is more than 89% efficient for all \(1 < \lambda_O < \infty\). As an indication of the robustness of these results against a badly misspecified model, recall that for a multiplicative model \(\lambda_O = \lambda_{O1}\lambda_{O2}\), where \(\lambda_{Oi}\) is the relative risk factor associated with the \(i\)th locus (Risch 1990a). If, for example, \(\lambda_O = 6\) would factor into \(\lambda_{O1} = 4\) and \(\lambda_{O2} = 1.5\), then the choice \(\lambda_O = 4\) would provide the optimal weights for detection of the first locus and would be 93% efficient for detection of the second.

Remarks. (i) A simple ad hoc statistic is for each locus \(t\) to count the total number of chromosomes where there is identity by descent at \(t\). For combining \(N\) sibling pairs and \(N'\) first cousin pairs this amounts to \((N/2)^{1/2} \times (1.2) + [(3N')^{1/2}/4] \times (1.1)\), which is the special case of (2.3) for \(\lambda_0 = 1\). As one would expect from the preceding calculations, this statistic is inefficient when \(\lambda_O\) is large. For example, for \(10 < \lambda_0 < 100\) and for \(N' \leq N \leq 4N'\) it is only about 75-80% efficient. (ii) A different strategy would be to choose the value of \(\lambda_0\) maximizing the log likelihood (2.1). This method would require substantial numerical computation and unfortunately would not generalize easily.
to deal with pedigrees containing more than two affecteds. In view of the satisfactory performance of the method suggested above, we have not pursued this alternative.

3. Pedigrees with Multiple Affecteds

To address the issue of combining information within pedigrees containing more than two affecteds, we recall the following model along the lines of James (1971) and Risch (1990a).

Let \( \varphi \) be the indicator of an individual's phenotype, i.e., \( \varphi = 1 \) or 0 according as the individual is affected or not, and let \( G \) denote the individual's genotype. We begin with a monogenic trait, so the penetrance of \( G = \{a, b\} \) can be expressed in the form

\[
E(\varphi|G) = f_a + f_b + d_{ab},
\]

where \( f_a \) denotes the additive contribution to the penetrance of the allele \( a \) and \( d_{ab} \) is the dominance deviation. The frequency of affecteds in the population is \( K = 2f_\cdot + d_\cdot \), where the dot (\( \cdot \)) as a subscript indicates an average with respect to the Hardy-Weinberg frequencies \( \{p_a\} \), so \( f_\cdot = \Sigma p_a f_a \) and \( d_\cdot = \Sigma p_a p_b d_{ab} \). By the standard analysis of variance decomposition we also have

\[
E(\varphi|G) = K + \tilde{f}_a + \tilde{f}_b + \tilde{d}_{ab},
\]

where \( \tilde{f}_a = f_a - f_\cdot + d_\cdot \cdot - d_\cdot \), \( \tilde{d}_{ab} = d_{ab} - d_a \cdot - d_b \cdot + d_\cdot \). A basic assumption is that the phenotypes of two or more individuals are conditionally independent given their genotypes, i.e.,

\[
E(\varphi_1 \varphi_2 \cdots \varphi_n|G_1, \cdots, G_n) = E(\varphi_1|G_1) \cdots E(\varphi_n|G_n).
\]

Then (James 1971, Risch 1990a) the probability that two relatives are both affected is

\[
E(\varphi_1 \varphi_2) = K^2 + (\frac{1}{2} \Sigma p_a f_a^2) \epsilon_{12} + V_{D} u_{12} \quad (3.1)
\]

where \( V_A = 2\Sigma p_a f_a^2 \) is the additive variance of the penetrances, \( V_D = \Sigma p_a p_b \tilde{d}_{ab}^2 \) is the dominance variance of the penetrances, \( \epsilon_{12} \) is the expected number of alleles shared
identical-by-descent by individuals 1, 2 and \( u_{12} \) is the probability that both alleles are shared identical-by-descent.

In an outbred population \( u_{12} = 0 \) except for siblings, double first cousins, double second cousins, etc. Indeed, for linkage analysis of recessive traits, for which \( V_D \gg V_A \), it is primarily pedigrees containing these relatives that contribute significant information. We have simplified our calculations by assuming \( V_D = 0 \), so within loci we in effect consider only a dominant trait etiology. Under this additivity assumption an analogue of (3.1) for three relatives is given by Feingold, Brown, and Siegmond (1993), and applied to sibling trios. Below we give the corresponding result for four relatives, which contains some new features. It will be clear that the calculation can be carried out for larger collections of relatives, subject only to one's willingness to deal with some cumbersome expressions.

It will be helpful to introduce the notation

\[
c_\nu = K^{-\nu} \sum p_a \tilde{f}_a
\]

for \( \nu = 2, 3, \ldots \). Also let \( \lambda_{12\ldots n} = K^{-n} E(\varphi_1 \varphi_2 \cdots \varphi_n) \) denote the relative risk for a particular configuration of \( n \) relatives all to be affected. With this notation (and the assumption \( V_D = 0 \)), the relation (3.1) can be written more compactly as \( \lambda_{12} = 1 + c_2 e_{12} \); and the Feingold, Brown, and Siegmond (1993) calculation for three relatives becomes \( \lambda_{123} = 1 + c_2 \sum_{i<j} e_{ij} + c_3 e_{123} \), where \( e_{123} \) is the expected number of alleles shared identical-by-descent by all three relatives.

As (3.1) shows, for pairs of affecteds the relative risk parameter \( \lambda_{12} \) depends only on the incidence \( K \) of the trait and the additive variance of the penetrances (since we have assumed that the dominance variance vanishes). For larger groups of affected relatives higher moments of the penetrances become involved. In order to develop examples involving a relatively small number of parameters in terms of which these moments can be computed, it is helpful to consider the standard two allele formulation, where genotypes \( aa, aA \) and \( AA \) have penetrances \( f_0, f_0 + f + d \) and \( f_0 + 2f \), respectively. Here \( f_0 \) is the percentage of phenocopies, \( f \) is the additive contribution to the penetrance of the disease susceptibility allele, and \( d \), which we are assuming to equal 0, is the dominance deviation.
If $p$ denotes the frequency of allele $A$ and $q = 1 - p$ the frequency of $a$, then $K = f_0 + 2pf$, and the $c_\nu$ are given by

$$c_2 = K^{-2}pqf^2, \quad c_3 = K^{-3}pq(q - p)f^3, \quad c_4 = K^{-4}pq(p^2 + q^2 - pq)f^4.$$ 

Note that if $f_0 = 0$, the $c_\nu$ do not depend on $f$, which cancels out of the numerator and denominator. Even for this very simple model, there are at least $2^3 = 8$ situations to consider: low or high penetrance, with or without phenocopies, and a low or not so low frequency of the disease susceptibility allele. Since we are interested primarily in complex diseases, we shall not consider the case of an extremely rare disease susceptibility allele except in the presence of phenocopies or heterogeneity, which make the detection of linkage more difficult.

We are also interested in heterogeneous traits, which as Risch (1990a) observes can be effectively modeled by assuming additivity of penetrances between loci. We consider briefly the case of two trait loci, which we assume lie on different autosomes, and we denote by a prime (') quantitites associated with the second locus. Then $K = 2\Sigma p_a f_a + 2\Sigma p'_a f'_a = 2(f. + f')$, and we put $c_\nu = K^{-\nu}\Sigma p_a f_a^\nu, \quad c'_\nu = K^{-\nu}\Sigma p'_a (f'_a)^\nu$. The basic relations for two and three affected relative pairs are $\lambda_{12} = 1 + (c_2 + c'_2)e_{12}$ and $\lambda_{123} = 1 + (c_2 + c'_2)\Sigma_{i<j} e_{ij} + (c_3 + c'_3)e_{123}$. In this case in the absence of phenocopies the relative risks depend on the ratio of the penetrances of the trait alleles, and are independent of the penetrances only when the penetrances are equal.

3.1. Sib triples.

To describe the statistic suggested by Sribney and Swift (1992), Whittemore and Halpern (1994b) and Feingold, Brown and Siegmund (1993) for sibling triples, let $X_{k,t}$ denote the number of pairwise comparisons of the three sibs for which $i$ paternal and $j$ maternal chromosomes are identical by descent at the locus $t$, where $(i + j)/2 = k$, $k = 1, 2, 3$. (The possible values of $i$ and $j$ are 1 and 3.) The log likelihood function at the trait locus $\tau$ is

$$X_{3,\tau}\log(1 + 3\alpha) + X_{2,\tau}\log(1 + \alpha) + X_{1,\tau}\log(1 - \alpha),$$
where for a monogenic trait \( \alpha = (c_2 + c_3/2)/(1 + 3c_2 + c_3/2) \) (Feingold, Brown and Siegmund 1993). A test for linkage can be based on the maximum over \( t \) of the score statistic, i.e., the derivative of the log likelihood with respect to \( \alpha \), evaluated at \( \alpha = 0 \) (the hypothesis of no linkage) and normalized to have unit variance when \( \alpha = 0 \). For a general discussion of the score statistic and its optimality properties see Cox and Hinkley (1974, Chapter 9). After some calculation we obtain

\[
Z_t = (2X_{3,t} + X_{2,t} - N/2)/(3N/8)^{1/2}.
\] (3.2)

This statistic amounts to making the three possible pairwise comparisons, which under the hypothesis of no linkage are pairwise independent although not mutually independent, and adding those results together as in (1.2). The non-centrality parameter is \( \xi = (3N/2)^{1/2} \alpha \). The corresponding parameter for \( N \) sibling pairs is \( (N/2)^{1/2} \alpha \), where \( \alpha = c_2/(1 + c_2) \). Sometimes the two values of \( \alpha \) are about equal, and then one sib triple, involving three pairwise comparisons, gives about as much information for detecting linkage as three sib pairs. For a heterogeneous trait involving two unlinked loci the value of \( \alpha \) at the first locus for sibling triples is \( \alpha = (c_2 + c_3/2)/(1 + 2(c_2 + c'_2) + (c_3 + c'_3)/2) \) and for sibling pairs is \( \alpha = c_2/(1 + c_2 + c'_2) \). The value \( \alpha' \) at the second locus is similar.

For ease of exposition in the following we give the parameters explicitly only in the monogenic case. The appropriate modifications for heterogeneous traits always follow the same pattern in the case of pedigrees containing two or three affecteds, but are more complicated when four or more affecteds are involved (cf. the Appendix).

Table 1 contains numerical examples designed to illustrate the complexity one immediately encounters when dealing with pedigrees containing more than two affecteds, particularly for heterogeneous traits. Since the concepts of efficiency defined above involve ratios of noncentrality parameters, the values given in the table are those for a single pair or triple. The noncentrality associated with \( N \) pairs or triples is \( N^{1/2} \) times that for a single pair or triple. A range of allele frequency, penetrance and phenocopy rates is given in the table. In those cases where there are phenocopies, they contribute 1/3 or 1/2 the total incidence of the trait. When there are two loci contributing to the trait, there are four cases considered: (i) equal penetrances and allele frequencies at the two
loci; (ii) equal penetrances at the two loci, but allele frequency at the second locus equals 1/2 that at the first; (iii) the penetrance at the second locus is 1/2 as large and allele frequency is twice as large as at the first locus, so the incidence of the trait associated with each locus is the same although the familial aggregation associated with the first locus is much higher than that associated with the second; (iv) a slightly more extreme example along the lines of (iii).

The principles given above also apply to the problem of combining the sibling pairs and the triples. In cases where the parameter $\alpha$ for sibling triples approximately equals the corresponding parameter for pairs, a single triple gives about as much information as three pairs. In this case the pairs and triples can be efficiently combined simply by considering the three pairwise combinations of the triples as if they involved separate pedigrees. With some qualifications this situation prevails when the trait alleles are relatively rare; but since this scheme for combining pairs and triples is so simple, one is tempted to use it unless it is strongly contraindicated. When the number of triples is small compared to the number of pairs, it is unlikely that one can make large gains in efficiency even if the theoretically optimal combination of pairs and triples is used. A numerical example is given below.

An interesting situation occurs for a heterogeneous trait when one trait allele has relatively low frequency and high penetrance and the other relatively high frequency and low penetrance. Then at the low frequency high penetrance locus a sibling triple can be more informative than three independent pairs, while at the low penetrance high frequency locus, a sibling triple can be less informative than a single sibling pair, a possibility that one can infer qualitatively by considering the functional form of $c_2, c'_2, c_3$ and $c'_3$ as functions of $p, p', f, f'$. Presumably the intuitive meaning is that under these conditions pedigrees with large numbers of affecteds are relatively unlikely to have their trait linked to the low penetrance locus. In this situation the theoretically optimal combination of sibling pairs and triples would be different at the two loci, a problem that does not arise when dealing with affected pairs. We return to this issue below.

4. Examples
In this section we give similar results for other pedigrees, which have been selected to provide what we believe are some general qualitative insights.

4.1. Two siblings and a child

Consider $N$ pedigrees consisting of two siblings and a child of one of the siblings, hence niece or nephew of the other. This is a very simple example of two "distant" relatives (i.e., the avuncular pair) with one or more intervening affected relatives (in this case the parent of the niece or nephew). It is natural to ask if one gains a significant amount of information by including the intervening relatives in the analysis, or if essentially all information is contained in the pair of distant relatives. In this case at least, we shall see that the avuncular pair usually contains the important information. A similar example leading to the same conclusion is two cousins and a parent of one of the cousins, hence uncle (or aunt) of the other.

Let $Y_{ij}$ denote the number of pedigrees in which the siblings are identical by descent on $i$ chromosomes and the avuncular pair are identical by descent on $j$ ($i = 0, 1, 2$, $j = 0, 1$). It may be shown by the methods of Feingold, Brown and Siegmund (1993) or the appendix of this paper that the log likelihood function at the trait locus is

$$Y_{21} \log(1 + \alpha + \delta) + Y_{00} \log(1 - \alpha - \delta) + Y_{11} \log(1 + \alpha) + Y_{10} \log(1 - \alpha),$$

where for a monogenic trait

$$\alpha = \frac{(c_2 + c_3)/2}{1 + 5c_2/2 + c_3/2}, \quad \delta = \frac{c_2}{1 + 5c_2/2 + c_3/2}.$$  

The fact that the likelihood function depends on two parameters makes this (and most of the following) examples fundamentally more complicated than pairs of relatives or sibling triples, where the likelihood function involves only a single unknown parameter. The derivative with respect to $\alpha$, evaluated at $\alpha = \delta = 0$ and standardized to have unit variance under the hypothesis of no linkage is

$$Z_{1,t} = (2Y_{1} - N)/N^{1/2},$$

where the dot indicates summation over the first subscript. This is just the avuncular comparison, ignoring the intervening parent. The derivative with respect to $\delta$, evaluated
at $\alpha = \delta = 0$ and standardized is

$$Z_{2,t} = 2^{1/2}(Y_{21} - Y_{00})/N^{1/2},$$

which is the sibling comparison, ignoring the child. The noncentrality parameters of these two statistics are

$$\xi_1 = E(Z_{1,r}) = N^{1/2}(\alpha + \delta/2), \quad \xi_2 = E(Z_{2,r}) = (N/2)^{1/2}(\alpha + \delta).$$

Obtaining an optimal combination of $Z_1$ and $Z_2$ can be regarded as essentially the same problem as the problem of combining different classes of relative pairs, each of which has its own noncentrality parameter; but unlike that case these two processes are correlated. To evaluate the optimal noncentrality parameter, we consider one of the processes, say $Z_1$ and the standardized residual of the regression of $Z_2$ on $Z_1$, namely $(Z_2 - \rho Z_1)/(1 - \rho^2)^{1/2}$, where $\rho$ is the correlation, under the hypothesis of no linkage, between the two processes. This residual is uncorrelated with $Z_1$. We can evaluate its expectation: $(\xi_2 - \rho \xi_1)/(1 - \rho^2)^{1/2}$, which can then be combined with the expectation of $Z_1$ as above. The result of some algebra is an optimal noncentrality of

$$N^{1/2}(\alpha^2 + \alpha \delta + \delta^2/2)^{1/2}. \quad (4.1)$$

Note, however, that the linear combination of $Z_1$ and $Z_2$ that has this noncentrality depends on the unknown parameters $\xi_1$ and $\xi_2$, hence cannot be used as a statistic. We could use the data to estimate these parameters, or what is essentially equivalent construct a chi-square statistic; but this would involve a second degree of freedom, hence a higher threshold for detection. Even so, one might reasonably consider this strategy if the entire sample consisted of such pairs. For the situation we envision, where there are several different kinds of pedigrees it is not clear how one might implement this idea. In addition, as numerical examples given below indicate, the optimal noncentrality is rarely significantly larger than the noncentrality $\xi_1$ for the avuncular comparison by itself. This can be demonstrated mathematically if we observe that typically $c_3 > c_2$, so $\alpha > \delta$, and it is reasonable to approximate the expression in (4.1) by taking two terms of a Taylor
series expansion, with $\delta$ assumed to be small. The result is $N^{1/2}(\alpha + \delta/2 + \delta^2/8\alpha)$, which is only slightly larger than $\xi_1$ when $\delta$ is small compared to $\alpha$. For the examples in Table 1, $Z_1$ is always at least 90% efficient, and usually more than 95% efficient. See Figure 1. Intuitively it is not surprising that we do not get substantially more information by considering explicitly the intervening affected parent, because we expect to find that parent carrying a trait allele, hence affected—except possibly when the penetrances are small.

For the same reason it is not surprising to find numerically that in many cases this pedigree gives roughly the same information as an avuncular pair with the status of the intervening parent unspecified. To the extent that this observation is generally true, it would be very useful when we turn to the problem of combining different kinds of pedigrees, because it would suggest that we simply ignore the intervening parent and treat these pedigrees as avuncular pairs. If there is a substantial frequency of phenocopies or if the trait is heterogeneous with say two trait loci, where at one locus the mutant allele has a low frequency and a high penetrance, while at the other these relations are reversed, then these pedigrees can be substantially more informative than avuncular pairs. Table 2 contains numerical examples.

4.2. Two siblings and a grandparent or first cousin

For the next example we examine the effect of having available a more distant affected relative of two affected siblings. It turns out that this situation can be quite powerful. We first consider $N$ pedigrees, each of which consists of two affected sibs and an affected (maternal) grandparent, then turn to the similar case of two affected sibs and their affected cousin.

At any given locus, for $i, j, k$ equal to 0 or 1 we define $Y_{ijk}$ as follows. Let $Y_{ijk}$ equal the number of pedigrees where the affected grandparent and the first sibling are identical by descent ($i = 1$) or not ($i = 0$), the grandparent and the second sibling are identical by descent ($j = 1$) or not ($j = 0$), and the siblings are identical by descent on their paternally inherited chromosome ($k = 1$) or not ($k = 0$). Thus $Y_{101}$ denotes the number of pedigrees where the affected grandparent and the first sibling are identical by descent,
the grandparent and the second sibling are not identical by descent, and the two siblings are identical by descent on their paternally inherited chromosome. Under the hypothesis of no linkage the \( Y_{ijk} \) are multinomial with probabilities equal to \( 1/8 \). It may be shown that the log likelihood function at a trait locus is

\[
Y_{111} \log(1 + 3 \alpha + \delta) + Y_{110} \log(1 + 3 \alpha - \delta) + (Y_{101} + Y_{011}) \log(1 - \alpha + \delta)
\]

\[
+ (Y_{100} + Y_{010}) \log(1 - \alpha - \delta) + Y_{001} \log(1 - \alpha + \delta) + Y_{000} \log(1 - \alpha - \delta),
\]

where for a monogenic trait

\[
\alpha = (c_2/2 + c_3/4)/(1 + 2c_2 + c_3/4), \quad \delta = c_2/2/(1 + 2c_2 + c_3/4).
\]

The derivatives with respect to \( \alpha \) and \( \delta \), evaluated at \( \alpha = \delta = 0 \), and standardized to have unit variance under the null hypothesis of no linkage are

\[
Z_{1,t} = 4[3Y_{11} - (Y_{10} + Y_{01} + Y_{00})]/(3N)^{1/2}
\]

and

\[
Z_{2,t} = (Y_{-,1} - Y_{-,0})/N^{1/2},
\]

respectively. The statistic \( Z_1 \) before standardization simply counts the number of pedigrees where all three affecteds are identical by descent. It ignores the status of the siblings on their paternal chromosome. It is also equivalent to taking the sum of the identity by descent comparisons of the affected grandparent with the first sib, with the second sib, and the identity by descent comparison of the sibs on their maternally inherited chromosome. The statistic \( Z_2 \) compares the siblings on their paternal chromosome as if they were half sibs and ignores the relation of the affected grandparent to the sibs. The noncentrality parameters are \( \xi_1 = (3N)^{1/2} \alpha \) and \( \xi_2 = N^{1/2} \delta \). On unlinked chromosomes \( Z_1 \) and \( Z_2 \) are uncorrelated. Hence the optimal noncentrality is \( N^{1/2}(3\alpha^2 + \delta^2)^{1/2} \). In all cases \( \alpha > \delta \). If it is much greater, as it will be whenever \( c_3 \) is large compared to \( c_2 \), the statistic \( Z_1 \) will be essentially optimal, as we expected. See Figure 1.

Table 2 gives some numerical examples of the noncentrality of \( Z_1 \), which show that these pedigrees can be very powerful—often as powerful as three independent second degree
relative pairs, and in some cases even more powerful. For example for the seventh row of Tables 1 and 2, all the relative pairs are about equally powerful at both loci. For either locus about 550-600 affected pairs would be required to have 90% power to detect linkage to that locus. Only about 100 pedigrees containing two sibs and an affected grandparent would be required to detect the rare mutant, high penetrance locus.

For a fully penetrant dominant trait, two affected siblings and an affected grandparent give two phase known meioses. In this case a recombination between the disease locus and a marker in a sibling has occurred if and only if the grandparent and that sibling are not identical by descent at the marker. The paternally inherited chromosome carries no linkage information (unless the father is also affected). The statistic \( Z_1 \) in effect counts recombinants in phase known meioses and hence in this special case is expected to be very powerful compared to a sibling pair statistic that ignores the affected grandparent. What perhaps is not expected is that this statistic is much more powerful than the customary sibling pair statistic (1.2) over a broad range of conditions involving the complexities of incomplete penetrance, phenocopies and heterogeneity.

The Whittemore-Halpern (1994b) statistic is easily calculated from their Table 2 and found to equal

\[
(4Y_{111} + 3Y_{110} + Y_{101} + Y_{011} - Y_{000} - N)/(5N/2)^{1/2}.
\]

Its noncentrality parameter is \((5\alpha + \delta)(N/10)^{1/2}\). When \(\delta\) is small compared to \(\alpha\), the relative efficiency of this statistic compared to \(Z_1\) is 5/6. If we force the Whittemore-Halpern statistic to neglect the relation of the sibs on their paternal chromosome by pretending they are half sibs, it then equals \(Z_1\).

Now consider the similar case of two affected siblings and their affected first cousin. To be specific assume the siblings are related to the cousin through their mothers. To simplify the analysis of this example we will from the beginning neglect the relation of the siblings on their paternal chromosome. At a given locus, for \(i, j, k = 0\) or \(1\), let \(Y_{ijk}\) denote the number of pedigrees where the two sibs are identical by descent on their maternally inherited chromosome \((i = 1)\) or not \((i = 0)\), sibling number one and the cousin are identical by descent \((j = 1)\) or not \((j = 0)\), and sibling number two and the
cousin are identical by descent \((k = 1)\) or not \((k = 0)\). It may be shown that the log likelihood function at the trait locus is

\[
Y_{111} \log(1 + 7\alpha + 2\delta) + (Y_{100} + Y_{010} + Y_{001}) \log(1 - \alpha) + Y_{000} \log(1 - \alpha - \delta),
\]

where for a monogenic trait

\[
\alpha = \frac{c_3/8}{1 + 3c_2/2 + c_3/8}, \quad \delta = \frac{c_2}{1 + 3c_2/2 + c_3/8}.
\]

The derivative with respect to \(\alpha\) evaluated at \(\alpha = \delta = 0\) and normalized to unit variance under the hypothesis of no linkage is

\[
Z_{1,t} = 8(Y_{111} - N/8)/(7N)^{1/2}.
\]

The corresponding derivative with respect to \(\delta\) is

\[
Z_{2,t} = 2(2Y_{111} - Y_{000})/(3N)^{1/2}.
\]

The statistic \(Z_1\) is equivalent to counting the number of pedigrees where all three comparisons show identity by descent, while \(Z_2\) counts the number of pairwise comparisons showing an allele inherited identical by descent. We can evaluate the optimal noncentrality as above to obtain \([28\alpha^2 + 3\delta^2 + 16\alpha\delta]^{1/2}/2\). The non-centrality of \(Z_1\) by itself is \((7\alpha + 2\delta)/7^{1/2}\), while that of \(Z_2\) is \([8\alpha + 3\delta]/2 \times 3^{1/2}\).

Usually \(\alpha\) will be larger than \(\delta\), and in the rare allele case it will be much larger. If we expand the optimal noncentrality in a Taylor series, we obtain \(7^{1/2}\alpha + 2\delta/7^{1/2} + O(\delta^2/\alpha)\), which suggests that \(Z_1\) will be close to optimal. Numerical calculations verify this and show that \(Z_1\) is somewhat better than \(Z_2\) for most of the conditions we have considered, although neither is uniformly better than the other. Moreover, each such pedigree is often roughly as powerful as two independent first cousin pairs. See Table 2 for numerical examples.

The Whittemore-Halpern (1994b) statistic is

\[
8(4Y_{111} + Y_{100} + Y_{010} + Y_{001} - 9N/8)/(87N)^{1/2}
\]
and has noncentrality

\[
(N/87)^{1/2}(23\alpha + 86).
\]

Hence it is about 87% efficient in the limiting case of a rare allele when \(\delta << \alpha\), and it is at least 89% efficient over the range of conditions we have studied numerically in Tables 1 and 2.

4.3. Two siblings and one affected parent

Risch (private communication) has considered incorporating parental status into the analysis of sibling pairs and triples. Here we show how our methods can be used to study this issue. We discuss the case of an affected mother, unaffected father, and two affected daughters, although the principles are the same for any specification of parental status. Denoting the mother, father and two sibs by the subscripts 1,2,3,4, respectively, the probability that the mother and both sibs are affected while the father is not equals

\[
E[\phi_1(1 - \phi_2)\phi_3\phi_4] = E[\phi_1\phi_3\phi_4] - E[\phi_1\phi_2\phi_3\phi_4],
\]

so the analysis of this case makes use of the calculations we have performed for three affecteds and also those for four affecteds given in the appendix. Let \(Y_{11}\) equal 1 if the sibs are identical by descent on both their maternally and paternally inherited chromosomes, and 0 otherwise, \(Y_{10}\) equal 1 if they are identical by descent on their maternal and not their paternal chromosome, and 0 otherwise, etc. It may be shown that the log likelihood function is

\[
Y_{11}\log(1 + \alpha) + Y_{10}\log(1 + \delta) + Y_{01}\log(1 - \delta) + Y_{00}\log(1 - \alpha),
\]

where for a monogenic trait

\[
\alpha = [(1 - K)c_2 + (0.5 - K)c_3]/[1 - K + (3 - 5K)c_2 + (0.5 - K)c_3 - 2Kc_2^2],
\]

\[
\delta = 0.5c_3/[1 - K + (3 - 5K)c_2 + (0.5 - K)c_3 - 2Kc_2^2].
\]

The standardized derivatives with respect to \(\alpha\) and \(\delta\) evaluated at \(\alpha = \delta = 0\) are respectively

\[
Z_{1,t} = (Y_{11} - Y_{00})/(N/2)^{1/2}
\]

and

\[
Z_{2,t} = (Y_{10} - Y_{01})/(N/2)^{1/2}.
\]
The statistic $Z_1$ is just the sib pair statistic (1.2) re-expressed in the present notation. Its expectation at the trait locus is $(N/2)^{1/2} \alpha$. The expectation of $Z_2$ at the trait locus is $(N/2)^{1/2} \delta$, and under the hypothesis of no linkage these two statistics are uncorrelated, so the optimal noncentrality is $(N/2)^{1/2}(\alpha^2 + \delta^2)^{1/2}$. It seems plausible that when we know that the mother is affected we might reasonably treat the daughters as if they were half siblings by considering only their identity by descent relation on their maternal chromosome. We can do this by using

$$Z_{3,t} = (Z_{1,t} + Z_{2,t})/2^{1/2} = (Y_1 - Y_0)/N^{1/2},$$

which has expectation at the trait locus equal to $N^{1/2}(\alpha + \delta)/2$. Although $\delta < \alpha$, for rare trait alleles, these two parameters are approximately equal. In that case $Z_3$ is approximately 100% efficient, while $Z_1$ is only about 50% efficient.

The Whittemore-Halpern statistic is

$$(6Y_{11} + 4Y_{10} + 2Y_{01} - 3N)/(5N)^{1/2};$$

its noncentrality parameter is $\xi = N^{1/2}(3\alpha + \delta)/2(5)^{1/2}$.

Figure 1 gives for the models in Table 1 the efficiencies of several of the statistics proposed above: (i) the avuncular comparison in the case that the intervening parent is also affected, (ii) the half sib comparison of two affected siblings when one parent is affected, (iii) the Whittemore-Halpern statistic and (iv) the ordinary sib pair statistic (2.1) for the case that one parent is affected. The first two are very efficient when the frequency of the mutant allele is less than about 0.05 and reasonably efficient otherwise. The Whittemore-Halpern statistic does not use the information as efficiently when the mutant allele frequencies are small, say less than about 0.05, but is very efficient when they are larger. The sib pair statistic (1.2) wastes information when the frequency of the mutant allele is small and is moderately efficient at larger frequencies.

For a dominant trait, sufficiently rare that we can safely assume there is only one trait allele among the four parental chromosomes, the statistic $Z_3$ provides the same linkage information as the phase unknown parametric statistic (although we are using it differently than it would ordinarily be used in a parametric analysis). Thus, we expect
it to be very powerful in this case. As in the earlier example of affected siblings and
an affected grandparent, it is perhaps surprising that \( Z_3 \) is substantially more powerful
than the ordinary sibling pair statistic (1.2) over a wide range of conditions involving
complexities of reduced penetrance, phenocopies and heterogeneity.

In Figure 2 the noncentralities associated with Locus 1 in Tables 1 and 2 have been
translated into sample sizes necessary to detect that locus with probability 0.90. The
number of pedigrees containing various configurations of three affecteds is plotted on
the vertical axis as a function of the number of sibling pairs, plotted on the horizontal
axis. To obtain these values we have used the approximations of Feingold (1993) and
Dupuis, Brown and Siegmund (1995), which are often more accurate than the Gaussian
approximations used for other calculations in this paper, but which also are not easily
derived for some of the other problems we have considered. It is readily apparent that
the pedigrees containing three affecteds can be much more powerful than sibling pairs.
The biggest difference occurs for heterogeneous traits when Locus 1 has comparatively
higher penetrance and lower trait allele frequency than Locus 2.

4.4. Three siblings, two of whom are affected

Although allele sharing methods of linkage analysis are often applied to affecteds
only, it is of some interest to consider the possibility of analysing both affecteds and
unaffecteds in a pedigree. As an example we consider the case of \( N \) pedigrees, each
containing one unaffected and two affected siblings. This case has also been studied by
Sribney and Swift (1992). As one might expect, when unaffecteds are included in the
analysis the values of the penetrances play an important role; when penetrances of trait
susceptibility alleles are large, including the unaffected can be very powerful, but when
penetrances are small very little is gained.

The probability that individuals 1 and 2 are affected and individual 3 is unaffected
is \( E[\varphi_1 \varphi_2 (1 - \varphi_3)] = E[\varphi_1 \varphi_2] - E[\varphi_1 \varphi_2 \varphi_3] \), so the analysis of this situation is a direct
application of the methods already discussed for pedigrees having 2 and 3 affecteds. Some
calculations show that the situation can be parameterized by

\[
\alpha = [c_2 - K(c_2 + c_3 / 2)]/[1 + c_2 - K(1 + 3c_2 + c_3 / 2)]
\]
and
\[ \delta = K[c_2 + c_3/2]/[1 + c_2 - K(1 + 3c_2 + c_3/2)]. \]

When the penetrance of the trait allele is small, \( K \) and hence also \( \delta \) are small, while \( \alpha \) is essentially the same as for an affected sib pair.

To be more precise we assume that siblings 1 and 2 are affected, while sibling 3 is unaffected. We let \( Y_{ijk} \) be one or zero according as siblings 1 and 2 share \( i \) alleles (0,1,or 2) identical by descent siblings 1 and 3 share \( j \), and siblings 2 and 3 share \( k \). For example, \( Y_{222} \) is one if and only if all siblings share two alleles identical by descent.

It may be shown that the log likelihood function is
\[ Y_{200} \log(1 + \alpha + 2\delta) + Y_{211} \log(1 + \alpha) + (Y_{110} + Y_{101}) \log(1 + \delta) + (Y_{020} + Y_{002}) \log(1 - \alpha) \]
\[ + Y_{222} \log(1 + \alpha - 2\delta) + (Y_{121} + Y_{112}) \log(1 - \delta) + Y_{011} \log(1 - \alpha). \]

The partial derivatives with respect to \( \alpha \) and \( \delta \), evaluated at \( \alpha = \delta = 0 \) and normalized to have unit variance are
\[ Z_{1,t} = (Y_{2...} - Y_{0...})/(N/2)^{1/2} \quad (4.2) \]
and
\[ Z_{2,t} = [2(Y_{200} - Y_{222}) + Y_{110} + Y_{101} - Y_{121} - Y_{112}]/(3N/2)^{1/2}. \quad (4.3) \]

The mean values of (4.2) and (4.3) are respectively \((N/2)^{1/2}\alpha\) and \(N^{1/2}\delta\). The correlation under the hypothesis of no linkage is zero, so the optimal non-centrality is \((N/2)^{1/2}[\alpha^2 + 2\delta^2]^{1/2}\). A Taylor series expansion of the optimal non-centrality shows that using (4.2) by itself as a test statistic, which is tantamount to neglecting the unaffected sib, is efficient when \( \delta \) is small compared to \( \alpha \). From numerical calculations one sees that this is true for a single locus trait or for two loci contributing equally to a heterogeneous trait when the incidence of the trait is in the range 0.01-0.2 and the penetrance of the trait susceptibility allele(s) is less than about 0.3. However, if the penetrances are larger than about 0.5, the value of \( \delta \) is not small compared to \( \alpha \) and there can be a substantial loss of efficiency in neglecting (4.3).

Sribney and Swift (1992) suggest the statistic
\[ Z_t = (Z_{1,t} + 2^{1/2}Z_{2,t})/3^{1/2}, \]

26
which has noncentrality $N^{1/2}(\alpha + 2\delta)/6^{1/2}$. This would be the score statistic when $\delta = \alpha$
and can be seen numerically, as expected, to be quite efficient when the penetrances of
the trait susceptibility alleles are large. In fact, for sufficiently large penetrances two
affecteds and one unaffected can actually provide more information for the detection of
linkage than three affecteds, whereas for penetrances so small that the unaffected sibling
can be neglected three affecteds provide up to three times as much information.

In view of the critical dependence of the value of an unaffected sibling on the pen-
etrance, no fixed linear combination of (4.2) and (4.3) can be completely satisfactory in
all situations. An alternative possibility would be a $\chi^2$ like statistic $[(\mathcal{Z}_{1,t}^2 + \mathcal{Z}_{2,t}^2)/2]^{1/2}$.
This would have the the optimal non-centrality but would require a higher detection
threshold because it has two degrees of freedom. In view of the new features involved
in the inclusion of non-affecteds in our analysis, we do not attempt a more systematic
discussion here.

4.5. Detecting linkage using sibling quadruples

To see the effect of a pedigree containing a larger number of affecteds, we consider
the case of $N$ sibling quadruples. The basic analysis to determine the likelihood function
is substantially more complicated than for the cases described above, which could for
the most part be treated by variations of the calculations given in Feingold, Brown and
Siegmund (1993). The details are given in an appendix.

For $i, j = 2, 3, 6$ and $i \geq j$, let $Y_{ij}$ denote the indicator that $i$ pairwise comparisons
show identity by descent at the given locus on one chromosome (maternal or paternal)
and $j$ do so on the other (paternal or maternal). For the subscripts $ij = 33$ and 22, we
also introduce a third subscript $k = 0$, or 1 according as the sib pairs having identity by
descent on their paternal chromosome are different from or the same as the pairs having
identity by descent on their maternal chromosome. The results in the Appendix show
that the log likelihood function at the trait locus is

$$Y_{66} \log(1 + 6\alpha + \delta) + Y_{63} \log(1 + 3\alpha) + Y_{62} \log(1 + 2\alpha + \delta) + Y_{331} \log(1 - \delta)$$

$$+ Y_{32} \log(1 - \alpha) + Y_{330} \log(1 - \gamma) + Y_{220} \log(1 - 2\alpha + \gamma) + Y_{221} \log(1 - 2\alpha + \delta).$$
We denote by $Z_{1,t}, Z_{2,t}$ and $Z_{3,t}$ the derivatives of the log likelihood with respect to \( \alpha, \delta \) and \( \gamma \), respectively, evaluated at \( \alpha = \delta = \gamma = 0 \), and normalized to have unit variance under the hypothesis of no linkage, namely \( \alpha = \delta = \gamma = 0 \). The index \( t \) is introduced to denote the position along the chromosome at which we make this evaluation. Some calculation shows that

\[
Z_{1,t} = (3Y_{66} + 3Y_{63}/2 + Y_{62} - Y_{32}/2 - Y_{220} - Y_{221})/(3N/4)^{1/2},
\]

\[
Z_{2,t} = (8/N)^{1/2}(Y_{66} - Y_{331} + Y_{62} + Y_{221}),
\]

\[
Z_{3,t} = (8/3N)^{1/2}(Y_{62} - Y_{330} + Y_{220}).
\]

Some additional calculation shows that at the trait locus \( \tau \)

\[
E(Z_{1,\tau}) = (3N)^{1/2}\alpha, \quad E(Z_{2,\tau}) = (N/8)^{1/2}\delta, \quad E(Z_{3,\tau}) = (3N/8)^{1/2}\gamma;
\]

and at unlinked loci these processes are pairwise uncorrelated. It follows that an optimal linear combination of the three processes would have noncentrality \( N^{1/2}(3\alpha^2 + \delta^2/8 + 3\gamma^2/8)^{1/2} \); but also, as above, this linear combination depends on the unknown values of \( \alpha, \delta \) and \( \gamma \) so cannot be used as a test statistic.

Statistics that have been proposed for this situation are (i) to add the 6 pairwise comparisons in a quadruple as if they came from different pedigrees, (ii) to score 1 or 0 for each quadruple according as all six pairwise comparisons show identity by descent on at least one chromosome or not (Thomas, Skolnick and Lewis 1994), and (iii) the Whittemore and Halpern (1994b) statistic. Based on the general considerations given above for combining different classes of affected relatives one might also consider (iv) a linear combination of \( Z_1, Z_2 \) and \( Z_3 \) based on some hypothetical values of \( \alpha, \delta \) and \( \gamma \) in the expectation that it is not important to have the values exactly correct to have a reasonably efficient statistic.

Some reflection shows that (i) is equivalent to using (4.4) by itself. This has noncentrality \( (3N)^{1/2}\alpha \). In the limiting case of a rare trait allele, \( c_4 \gg c_3 \gg c_2 \), so \( \alpha, \delta \) and \( \gamma \) are about equal, and the efficiency of this statistic is \( 6/7 \approx 0.86 \). The statistic (ii) is

\[
Z_{4,t} = 64(Y_{66} + Y_{63} + Y_{62} - 15N/64)/7 \times (15N)^{1/2},
\]

\[ (4.5) \]
which has noncentrality \((N/15)^{1/2}(6\alpha + \delta/7 + 6\gamma/7)\). Its efficiency is \(98/105 \approx 0.93\) for a rare allele. The Whittemore-Halpern statistic is

\[
36Y_{66} + 18Y_{63} + 14Y_{62} - 3Y_{32} + Y_{330} - 6Y_{221} - 7Y_{220}
\]

standardized to have mean 0 and variance 1 under the hypothesis of no linkage. Its noncentrality parameter is

\[
N^{1/2}(504\alpha + 9\delta + 15\gamma)/4(5370)^{1/2};
\]

it is about 93% efficient in the rare allele case.

The linear combination

\[
(3^{1/2}Z_1 + Z_2/8^{1/2} + (3/8)^{1/2}Z_3)/(7/2)^{1/2}
\]

is 100% efficient in the rare allele case, when \(\alpha, \delta\) and \(\gamma\) are all about equal, but is less so otherwise. Typically \(\alpha\) is larger than \(\delta\), which is in turn larger than \(\gamma\). A statistic that would be optimal in the case that \(\alpha : \delta : \gamma = 1:0.7:0.5\) is to multiply the coefficients of \(Z_2\) and \(Z_3\) in (4.6) by 0.7 and 0.5 respectively (and restandardize the resulting statistic to have unit variance). Some numerical experimentation with this and other combinations showed that it works well over a range of conditions and is still more than 97% efficient in the rare allele case.

Table 3 contains numerical values of noncentrality parameters using the two allele model introduced above. They show that over a wide range of allele frequencies and penetrances the different statistics suggested above are all reasonably efficient, although (iii) and (iv) are slightly better than (i) and (ii). Under the most likely scenario that sibling quadruples form a relatively small part of our sample, it seems unlikely that the exact choice will make a significant difference. A numerical example is given below.

It is interesting to note the contrast with the cases of sibling triples and sibling pairs, where (i), (iii) and (iv) all lead to the same statistic, which is 100% efficient. For triples the efficiency of (ii) is 6/7, while for pairs its efficiency is only 2/3. (cf. Feingold, Brown and Siegmund 1993).
5. Combining pairs and pedigrees containing larger numbers of affecteds.

In this section we consider the problem of combining pedigrees containing different numbers and configurations of affecteds. We assume that most pedigrees are affected sib pairs, accompanied by smaller numbers of other kinds of pedigrees. The general principles given earlier for combining pedigrees can in principle be applied here, but we are no longer able to evaluate different strategies in terms of the single parameter $\lambda_O$. Instead we consider the two allele (per trait locus) model with varying allele frequencies, penetrances, and phenocopy frequencies. Of course, if we are willing to specify the mode of inheritance as well as allele frequencies and penetrances, we can compute optimal weights. Here we explore the possibility of developing some simple general rules that can be useful without making specific assumptions.

We begin with the case of a sample consisting of a number of sibling pairs with smaller numbers of sibling triples and quadruples. An interesting example in the recent literature is Schwab et al. (1995), who used twenty-four polymorphic markers in a sixty cM region of chromosome 6 in an effort to replicate a finding of linkage of schizophrenia. There were 48 sib pairs, 4 triples and 3 quadruples, which were analysed by making the 78 ($=48+12+18$) possible pairwise comparisons. Let $N_2$ denote the number of pairs, $N_3$ the number of triples, and $N_4$ the number of quadruples. Adding up all possible pairwise comparisons amounts to taking the statistic

$$[N_2^{1/2} \times (1.2) + (3N_3)^{1/2} \times (3.2) + (6N_4)^{1/2} \times (4.4)]/[N_2 + 3N_3 + 6N_4]^{1/2}. \quad (5.1)$$

The noncentrality parameter of (5.1) and the optimal noncentrality, hence the efficiency of (5.1) among statistics that combine (1.2), (3.2) and (4.4) are easily evaluated. Schwab et al. assume a large frequency of $p = 0.2$ for the trait allele and no phenocopies, so the assumed penetrance plays no role in the noncentrality parameter. (Risch (1990a) has argued on the basis of pedigree analysis that a multiplicative model might be appropriate for schizophrenia. This would justify consideration of the putative locus on chromosome 6 by itself as if the trait were monogenic.) For this large allele frequency sibling quadruples are essentially no more powerful than sibling triples, which in turn are only slightly more powerful than sibling pairs. Consequently the statistic (5.1), which is implicitly
predicated on the assumption that triples are three times as informative as pairs and quadruples six times as informative, puts substantially more weight on sibling triples and quadruples than the optimal weights, which are approximately in the ratios $1 : 2/3^{1/2} : 2/3^{1/2}$. It is, nevertheless, about 95% efficient. For different combinations of triples and quadruples, (5.1) might be less efficient, but a numerical search indicates that it is always at least 90% efficient, and the minimum efficiency occurs near the unlikely combination $N_2 = 48, N_3 = 0, N_4 = 16$.

From numerical examples given above it seems plausible that lighter weights on triples and quadruples would produce a more satisfactory overall statistic. For example, the ratios $1 : 2^{1/2} : 3^{1/2}$, found after some numerical experimentation, are more than 98% efficient for all the models of Table 3 and all sample sizes considered. If we assume that our sample contains only sibling pairs and quadruples, which presumably provides the opportunity for the greatest gains/losses of efficiency in extreme cases, we can repeat the analysis given earlier for comparing sibling pairs and first cousin pairs. The extremes of $c = 1$ and $c = 6^{1/2}$ correspond to the cases where quadruples are no more informative than pairs and six times as informative, respectively. The weight $\hat{c} = 3^{1/2}$ suggested above is more than 93% efficient over this range, while the minimum efficiency of (5.1) is 82%. As above these minimum efficiencies occur when the number of quadruples is about 1/3 the number of pairs. In the more likely event that the numbers of triples and quadruples are small compared to the number of pairs there is less room to improve on (5.1).

The other examples in Section 4 can similarly be incorporated into an overall statistic. For two siblings and a niece we found that we lost very little power by using only the aunt-niece comparison and ignoring the sibling comparison; and in many cases this triple is about as powerful as an avuncular pair. Hence the triples could simply be included with the other second degree relatives. We found that two siblings and a cousin are roughly as powerful as two independent first cousin pairs, so the statistic $Z_2$ could be multiplied by $(2N)^{1/2}$, the two in order to give it the weight of two comparisons, and then treated as a first cousin pair. Thus we consider each larger pedigree to be equivalent to a certain number of pairs of a particular kind. We then combine the "pairs," in which the larger
pedigrees have been imbedded, by the prescription suggested earlier in this paper.

To evaluate this overall strategy we considered a sample consisting of $N_1$ sibling pairs, $N_2$ sibling pairs with one affected parent, $N_3$ sibling triples, $N_4$ sibling pairs with an affected grandparent, $N_5$ second degree relatives, $N_6$ siblings with an affected offspring of one of the siblings, $N_7$ cousins and $N_8$ sibling pairs with an affected cousin. We treated the sibling triples as if they were three independent sibling pairs, the sibling pairs with an affected parent as if they were half siblings, the siblings with an affected offspring as if they were avuncular pairs. In addition each sibling pair with an affected grandparent was regarded as equivalent to three second degree relative pairs, and each pair of siblings with an affected cousin was regarded as two independent cousin pairs. The sibling pairs, second degree relative pairs, and cousin pairs were weighted according to the prescription given above with the value $\hat{\lambda}_0 = 4$. For the values $N_1 = N_2 = 20, N_3 = 10, N_4 = 5, N_5 = 7, N_6 = N_7 = N_8 = 5$ and for the models of Table 1, this statistic was always at least 93% efficient. Similar efficiencies were obtained for other sample sizes.

In Figure 2 we compare the power of a mixture of these pedigrees with the power of an equal number $N = \Sigma N_i$ of sib pairs. The total sample size was adjusted so that the power of the sib pairs is 0.5 in all cases. The relative sizes of the other pedigrees, $N_i/N$, were taken to be 0.4, 0.2, 0.12, 0.05, 0.08, 0.05, 0.05, 0.05 for $i = 1, \cdots, 8$, respectively. The models are from rows 3, 5, 6, 7, 10 and 12 of Table 1, ordered according to row, then according to locus within a row. With one exception the power of the mixture is more than 0.5, sometimes substantially. The exceptional case involved the second locus in row ten, where the pedigrees containing three affecteds give no more linkage information than a sibling pair. To increase the power of $N$ sib pairs from 0.5 to about 0.9 would require a sample roughly 80% larger, so in some cases the mixture of different pedigrees is quite valuable.

**Discussion**

We have shown that for a trait for which the penetrances are additive within loci (effectively dominant) and between loci (heterogeneous), affected pairs of different relations can be efficiently combined by consideration of the relative risk parameter $\lambda_0$. In
particular, an assumed value of 4 leads to a statistic that is least 93\% efficient for a wide range of true values of $\lambda_O$ and relative sample sizes of the different relative pairs involved in the study.

We have studied a few examples of pedigrees containing three and four affecteds. The power of these pedigrees usually cannot be assessed in terms of a a single parameter, so there is no simple statistic that is fully efficient; but in each case we exhibited one or more statistics that are very efficient over a broad range of conditions. In concentrating on statistics based on the likelihood function, we are able to incorporate assumptions about the mode of inheritance and allele frequencies to obtain quite efficient statistics, although this requires a case by case analysis. The Whittemore-Halpern (1994b) statistic is not based on the likelihood function, but it seems to perform well under a variety of conditions and has the advantage that there is a definite algorithm to compute it in all cases (except if we want to include unaffected pedigree members in the analysis).

Using a two allele model, we have shown that each of the larger pedigrees is often roughly as powerful as a particular number of affected pair pedigrees. This suggests an overall strategy where the larger pedigrees are given an appropriate weight, then combined with the identified relative pairs, which in turn are weighted according to the selected value of $\lambda_O$ for use in the overall statistic. Numerical examples show that the performance is robust with respect to the exact values of the weights chosen, and that the strategy is very efficient.

In the course of our studies we found that a pair of siblings together with a more distant relative can be particularly powerful. It may be useful to target such pedigrees for recruitment. This remark should be qualified with the reminder that the use of distant relatives may involve marker genotyping of several unaffected relatives to extract the best possible identity by descent information. If Genomic Mismatch Scanning (Nelson, et al. 1993) were to be used however, the total amount of laboratory effort would not increase. Not only does the GMS method as originally described in Nelson et al. involve only the affected individuals, but it is most easily implemented when the affecteds as a group are unilineally related.

A particularly interesting, difficult case is a heterogeneous trait involving, say, two
unlinked loci, where at one locus the trait susceptibility allele is relatively rare and the penetrance large while at the other the allele is more common and the penetrance comparatively small. In that case pedigrees containing more than two affecteds are particularly powerful for detecting linkage to the high penetrance locus; but the noncentrality at the low penetrance locus can be even smaller in pedigrees containing three or four affecteds than in similar pedigrees containing only an affected pair. The intuitive explanation is that if there are multiple affecteds in a pedigree, it is relatively unlikely that the trait is segregating at the low penetrance locus. Consequently the locus with the high penetrance low frequency allele will almost invariably be detected first, especially if our sample contains a substantial fraction of pedigrees having more than two affecteds. One might then consider a conditional search (Dupuis, Brown and Siegmund 1995), which involves conditioning on the observed data at the detected locus to obtain a larger noncentrality, hence greater power to detect the weaker locus. For example, for the fifth row of Tables 1 and 2 the sibs-grandparent triples are actually more powerful than three second degree relatives at the first locus, and comparable to a single pair at the second. Given that the first locus has been detected, it follows from Dupuis, Brown and Siegmund (1995) that for a conditional search the noncentrality for sibs-grandparent triples at the second locus is 0.33, while that for second degree relatives is 0.25. Hence sibs-grandparent triples receive a bigger boost from the conditional search than second degree pairs. Even so, for differences between the loci as large as this example and some of the others given in Tables 1 and 2, detection of the second locus will usually require a substantially larger sample size than detection of the first.

Acknowledgment

This research has been supported in part by NIH Grant HG00848.

Appendix

Here we perform some basic calculations for pedigrees having four affecteds and apply the results to the specific examples of half siblings, which are particularly simple, and the more interesting case of siblings. It is convenient to introduce the notation
$\tilde{\varphi}_i = E(\varphi_i | G_i)$. A straightforward calculation gives

$$
\lambda_{1...4} = K^{-4} E(\tilde{\varphi}_1 \tilde{\varphi}_2 \tilde{\varphi}_3 \tilde{\varphi}_4) = 1 + K^2 \sum_{i<j} \text{cov}(\tilde{\varphi}_i, \tilde{\varphi}_j)
$$

$$
+ K^{-3} \sum_{i<j<k} E[(\tilde{\varphi}_i - K)(\tilde{\varphi}_j - K)(\tilde{\varphi}_k - K)] + K^{-4} E[(\tilde{\varphi}_1 - K) \cdots (\tilde{\varphi}_4 - K)].
$$

There are three terms in the first sum and four terms in the second. To evaluate this expression we observe that as above $K^{-2} \text{cov}(\tilde{\varphi}_i, \tilde{\varphi}_j) = c_2 \epsilon_{ij}$. Here

$$
\epsilon_{ij} = P(X_i \equiv X_j) + P(X_i \equiv Y_j) + P(Y_i \equiv X_j) + P(Y_i \equiv Y_j),
$$

where $X_i(Y_i)$ denotes the allele received by individual $i$ from the homo (hetero)gametic parent and $\equiv$ denotes identity-by-descent. Similarly

$$
K^{-3} E[(\tilde{\varphi}_i - K)(\tilde{\varphi}_j - K)(\tilde{\varphi}_k - K)] = c_3 \epsilon_{ijk},
$$

where $\epsilon_{ijk} = P(X_i \equiv X_j \equiv X_k) + \cdots + P(Y_i \equiv Y_j \equiv Y_k)$ (eight terms). Finally

$$
K^{-4} E[(\tilde{\varphi}_1 - K) \cdots (\tilde{\varphi}_4 - K)] = c_4 \epsilon_{1234} + c_2^2 \epsilon_{1234}^*,
$$

where $\epsilon_{1234}$ is the sum of $2^4 = 16$ terms yielding the expected number of alleles that all four relatives inherit identical-by-descent and $\epsilon_{1234}^*$ is the sum of $3 \cdot 2^4 = 48$ terms of the form $P(X_1 \equiv X_2, X_2 \not\equiv X_3, X_3 \equiv X_4)$, i.e., the expected number of alleles that two pairs share identical-by-descent but are not shared by all four. Using these results, we obtain

$$
\lambda_{1...4} = 1 + c_2 \sum_{i<j} \epsilon_{ij} + c_3 \sum_{i<j<k} \epsilon_{ijk} + c_4 \epsilon_{1234} + c_2^2 \epsilon_{1234}^*.
$$

(A.1)

As a simple example we consider four half siblings having a common mother, but different fathers. Then $P(X_i \equiv X_j) = 1/2, P(X_i \equiv Y_j) = P(Y_i \equiv Y_j) = 0, P(X_i \equiv X_j \equiv X_k) = 1/4, P(X_i \equiv X_2 \equiv X_3 \equiv X_4) = 1/8, P(X_i \equiv X_j, X_j \not\equiv X_k, X_k \equiv X_l) = 1/8$, etc., so $\lambda_{1...4} = 1 + 3c_2 + c_3 + c_4/8 + 3c_2^2/8$. For full sibs except for $\epsilon_{1...4}^*$ the coefficients are doubled since there can be identity-by-descent on the maternally or paternally derived chromosome; thus $\lambda_{1...4} = 1 + 6c_2 + 2c_3 + c_4/4 + 9c_2^2/4$. For a second example we consider
a father (1), mother (2), and two children (3, 4). Then \(e_{12} = 0, e_{13} = e_{14} = e_{23} = e_{24} = e_{34} = 1, e_{134} = e_{234} = 1/2, e_{1234} = 0, e_{i1234} = 2\), so

\[\lambda_{1\ldots 4} = 1 + 5c_2 + c_3 + 2c_2^2.\]

**Remark.** For a heterogeneous trait another quantity similar to the final term in (A.1) appears. We assume that penetrances are additive within and between (two) unlinked loci. In evaluating \(K^{-4} E(\varphi_1 \varphi_2 \varphi_3 \varphi_4)\) we obtain in addition to the terms involving each locus separately a cross product term of the form

\[c_2c'_2(e_{12}e'_{34} + e_{13}e'_{24} + e_{14}e'_{23} + e'_{12}e_{34} + e'_{13}e_{24} + e'_{14}e_{23}).\]

For example, for four siblings all the \(e_{i,j}\) and \(e'_{i,j}\) equal 1, so

\[\lambda_{1\ldots 4} = 1 + 6(c_2 + c'_2) + 2(c_3 + c'_3) + (c_4 + c'_4)/4 + 9(c_2^2 + c'_2^2)/4 + 6c_2c'_2.\]

For half siblings the \(e_{i,j}\) and \(e'_{i,j}\) all equal 1/2. For two siblings and their parents

\[\lambda_{1\ldots 4} = 1 + 5(c_2 + c'_2) + c_3 + c'_3 + 2(c_2 + c'_2)^2.\]

For four affected half siblings, we denote by \(Y_i\) the number of pairs \(i = 2, 3, 6\) that are identical by descent at a given locus. We also let \(\Phi = \{\prod \varphi_s = 1\}\) denote the event that all relatives under discussion have the phenotype of interest. Calculations similar to those given above yield for \(\lambda_{1\ldots 4}^{(i)} = K^{-4} P[\Phi|Y_i = 1]\) the values \(\lambda_{1\ldots 4}^{(6)} = 1 + 6c_2 + 4c_3 + c_4, \lambda_{1\ldots 4}^{(3)} = 1 + 3c_2 + c_3, \lambda_{1\ldots 4}^{(2)} = 1 + 2c_2 + c'_2.\) From these results and Bayes formula one can evaluate the probability that \(i\) pairwise comparisons of four affected half siblings show identity by descent at the given locus. Letting \(Q_i = P\{Y_i = 1|\Phi\}\), we see that at the trait locus

\[Q_6 = 8^{-1} \lambda_{1\ldots 4}^{(6)}/\lambda_{1\ldots 4} = 8^{-1} \left[ \frac{1 + 6c_2 + 4c_3 + c_4}{1 + 3c_2 + c_3 + c_4/8 + 3c_2^2/8} \right].\]

Putting

\[\alpha = \frac{(c_2 + c_3)/2 + (c_4 - c_2^2)/8}{1 + 3c_2 + c_3 + c_4/8 + 3c_2^2/8}, \quad \delta = \frac{c_4/8 + 3c_2^2/8}{1 + 3c_2 + c_3 + c_4/8 + 3c_2^2/8},\]

we can rewrite this conditional probability as \(Q_6 = (1 + 6\alpha + \delta)/8.\) Similarly \(Q_3 = (1 - \delta)/2, \quad Q_2 = 3(1 - 2\alpha + \delta)/8.\)

36
The necessity to use two parameters to describe these conditional probabilities makes the case of four (or more) half siblings fundamentally more difficult than those of two and three half siblings, which required only a single parameter. The case of siblings is still more complicated. For \( i, j = 2, 3, 6 \) and \( i \geq j \), let \( Y_{ij} \) denote the indicator that \( i \) pairwise comparisons show identity by descent at the given locus on one chromosome (maternal or paternal) and \( j \) do so on the other (paternal or maternal), and put \( Q_{ij} = P\{Y_{ij} = 1|\Phi\} \). For the subscripts \( ij = 33 \) and \( 22 \), we also introduce a third subscript \( k = 0, \) or \( 1 \) according as the sib pairs having identity by descent on their paternal chromosome are different from or the same as the pairs having identity by descent on their maternal chromosome. In those cases we let \( Q_{ijk} = P\{Y_{ijk} = 1|\Phi\} \). We introduce the three parameters \( \alpha, \delta \) and \( \gamma \) defined by

\[
\alpha = \frac{c_2 + c_3 + (c_4 + c_2^2)/4}{D}, \quad \delta = \frac{c_4/4 + 9c_2^2/4}{D}, \quad \gamma = \frac{(c_4 + c_2^2)/4}{D},
\]

where \( D = 1 + 6c_2 + 2c_3 + c_4/4 + 9c_2^2/4 \). Then at the trait locus

\[
Q_{66} = \frac{(1 + 6\alpha + \delta)}{64}, \quad Q_{63} = \frac{(1 + 3\alpha)}{8}, \quad Q_{330} = \frac{3(1 - \gamma)}{16}, \quad Q_{331} = \frac{(1 - \delta)}{16},
\]

\[
Q_{82} = \frac{3(1 + 2\alpha + \gamma)}{32}, \quad Q_{220} = \frac{3(1 - 2\alpha + \gamma)}{32}, \quad Q_{221} = \frac{3(1 - 2\alpha + \delta)}{64}, \quad Q_{32} = \frac{3(1 - \alpha)}{8}.
\]

References


Feingold E and Siegmund D (1996) Linkage analysis for recessive and partially recessive 
traits using affected sibling pairs and inbred individuals. Unpublished manuscript
James J W (1971) Frequency in relatives for an all-or-none trait. Ann Hum Genet 35: 
47-48.
Kempthorne O (1957) An Introduction to Genetic Statistics. John Wiley and Sons, New 
York.
Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and non-parametric 
linkage analysis: a unified multipoint approach. Unpublished manuscript.
Kruglyak L, Lander ES (1995) Genetic dissection of complex traits: guidelines for inter-
preting and reporting linkage results. Nature Genetics 11: 241-247
Lander ES, Botstein D (1986) Strategies for studying heterogeneous genetic traits in 
humans by using a linkage map of restriction fragment length polymorphims. Proc 
Nat Acad Sci USA 83: 7353-7357.
versity Press, Baltimore.
Risch N (1990a,b,c) Linkage strategies for genetically complex traits I, II, III. The power 
Schwab SG, Albus M, Hallmayer J, Hönig S, Borrmann M, Lichtermann D, Ebstein RP, 
susceptibility gene for schizophrenia on chromosome 6p by multipoint affected sib-pair 
linkage analysis. Nature Genetics 11: 325-327
Sribney WM and Swift M (1992) Power of sib-pair and sib-trio linkage analysis with 
assortative mating and multiple disease loci. Am J Hum Genet 51: 773-784
IMA J of Math Appl in Med and Bio 11: 1-16


<table>
<thead>
<tr>
<th>$f_0$</th>
<th>$f$</th>
<th>$f'$</th>
<th>$p$</th>
<th>$p'$</th>
<th>$K$</th>
<th>Pairs 1</th>
<th>Pairs 2</th>
<th>Triples 1</th>
<th>Triples 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.50</td>
<td>0.50</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.33</td>
<td>0.33</td>
<td>0.53</td>
<td>0.53</td>
</tr>
<tr>
<td>0.01</td>
<td>0.50</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.02</td>
<td>0.61</td>
<td>0.00</td>
<td>1.05</td>
<td>0.00</td>
</tr>
<tr>
<td>0.00</td>
<td>0.50</td>
<td>0.50</td>
<td>0.02</td>
<td>0.01</td>
<td>0.03</td>
<td>0.42</td>
<td>0.21</td>
<td>0.66</td>
<td>0.34</td>
</tr>
<tr>
<td>0.00</td>
<td>0.50</td>
<td>0.25</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>0.43</td>
<td>0.21</td>
<td>0.82</td>
<td>0.21</td>
</tr>
<tr>
<td>0.01</td>
<td>0.50</td>
<td>0.25</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>0.38</td>
<td>0.19</td>
<td>0.75</td>
<td>0.20</td>
</tr>
<tr>
<td>0.05</td>
<td>0.50</td>
<td>0.00</td>
<td>0.05</td>
<td>0.00</td>
<td>0.10</td>
<td>0.39</td>
<td>0.00</td>
<td>0.65</td>
<td>0.00</td>
</tr>
<tr>
<td>0.03</td>
<td>0.50</td>
<td>0.25</td>
<td>0.01</td>
<td>0.04</td>
<td>0.06</td>
<td>0.21</td>
<td>0.20</td>
<td>0.43</td>
<td>0.24</td>
</tr>
<tr>
<td>0.00</td>
<td>0.50</td>
<td>0.50</td>
<td>0.05</td>
<td>0.05</td>
<td>0.10</td>
<td>0.25</td>
<td>0.25</td>
<td>0.35</td>
<td>0.35</td>
</tr>
<tr>
<td>0.00</td>
<td>0.50</td>
<td>0.50</td>
<td>0.10</td>
<td>0.05</td>
<td>0.15</td>
<td>0.28</td>
<td>0.15</td>
<td>0.37</td>
<td>0.19</td>
</tr>
<tr>
<td>0.00</td>
<td>0.50</td>
<td>0.25</td>
<td>0.05</td>
<td>0.10</td>
<td>0.10</td>
<td>0.31</td>
<td>0.14</td>
<td>0.50</td>
<td>0.15</td>
</tr>
<tr>
<td>0.05</td>
<td>0.50</td>
<td>0.25</td>
<td>0.05</td>
<td>0.10</td>
<td>0.15</td>
<td>0.21</td>
<td>0.10</td>
<td>0.38</td>
<td>0.12</td>
</tr>
<tr>
<td>0.00</td>
<td>0.50</td>
<td>0.00</td>
<td>0.10</td>
<td>0.00</td>
<td>0.05</td>
<td>0.49</td>
<td>0.00</td>
<td>0.67</td>
<td>0.00</td>
</tr>
<tr>
<td>0.00</td>
<td>0.25</td>
<td>0.00</td>
<td>0.20</td>
<td>0.00</td>
<td>0.05</td>
<td>0.35</td>
<td>0.00</td>
<td>0.45</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Table 2. Noncentrality parameters at each of two (unlinked) loci for various configurations of affecteds. The models are identical to those in Table 1. Sibs-1P denotes two affected sibs with one affected and one unaffected parent; 2nd Deg denotes a pair of second degree relatives; Sibs-O denotes two siblings and the offspring of one of them; Sibs-GP denotes two siblings and a grandparent; 1st C denotes a pair of first cousins; and Sibs-1st C denotes two siblings and their first cousin.

<table>
<thead>
<tr>
<th></th>
<th>Sibs-1P</th>
<th>2nd Deg</th>
<th>Sibs-O</th>
<th>Sibs-GP</th>
<th>1st C</th>
<th>Sibs-1st C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2</td>
<td>0.44 0.44</td>
<td>0.43 0.43</td>
<td>0.45 0.45</td>
<td>0.70 0.70</td>
<td>0.65 0.65</td>
<td>0.95 0.95</td>
</tr>
<tr>
<td>1 2</td>
<td>0.86 0.00</td>
<td>0.76 0.00</td>
<td>0.89 0.00</td>
<td>1.38 0.00</td>
<td>1.05 0.00</td>
<td>1.88 0.00</td>
</tr>
<tr>
<td>1 2</td>
<td>0.55 0.28</td>
<td>0.53 0.27</td>
<td>0.56 0.29</td>
<td>0.84 0.43</td>
<td>0.77 0.39</td>
<td>1.11 0.56</td>
</tr>
<tr>
<td>1 2</td>
<td>0.67 0.17</td>
<td>0.55 0.27</td>
<td>0.69 0.18</td>
<td>1.07 0.28</td>
<td>0.81 0.40</td>
<td>1.43 0.38</td>
</tr>
<tr>
<td>1 2</td>
<td>0.61 0.16</td>
<td>0.45 0.22</td>
<td>0.64 0.17</td>
<td>0.94 0.25</td>
<td>0.59 0.29</td>
<td>1.20 0.32</td>
</tr>
<tr>
<td>1 2</td>
<td>0.51 0.00</td>
<td>0.37 0.00</td>
<td>0.58 0.00</td>
<td>0.71 0.00</td>
<td>0.40 0.00</td>
<td>0.77 0.00</td>
</tr>
<tr>
<td>1 2</td>
<td>0.37 0.19</td>
<td>0.20 0.20</td>
<td>0.34 0.18</td>
<td>0.53 0.29</td>
<td>0.22 0.22</td>
<td>0.59 0.33</td>
</tr>
<tr>
<td>1 2</td>
<td>0.29 0.29</td>
<td>0.27 0.27</td>
<td>0.32 0.32</td>
<td>0.40 0.40</td>
<td>0.32 0.32</td>
<td>0.45 0.45</td>
</tr>
<tr>
<td>1 2</td>
<td>0.31 0.18</td>
<td>0.28 0.15</td>
<td>0.34 0.19</td>
<td>0.39 0.22</td>
<td>0.31 0.17</td>
<td>0.43 0.24</td>
</tr>
<tr>
<td>1 2</td>
<td>0.41 0.10</td>
<td>0.32 0.15</td>
<td>0.45 0.13</td>
<td>0.55 0.16</td>
<td>0.36 0.17</td>
<td>0.60 0.18</td>
</tr>
<tr>
<td>1 2</td>
<td>0.29 0.08</td>
<td>0.19 0.09</td>
<td>0.34 0.11</td>
<td>0.38 0.12</td>
<td>0.19 0.09</td>
<td>0.38 0.12</td>
</tr>
<tr>
<td>1 2</td>
<td>0.56 0.00</td>
<td>0.53 0.00</td>
<td>0.61 0.00</td>
<td>0.75 0.00</td>
<td>0.62 0.00</td>
<td>0.85 0.00</td>
</tr>
<tr>
<td>1 2</td>
<td>0.30 0.00</td>
<td>0.33 0.00</td>
<td>0.41 0.00</td>
<td>0.45 0.00</td>
<td>0.35 0.00</td>
<td>0.47 0.00</td>
</tr>
</tbody>
</table>
Table 3. Noncentrality parameters for a monogenic trait having the indicated penetrances and allele frequency. 6P is the noncentrality of six independent sibling pairs; Opt is the optimal noncentrality for a sibling quadruple; APW is the noncentrality of (4.4), i.e., the sum of the six pairwise comparisons of a quadruple; All is the noncentrality of the Thomas-Skolnick-Lewis statistic (4.5); W-H is noncentrality of the Whittemore-Halpern statistic; and WS is the noncentrality of the weighted score statistic with weights 0.7 and 0.5 as described in the text.

<table>
<thead>
<tr>
<th>f₀</th>
<th>f</th>
<th>p</th>
<th>K</th>
<th>λ₀</th>
<th>6P</th>
<th>Opt</th>
<th>APW</th>
<th>All</th>
<th>W-H</th>
<th>WS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.50</td>
<td>0.01</td>
<td>0.01</td>
<td>25.8</td>
<td>1.66</td>
<td>1.61</td>
<td>1.50</td>
<td>1.55</td>
<td>1.55</td>
<td>1.59</td>
</tr>
<tr>
<td>0.01</td>
<td>0.50</td>
<td>0.01</td>
<td>0.02</td>
<td>7.2</td>
<td>1.49</td>
<td>1.49</td>
<td>1.40</td>
<td>1.44</td>
<td>1.45</td>
<td>1.48</td>
</tr>
<tr>
<td>0.03</td>
<td>0.50</td>
<td>0.01</td>
<td>0.04</td>
<td>2.5</td>
<td>1.05</td>
<td>1.29</td>
<td>1.23</td>
<td>1.24</td>
<td>1.27</td>
<td>1.28</td>
</tr>
<tr>
<td>0.00</td>
<td>0.50</td>
<td>0.05</td>
<td>0.05</td>
<td>5.8</td>
<td>1.43</td>
<td>1.05</td>
<td>1.00</td>
<td>1.01</td>
<td>1.03</td>
<td>1.05</td>
</tr>
<tr>
<td>0.05</td>
<td>0.50</td>
<td>0.05</td>
<td>0.10</td>
<td>2.2</td>
<td>0.94</td>
<td>0.83</td>
<td>0.81</td>
<td>0.79</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>0.05</td>
<td>0.25</td>
<td>0.04</td>
<td>0.07</td>
<td>1.5</td>
<td>0.57</td>
<td>0.69</td>
<td>0.68</td>
<td>0.65</td>
<td>0.69</td>
<td>0.69</td>
</tr>
<tr>
<td>0.00</td>
<td>0.50</td>
<td>0.10</td>
<td>0.10</td>
<td>3.3</td>
<td>1.20</td>
<td>0.75</td>
<td>0.73</td>
<td>0.71</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>0.10</td>
<td>0.50</td>
<td>0.10</td>
<td>0.20</td>
<td>1.6</td>
<td>0.62</td>
<td>0.53</td>
<td>0.52</td>
<td>0.49</td>
<td>0.53</td>
<td>0.53</td>
</tr>
<tr>
<td>0.00</td>
<td>0.25</td>
<td>0.20</td>
<td>0.10</td>
<td>2.0</td>
<td>0.87</td>
<td>0.48</td>
<td>0.47</td>
<td>0.45</td>
<td>0.48</td>
<td>0.48</td>
</tr>
</tbody>
</table>
Figure Legends

Figure 1. Efficiency of the statistics suggested in the text for (i) sibling pairs with one affected parent □ and (ii) sibling pairs with an affected child ■; (iii) efficiency of the Whittemore-Halpern statistic for sibling pairs with one affected parent ◆ and (iv) efficiency of the usual sib pair statistic (2.1) when the sibling pairs have one affected parent ◊. The models are those of Table 1 ordered according to row and according to loci within rows. Two identical loci within a row are represented only once.

Figure 2. Comparison of sample sizes required for 90% power to detect linkage to Locus 1 for the models in Table 1. Horizontal axis gives the number of affected sibling pairs. Vertical axis gives (i) Sibling triples ■, (ii) Sibling pairs with one affected and one unaffected parent □, (iii) Sibling pairs with an affected grandparent ◆, and (iv) Sibling pairs with an affected first cousin ◊.

Figure 3. Power for the models in Table 1 ordered according to row and according to loci within each row. The relative numbers of the different pedigrees, taken from Tables 1 and 2, are given in the text. An equal number of pedigrees consisting entirely of pairs of siblings would have power of 50%. (Two identical loci within a row are represented only once.)