The Effect of Family Structure on Linkage Tests Using Allelic Association

J. C. Whittaker[1] and C. M. Lewis[2]

[1]Department of Applied Statistics, University of Reading, Reading, United Kingdom; and [2]Division of Medical and Molecular Genetics, United Medical and Dental Schools, Guy's Hospital, London

Summary

We considered the problem of testing a marker for linkage with a disease, using tests based on the transmission-disequilibrium test (TDT). The power of such tests was investigated for a number of possible family types, for which the families were classified by the disease status of family individuals. We show that parental disease status greatly affects the power, with families containing a single affected parent often preferred over families in which neither parent is affected. Families with a pair of affected sibs are of great value for all situations considered, but extension of the TDT to allow inclusion of information from unaffected sibs rarely increases power, if the parents have been genotyped.

Introduction

The transmission-disequilibrium test (TDT) has been used extensively to detect linkage disequilibrium, without the problems of ascertainment of appropriate populations of cases and controls for population-association studies (Spielman and Ewens 1996). With the increased availability of markers and genotyping capacity, this type of study will become more common.

The TDT has been used to narrow candidate regions identified, through linkage analysis, in a genome screen and to test polymorphisms in candidate genes; in the near future, genomewide screening for linkage disequilibrium may become feasible (Risch and Merikangas 1996). In this article, we consider the power of tests that are based on the TDT but that use different family structures.

The original formulation of the TDT used a single affected offspring and two parents, for whom affectation status was ignored by the test statistic. We show that, under a wide range of genetic models, the sampling of families in which one parent is affected can improve substantially the power of the TDT. In addition, we consider the contribution of a second sibling, either affected or unaffected. However, we must remember that, for independent families with a single affected child, the TDT tests the null hypothesis of no linkage *or* no allelic association between marker and disease loci (Spielman and Ewens 1996), and, therefore, the TDT often is described as a test for linkage in the presence of allelic association. In this article, we use the term "linkage disequilibrium" to imply the presence of both linkage and allelic association in the sample, and, therefore, we refer to the TDT as a test for linkage disequilibrium. For families containing multiple affected children, the situation is slightly more complicated, and this is discussed further in the Sib-Pair Data section.

Theory

We examined the power of tests based on the TDT, for a number of different familial patterns of disease. Consider a locus with two alleles, M and m, for which the allele frequencies are p and $1 - p$, respectively. We used f_{ij} to specify the influence of the locus on the disease, where, for $i, j =$ M, m, f_{ij} is the probability that an individual with genotype ij is affected by the disease. We assume that f_{ij} is the same for both the parental and the offspring generations. This is exactly true only if the locus is a candidate gene; if the locus is in fact a marker linked to a disease locus with, for instance, allele M associated with disease allele D, $f_{\mathrm{MM}} = P(\mathrm{affected}\,|\,\mathrm{MM})$, for example, can be seen as the combination of the penetrances at the disease locus and the association of M with D (Schaid 1996). Recombinations between parent and child result in the weakening of association between M and D and, therefore, cause f_{ij} to differ between the two generations. However, this effect is small over the genetic distances for which allelic association is maintained and can be ignored safely. Also note that the assumption that marker penetrances f_{MM}, f_{Mm}, and f_{mm} are constant over the two generations is needed only for power calculations: even if this assumption is vio-

Received February 11, 1998; accepted for publication July 9, 1998; electronically published August 21, 1998.

Address for correspondence and reprints: Dr. John C. Whittaker, Department of Applied Statistics, University of Reading, P.O. Box 240, Whiteknights Road, Reading RG6 2FN, United Kingdom. E-mail: j.c.whittaker@reading.ac.uk

0002-9297/98/6303-0031$02.00

lated, all the test statistics described here remain valid, since they have the specified null distribution.

We also assume that individuals can be diagnosed, without error, as affected or unaffected by the disease of interest, so that any problems due to variable age at onset of the disease are ignored. Most of the theoretical development of linkage tests using allelic association has concentrated on families with one affected child; we consider these families first and then examine the effect of extra sibs.

Table 1

Contingency Table of Transmitted and Nontransmitted Alleles

Transmitted	Nontransmitted Allele[a]	
Allele	M	m
M	*a*	*b*
m	*c*	*d*

[a] Letters *a*–*d* represent no. of parents.

One Affected Child

Suppose we have *n* unrelated families in our sample, each consisting of a single affected child and both parents. We ignore the problems caused by missing parental information and false paternity. For families $i = 1, 2, \ldots, n$, we define z_i to be 2 (−2) if both parents are heterozygous and transmit an M (m) allele, so that the affected child has genotype MM (mm); 1 (−1) if one parent is heterozygous and transmits an M (m) allele, while the other parent is homozygous; and 0 if both parents are homozygous or if both are heterozygous and transmit different alleles. It is easy to show that, under the null hypothesis of no linkage disequilibrium between the locus and the disease, $\mathbf{E}(z_i) = 0$ and $\mathrm{Var}(z_i) = h$, where *h* is the expected number of parents of an affected child who are heterozygous. Families are independent; therefore, asymptotically,

$$\frac{\sum_{i=1}^{n} z_i}{\sqrt{nh}} \sim N(0,\ 1)\ ,$$

under the null hypothesis of no linkage disequilibrium. Usually, *nh* is unknown and is estimated by consideration of the actual number of heterozygous parents in the sample, denoted as $b + c$ in the contingency table of transmitted and nontransmitted alleles (table 1). This gives the test statistic

$$T = \frac{\sum_{i=1}^{n} z_i}{\sqrt{b+c}}\ ,$$

and the squaring of *T* gives the familiar TDT $[(b-c)^2]/(b+c)$, with an asymptotic χ_1^2 distribution. Alternatively, the TDT can be viewed as the McNemar statistic for the testing of the equality of binomial proportions, after conditioning on the observed number of heterozygous parents in the sample.

To calculate the power of the test, we need the distribution of *T* under the alternative hypothesis. We define τ_{MmMM} to be the event that the father is heterozygous and transmits M to the child, while the mother is an MM homozygote. The other possible genotype/transmission events are written similarly as $\tau_{\mathbf{x}}$, where $\mathbf{x} = (x_1x_2x_3x_4) \in \mathbf{A}$, where $\mathbf{A}$ is the set of ordered parental genotypes $\mathbf{A} = \{\mathrm{MMMM}, \mathrm{MMMm}, \ldots, \mathrm{mmmm}\}$. Let $\mathrm{C_A}$ be the event that the child is affected. Under the alternative hypothesis, we get

$$\begin{aligned}\mathbf{E}(z_i) = {} & 2P(\tau_{\mathrm{MmMm}}|\mathrm{C_A}) + P(\tau_{\mathrm{MmMM}}|\mathrm{C_A}) + P(\tau_{\mathrm{Mmmm}}|\mathrm{C_A}) \\ & + P(\tau_{\mathrm{MMMm}}|\mathrm{C_A}) + P(\tau_{\mathrm{mmMm}}|\mathrm{C_A}) - P(\tau_{\mathrm{MMmM}}|\mathrm{C_A}) \\ & - P(\tau_{\mathrm{mmmM}}|\mathrm{C_A}) - P(\tau_{\mathrm{MmMM}}|\mathrm{C_A}) - P(\tau_{\mathrm{mMmm}}|\mathrm{C_A}) \\ & - 2P(\tau_{\mathrm{mMmM}}|\mathrm{C_A})\ ,\end{aligned}$$

with corresponding formulas for $\mathrm{Var}(z_i)$ and *h* derived easily. Under the assumptions of random mating and Hardy-Weinberg equilibrium and when the frequency of allele x_i is written as $P(x_i)$,

$$\begin{aligned}P(\tau_{\mathbf{x}}|\mathrm{C_A}) &= \frac{P(\mathrm{C_A}|\tau_{\mathbf{x}})P(x_1)P(x_2)P(x_3)P(x_4)}{\sum_{\mathbf{y}\in\mathbf{A}} P(\mathrm{C_A}|\tau_{\mathbf{y}})P(y_1)P(y_2)P(y_3)P(y_4)} \\ &= \frac{f_{x_1x_3}P(x_1)P(x_2)P(x_3)P(x_4)}{\sum_{\mathbf{y}\in\mathbf{A}} f_{y_1y_3}P(y_1)P(y_2)P(y_3)P(y_4)}\ ,\end{aligned}$$

by Bayes theorem. This enables us to calculate $\mathbf{E}(z_i)$, $\mathrm{Var}(z_i)$, and *h* for any penetrances and allele frequencies. Asymptotically, *T* has distribution

$$T \sim N\left[\frac{\sqrt{n}\mathbf{E}(z_i)}{\sqrt{h}}, \frac{\mathrm{Var}(z_i)}{h}\right]\ ,$$

and, therefore, this allows us to work out the power of the test. Note that, although the assumption of Hardy-Weinberg equilibrium is convenient for power calculations, Hardy-Weinberg equilibrium is *not* necessary in order for the tests to be valid.

The above discussion ignores any information on the disease status of the parents. For example, if $\mathrm{P_{AN}}$ indicates that the father has the disease but the mother does not, we can condition on parental disease status to get

$$P(\tau_{\mathbf{x}}|\mathrm{C_A},\ \mathrm{P_{AN}})$$

$$= \frac{f_{x_1x_3}f_{x_1x_2}(1-f_{x_3x_4})P(x_1)P(x_2)P(x_3)P(x_4)}{\sum_{\mathbf{y}\in\mathbf{A}}f_{y_1y_3}f_{y_1y_2}(1-f_{y_3y_4})P(y_1)P(y_2)P(y_3)P(y_4)}\ ,$$

as above, with similar expressions for $\mathrm{P_{AA}}$ and $\mathrm{P_{NN}}$. This allows us to compare the contributions, to power, of families in which neither, one, or both of the parents is affected by the disease, for any set of parameter values.

Multiplicative Disease Model

A useful and commonly used disease model (Self et al. 1991; Risch and Merikangas 1996; Schaid 1996) is given when haplotypes are assumed to act multiplicatively on the risk of disease, so that $f_{ab}=f_af_b$ for a, $b=\mathrm{M}$, m. Under this model, the alleles transmitted from the parents of an affected child are independent, and we can consider parents rather than families (Sham and Curtis 1995; Curnow et al. 1998). This allows us to examine analytically the value of affected parents. Let $\mathrm{P_A}$ and $\mathrm{P_N}$ indicate that a parent of an affected child is affected or unaffected, respectively, and let τ_{Mm} be the event that a heterozygous parent transmits an M allele to the child, with $\tau_{\mathbf{x}}$ defined similarly for $\mathbf{x}\in\mathbf{B}=$ {MM, Mm, mM, mm}. We wish to compare the values for affected and unaffected parents. We need to compare only the probability of heterozygosity for affected and unaffected parents, because, if we condition on on parental genotype, then parental disease status is irrelevant. Therefore, we define

$$\gamma = \frac{P(\text{parent heterozygous}\mid \mathrm{C_A},\ \mathrm{P_A})}{(\text{parent heterozygous}\mid \mathrm{C_A},\ \mathrm{P_N})}$$

$$= \frac{P(\tau_{\mathrm{Mm}}|\mathrm{C_A},\ \mathrm{P_A})+P(\tau_{\mathrm{mM}}|\mathrm{C_A},\ \mathrm{P_A})}{P(\tau_{\mathrm{Mm}}|\mathrm{C_A},\ \mathrm{P_N})+P(\tau_{\mathrm{mM}}|\mathrm{C_A},\ \mathrm{P_N})}\ .$$

Power increases with heterozygosity; therefore, if $\gamma>1$, affected parents are more valuable than unaffected parents and vice versa. Now,

$$P(\tau_{\mathrm{Mm}}|\mathrm{P_A},\ \mathrm{C_A}) = \frac{f_{\mathrm{M}}^2 f_{\mathrm{m}} pq}{\sum_{\mathbf{y}\in\mathbf{B}} f_{y_1}^2 f_{y_2} P(y_1)P(y_2)}\ ,$$

with similar expressions for $P(\tau_{\mathrm{mM}}\mid \mathrm{C_A},\ \mathrm{P_A})$, $P(\tau_{\mathrm{Mm}}\mid \mathrm{C_A},\ \mathrm{P_N})$, and $P(\tau_{\mathrm{mM}}|\mathrm{C_A},\ \mathrm{P_N})$. Thus,

$$\gamma = \left[f_{\mathrm{m}}f_{\mathrm{M}}\sum\nolimits_{\mathbf{y}\in\mathbf{B}}(1-f_{y_1}f_{y_2})f_{y_1}P(y_1)P(y_2)\right]\Big/$$

$$\left[(1-f_{\mathrm{M}}f_{\mathrm{m}})\sum\nolimits_{\mathbf{y}\in\mathbf{B}}f_{y_1}^2f_{y_2}P(y_1)P(y_2)\right]$$

$$=\{f_{\mathrm{m}}f_{\mathrm{M}}[f_{\mathrm{M}}(1-f_{\mathrm{M}}^2)p^2+(1-f_{\mathrm{m}}f_{\mathrm{M}})(f_{\mathrm{m}}+f_{\mathrm{M}})$$

$$\times p(1-p)+f_{\mathrm{m}}(1-f_{\mathrm{m}}^2)(1-p)^2]\}\Big/$$

$$\{(1-f_{\mathrm{M}}f_{\mathrm{m}})[f_{\mathrm{M}}^3p_2+f_{\mathrm{M}}f_{\mathrm{m}}(f_{\mathrm{m}}+f_{\mathrm{M}})$$

$$\times p(1-p)+f_{\mathrm{m}}^3(1-p)^2]\}$$

$$=\{f_{\mathrm{m}}f_{\mathrm{M}}[(1-f_{\mathrm{M}}^2)p+(1-f_{\mathrm{m}}^2)(1-p)]\}\Big/$$

$$(1-f_{\mathrm{m}}f_{\mathrm{M}})[f_{\mathrm{M}}^2p+f_{\mathrm{m}}^2(1-p)]\ ,$$

which is >1 if and only if $p<p^*=f_{\mathrm{m}}/(f_{\mathrm{M}}+f_{\mathrm{m}})$; that is, an affected parent is more informative than an unaffected parent if the associated allele is sufficiently rare. An alternative phrasing would be that an affected parent is more informative than an unaffected parent if an affected individual selected at random is more likely to be homozygous for the unassociated allele m than for the associated allele M. Note that we are assuming implicitly that $f_{\mathrm{M}}>f_{\mathrm{m}}>0$, so that $p^*<.5$; p^* is largest when f_{M} and f_{m} are of similar size, so that the locus has little effect on the risk of disease, and p^* is smallest when f_{M} and f_{m} are very different in size, so that the locus has a large effect on the risk of disease. An obvious special case is a fully penetrant recessive disease: then, $f_{\mathrm{M}}=1$, and $f_{\mathrm{m}}=0$. Thus, unaffected parents are always preferred. This is intuitively obvious: affected parents are MM homozygous and, therefore, can provide no information about allele transmission. Power calculations for multiplicative models are given in the Results and Discussion section, along with results for more-general disease models. Intuitively, we expect similar results for these more-general disease models: affected parents will be of most value when associated marker alleles are rare. This is discussed in more detail in the Results and Discussion section.

Sib-Pair Data

For families with more than one child, the above power calculations are insufficient, even if only one child from the family is included in the analysis, because the presence of other children changes the probabilities of the parents being heterozygous. For example, consider a fully penetrant recessive disease: in a family comprising one affected and three unaffected children, the probability that the father is heterozygous is higher than that for a family with a single affected child. Therefore, the formulas in the previous section represent “averages” over families of various types. In this section, we consider the value of families with two sibs, one affected and the other either affected or unaffected, by using tests based on the TDT.

Obviously, tests such as the TDT can be applied to families in which both sibs are affected: the TDT remains a test for linkage in the presence of allelic association; however, now, allelic association in the sample may be because of the common parentage of the affected sibs, rather than because of population allelic associations. Therefore, the TDT usually is described as a test for linkage when it is applied to pairs of affected sibs (Spielman and Ewens 1996). As expected, the use of affected sib pairs reduces the number of families required for a particular size and power (Risch and Merikangas 1996). Whether unaffected sibs are of any value in TDT-based tests is less obvious (Boehnke and Langefeld 1997), apart from the use of information on unaffected sibs to infer missing parental genotypes. Indeed, if unaffected sibs are to be included, the appropriate test statistic is not obvious. The most obvious statistic involves use of z_i, introduced above, but with the coding for unaffected sibs reversed, so that, for example, z_i is 2 if both parents are heterozygous and transmit an M allele to an affected child and -2 if both parents are heterozygous and transmit an M allele to an unaffected child. This gives the test statistic

$$T_{\mathrm{equal}}=\frac{\sum_{i=1}^{n}(z_{\mathrm{A}i}+z_{\mathrm{N}i})}{\sqrt{b+c}}\ ,$$

where $z_{\mathrm{A}i}$ and $z_{\mathrm{N}i}$ describe the parental transmissions to the affected and unaffected children in the ith family. However, this statistic clearly is not optimal, because transmissions to affected children will be more informative than transmissions to unaffected children and should be given a correspondingly higher weight in the test statistic.

Schaid (1996) derived score tests for families with a single affected child, under several disease models; in particular, he shows that the TDT is the score test for a biallelic marker, under the multiplicative disease model $f_{ij}=f_if_j$, and that this test often is preferable to the score test for the general disease model, because it uses fewer parameters (f_{M} and f_{m} in our notation) than the more general model (f_{MM}, f_{Mm}, and f_{mm}). Here, we derive a test statistic that includes information from unaffected children, by considering the locally most powerful test around the null hypothesis that the locus is not linked to the disease. We derive the test statistic for the multiplicative model $f_{ij}=f_if_j$.

Suppose we have n independent families with one affected and one unaffected child. Let $s_{\mathrm{A}i}$ be the number of M alleles in the affected sib of the ith family and $s_{\mathrm{N}i}$ be the number of M alleles in the unaffected sib, and let $\mathbf{g}_i=(g_{\mathrm{F}i},\ g_{\mathrm{M}i})$ describe the genotypes of the father and mother of the sibs. Under the disease model $f_{ij}=f_if_j$, the probability that the sibs having genotype $\mathbf{s}_i=(s_{\mathrm{A}i},\ s_{\mathrm{N}i})$, conditional on parental genotypes, is

$$P_1(\mathbf{s}_i|\mathbf{g}_i,\ \mathrm{C}_{\mathrm{AN}})$$
$$=\frac{f_{\mathrm{M}}^{s_{\mathrm{A}i}}f_{\mathrm{m}}^{2-s_{\mathrm{A}i}}(1-f_{\mathrm{M}}^{s_{\mathrm{N}i}}f_{\mathrm{m}}^{2-s_{\mathrm{N}i}})P(\mathbf{s}_i|\mathbf{g}_i)}{\sum_{\mathbf{y}\in\mathbf{S}}f_{\mathrm{M}}^{y_{\mathrm{A}}}f_{\mathrm{m}}^{2-y_{\mathrm{A}}}(1-f_{\mathrm{M}}^{y_{\mathrm{N}}}f_{\mathrm{m}}^{2-y_{\mathrm{N}}})P(\mathbf{y}|\mathbf{g}_i)}\ ,$$

where $\mathbf{S}$ is the set of possible genotypes for the sibs. The corresponding probability under the null hypothesis of no linkage is

$$P_0(\mathbf{s}_i|\mathbf{g}_i,\ \mathrm{C}_{\mathrm{AN}})=\frac{P(\mathbf{s}_i|\mathbf{g}_i)}{\sum_{\mathbf{y}\in\mathbf{S}}P(\mathbf{y}|\mathbf{g}_i)}\ .$$

By standard statistical theory (Cox and Hinkley 1974), the most powerful test for any set of parameter values is given by rejection of the null hypothesis, if

$$c<\frac{\Pi_{i=1}^{n}P_1(\mathbf{s}_i|\mathbf{g}_i,\mathrm{C}_{\mathrm{AN}})}{\Pi_{i=1}^{n}P_0(\mathbf{s}_i|\mathbf{g}_i,\mathrm{C}_{\mathrm{AN}})}\ ,$$

for c chosen to give the desired significance level. The denominator of this expression depends only on the parental genotypes; therefore, equivalently, we reject if

$$c_1<\sum_{i=1}^{n}\ln\,[P_1(\mathbf{s}_i|\mathbf{g}_i,\ \mathrm{C}_{\mathrm{AN}})\ ,$$

with c_1 determined by the parental genotypes and the required significance level. We derive an approximation for this expression that is valid near the null hypothesis, by setting $e^{\beta}=f_{\mathrm{M}}/f_{\mathrm{m}}$ so that

$$P_1(\mathbf{s}_i|\mathbf{g}_i,\ \mathrm{C}_{\mathrm{AN}})=\frac{e^{\beta s_{\mathrm{A}i}}(1-f_{\mathrm{m}}^2e^{\beta s_{\mathrm{N}i}})}{\sum_{\mathbf{y}\in\mathbf{S}}e^{\beta y_{\mathrm{A}i}}[1-f_{\mathrm{m}}^2e^{\beta y_{\mathrm{N}i}}P(\mathbf{y}|\mathbf{g}_i)]}\ .$$

Furthermore, $e^x\approx 1+x$, so that

$$P_1(\mathbf{s}_i|\mathbf{g}_i,\ \mathrm{C}_{\mathrm{AN}})$$
$$\approx\frac{(1+\beta s_{\mathrm{A}i})[1-f_{\mathrm{m}}^2(1+\beta s_{\mathrm{N}i})]}{\sum_{\mathbf{y}\in\mathbf{S}}(1+\beta y_{\mathrm{A}i})[1-f_{\mathrm{m}}^2(1+\beta y_{\mathrm{N}i})]P(\mathbf{y}|\mathbf{g}_i)}\ .$$

Around the null hypothesis, $\beta\approx 0$; we therefore can ignore β^2 terms, to get

$$P_1(\mathbf{s}_i|\mathbf{g}_i,\ \mathrm{C}_{\mathrm{AN}})$$
$$\approx\frac{1+\beta s_{\mathrm{A}i}-f_{\mathrm{m}}^2(1+\beta s_{\mathrm{A}i}+\beta s_{\mathrm{N}i})}{1+\beta\mathbf{E}_0(S_{\mathrm{A}i}|\mathbf{g}_i)-f_{\mathrm{m}}^2[1+\beta\mathbf{E}_0(S_{\mathrm{A}i}|\mathbf{g}_i)+\mathbf{E}_0(S_{\mathrm{N}i}|\mathbf{g}_i)]}\ ,$$

where $\mathbf{E}_0(S_{\mathrm{A}i}|\mathbf{g}_i)=\mathbf{E}_0(S_{\mathrm{N}i}|\mathbf{g}_i)$ are the expected genotypes

Table 2

Disease Models for Dominant, Recessive, Multiplicative, and Additive Models

Model	Model Type[a]	Disease Frequency	p_M	f_{MM}	f_{Mm}	f_{mm}	Attributable Risk
1	Dom	.100	.050	.76923	.76923	.02770	.72
2	Dom	.001	.050	.00513	.00513	.00055	.45
3	Dom	.100	.100	.13158	.13158	.09259	.07
4	Rec	.010	.400	.03125	.00595	.00595	.41
5	Rec	.001	.010	.50000	.00095	.00095	.05
6	Rec	.001	.075	.01778	.00091	.00091	.09
7	Rec	.001	.200	.01875	.00026	.00026	.74
8	Mult	.100	.125	.54903	.18936	.06531	.35
9	Mult	.010	.025	.28719	.04760	.00789	.21
10	Mult	.001	.025	.00421	.00200	.00095	.05
11	Add	.010	.150	.02745	.01719	.00692	.31
12	Add	.001	.050	.00421	.00252	.00083	.17

[a] Dom = dominant, Rec = recessive, Mult = multiplicative, and Add = additive.

of the offspring, given the parental genotypes. Using $\ln(1+x) \approx x$, we get

$$\ln[P_1(\mathbf{s}_i|\mathbf{g}_i,\ \mathrm{C}_{\mathrm{AN}})] \approx \beta\{(1-f_{\mathrm{m}}^2)[s_{\mathrm{A}i} - \mathbf{E}_0(S_{\mathrm{A}i}|\mathbf{g}_i)] - f_{\mathrm{m}}^2[s_{\mathrm{N}i} - \mathbf{E}_0(S_{\mathrm{N}i}|\mathbf{g}_i)]\}\ .$$

Thus, an approximation to the most powerful test near the null hypothesis is to reject the null hypothesis if

$$c_1 < \sum_{i=1}^{n}(1-f_{\mathrm{m}}^2)[s_{\mathrm{A}i} - \mathbf{E}_0(S_{\mathrm{A}i}|\mathbf{g}_i)] - f_{\mathrm{m}}^2[s_{\mathrm{N}i} - \mathbf{E}_0(S_{\mathrm{N}i}|\mathbf{g}_i)]\ ,$$

with c_1 determined by the parental genotypes and the required significance level. In fact, we use the equivalent test

$$c_2 < T_{\mathrm{sib}} = \frac{\sum_{i=1}^{n}\{(1-f_{\mathrm{m}}^2)[s_{\mathrm{A}i} - \mathbf{E}_0(S_{\mathrm{A}i}|\mathbf{g}_i)]\} - f_{\mathrm{m}}^2[s_{\mathrm{N}i} - \mathbf{E}_0(S_{\mathrm{N}i}|\mathbf{g}_i)]}{\sqrt{h[(1-f_{\mathrm{m}}^2)^2 + f_{\mathrm{m}}^4]}}\ ,$$

where h is the expected number of parents who are heterozygous, under the null hypothesis. This is a version of the TDT with transmissions from parents to affected offspring and to unaffected offspring, coded as above but weighted by $1 - f_{\mathrm{m}}^2$ and f_{m}^2, respectively. The expression contains two unknowns, h and f_{m}^2. We estimate h from the data, as for the usual TDT, and obtain a value for f_{m}^2 by assuming that the population prevalence of the disease is f_{m}^2. Under the null hypothesis, $T \sim N(0,1)$, asymptotically; under the alternative hypothesis, the distribution of T can be determined as described above, by calculation of $P(\mathbf{g},\ \mathrm{s}|\mathrm{C}_{\mathrm{AN}})$, and, thus, power can be calculated. Again, this is easily extended to take into account the disease status of the parents.

Note that, although T_{sib} is derived under the assumptions of a multiplicative model and n independent families, the test statistic remains valid as a test for linkage, in the sense that it has the specified distribution, for any disease model and for more-general family structures. However, tests derived under the correct disease model probably will have more power than T_{sib}. In this sense, T_{sib} resembles the TDT; Schaid and Sommer (1994) showed that the TDT can be derived as the score test under the multiplicative model, and they derived corresponding test statistics for recessive and dominant models. Schaid and Sommer (1994) also showed that these test statistics can have greater power than the TDT if the mode of inheritance of the disease is recessive or dominant.

Table 3

Results for Families Ascertained on the Basis of a Single Affected Child, by Parental Disease Status

	No. of Families Needed, by Parental Status				Population Frequency, by Parental Status		
Model	NN	AN	AA	XX	NN	AN	AA
1	108	45	126	64	.3774	.5737	.0488
2	134	87	118	134	.9963	.0037	.0000
3	4,163	3,694	3,401	4,062	.8080	.1818	.0102
4	181	171	242	181	.9752	.0246	.0002
5	562	148	202	559	.9976	.0024	.0000
6	564	310	258	563	.9978	.0022	.0000
7	39	48	349	39	.9938	.0062	.0000
8	258	196	158	240	.7643	.2199	.0158
9	260	118	77	253	.9705	.0293	.0002
10	2,121	1,442	1,092	2,120	.9979	.0021	.0000
11	394	339	317	392	.9774	.0224	.0001
12	481	313	258	480	.9978	.0022	.0000

Note.—NN = two unaffected parents, AN = one affected and one unaffected parent, and AA = two affected parents. "XX" indicates that the disease status of each parent was not considered when the family was ascertained.

Results and Discussion

The formulas given above allowed us to compare the value of families with different patterns of disease status, under a variety of genetic models. We considered families ascertained on the basis of a single affected child, two affected children, or one affected and one unaffected child, and, for each case, we considered the power for 0, 1, or 2 affected parents. Calculations were performed for a large number of genetic models; for the sake of brevity, we give the results for a relatively small number of models. The general conclusions given below apply to all models considered.

The disease models used (table 2) were parameterized in terms of the frequency of allele M and the genotype penetrances, f_{MM}, f_{Mm}, and f_{mm}. Four classes of models were considered: dominant, recessive, multiplicative, and additive; the additive model has $f_{ij} = f_i + f_j$ for $i, j = \mathrm{M, m}$. We also determined the attributable risk—$(K - f_{\mathrm{mm}})/K$, where K is disease prevalence—which is the proportion of cases that can be attributed to the increased risk conferred by allele M (e.g., see Boehnke and Langefeld 1998). The remaining proportion of cases reflects environmental causes and the influence of other genetic loci.

Results are displayed as the number of families of each type required in order to obtain a power of 0.8 and a size of 5×10^{-8} (tables 3–5). This size and power were chosen by Risch and Merikangas (1996) to be appropriate for a genome scan: since we were interested mainly in a candidate gene, a rather higher power and smaller size would have been more suitable, but we retained the Risch and Merikangas (1996) values, to allow comparison with their article. The pattern of the results was the same for other type 1 and type 2 errors, although, of course, the number of families changed greatly. In practice, studies would not be restricted to a single family type, but use of this restriction gives a guide to the power contributed by each type of family.

Table 3 shows the results for families with a single affected child, table 4 for families with one affected and one unaffected child, and table 5 for affected sib pairs. For families with one affected and one unaffected child, we considered three possible test statistics: T_{only}, the usual TDT test statistic, which uses only transmissions to the affected sib, and T_{sib} and T_{equal}, derived above. For each situation, we also give the number of families required for a power of 0.8 and a size of 5×10^{-8} if the families are ascertained solely on the basis of a single affected child, ignoring parental disease status. We also provide the frequencies of the various parental disease types, conditional on the specified offspring; for example, the frequencies of the parental disease configurations in table 3 are conditional on the family having a single affected child.

Table 4

Results for Families Ascertained on the Basis of One Affected and One Unaffected Child, by Parental Disease Status

Model and Statistic	No. of Families Needed, by Parental Status				Population Frequency, by Parental Status		
	NN	AN	AA	XX	NN	AN	AA
1:							
T_{sib}	129	38	87	64	.4577	.5081	.0341
T_{equal}	112	33	54	55			
T_{only}	146	43	107	73			
2:							
T_{sib}	134	87	118	134	.9963	.0037	.0000
T_{equal}	274	181	244	273			
T_{only}	134	87	118	134			
3:							
T_{sib}	4,149	3,676	3,380	4,048	.8082	.1816	.0102
T_{equal}	6,645	5,887	5,412	6,483			
T_{only}	4,215	3,734	3,434	4,112			
4:							
T_{sib}	182	171	241	182	.9753	.0245	.0002
T_{equal}	352	332	468	351			
T_{only}	182	171	241	182			
5:							
T_{sib}	661	184	189	658	.9977	.0023	.0000
T_{equal}	990	230	230	986			
T_{only}	661	184	189	658			
6:							
T_{sib}	568	312	259	567	.9978	.0022	.0000
T_{equal}	1,103	610	508	1,101			
T_{only}	568	312	259	567			
7:							
T_{sib}	39	48	346	39	.9938	.0062	.0000
T_{equal}	78	100	683	79			
T_{only}	39	48	346	39			
8:							
T_{sib}	267	197	154	247	.7733	.2124	.0143
T_{equal}	382	270	194	349			
T_{only}	275	205	162	255			
9:							
T_{sib}	266	120	78	259	.9709	.0289	.0002
T_{equal}	496	223	138	482			
T_{only}	266	121	78	259			
10:							
T_{sib}	2,122	1,442	1,092	2,121	.9979	.0021	.0000
T_{equal}	4,235	2,879	2,180	4,231			
T_{only}	2,122	1,442	1,092	2,121			
11:							
T_{sib}	394	339	317	393	.9775	.0224	.0001
T_{equal}	773	664	621	770			
T_{only}	394	339	317	393			
12:							
T_{sib}	481	313	258	480	.9978	.0022	.0000
T_{equal}	964	629	519	963			
T_{only}	481	313	258	480			

Note.—NN = two unaffected parents, AN = one affected and one unaffected parent, and AA = two affected parents. "XX" indicates that the disease status of each parent was not considered when the family was ascertained.

Table 5

Results for Families Ascertained on the Basis of Two Affected Children, by Parental Disease Status

	No. of Families Needed, by Parental Status				Population Frequency, by Parental Status		
Model	NN	AN	AA	XX	NN	AN	AA
1	26	24	74	26	.2207	.7018	.0775
2	45	42	65	45	.9946	.0054	.0000
3	1,879	1,688	1,573	1,838	.8058	.1837	.0105
4	71	81	152	71	.9682	.0316	.0003
5	11	16	133	11	.9809	.0190	.0001
6	46	35	61	46	.9963	.0037	.0000
7	16	29	363	16	.9878	.0122	.0000
8	90	81	74	87	.7027	.2712	.0262
9	52	37	28	51	.9465	.0527	.0007
10	722	503	386	721	.9979	.0021	.0000
11	171	161	159	171	.9745	.0253	.0002
12	159	120	108	159	.9974	.0026	.0000

Note.—NN = two unaffected parents, AN = one affected and one unaffected parent, and AA = two affected parents. "XX" indicates that the disease status of each parent was not considered when the family was ascertained.

Note that families with one affected and one unaffected child for which only the affected child was used in the analysis often gave lower power than families for which only a single affected child was used, because the presence of an unaffected sib increases the chance that an affected child is a nongenetic case. Thus, T_{sib} and T_{equal} in table 4 should be compared with T_{only}, rather than with the results in table 3. First, we examine the effect of parental disease status on families with a single affected child. For multiplicative models, we showed above that, if the disease allele is sufficiently rare, affected parents are more valuable than unaffected parents. For the multiplicative models given in table 2 (models 8–10), affected parents are always preferred, because disease-allele frequencies are below this rarity threshold. Note that the sample-size reductions gained by selection of affected parents can be considerable; for example, for model 9, samples of 77 families with both parents affected, 118 families with a single affected parent, or 260 families with neither parent affected are required in order to obtain the specified size and power.

A similar trend can be seen for other, nonmultiplicative disease models: affected parents are of the most value when the disease allele is rare. There is, however, one crucial difference: because parental transmissions are no longer independent, families with a single affected parent possibly may be more powerful than families with 0 or 2 affected parents. This trend is depicted in figure 1, which plots power for the disease model $f_{\mathrm{MM}} = .1$, $f_{\mathrm{Mm}} = .07$, and $f_{\mathrm{mm}} = .01$ and a range of allele frequencies. For rare alleles ($p < .03$), families with two affected parents are optimal; for allele frequencies of >~.16, families with no affected parents are optimal; and, for frequencies in the intervening range, families with a single affected parent are favored.

Families with a single affected parent seem to be optimal for dominant disease models, in most cases (see models 1 and 2). The fact that the presence of an affected parent increases power for dominant models is not surprising: for most dominant models, affected parents are more likely than unaffected parents to be Mm heterozygotes. However, the same argument would lead us to expect families with two affected parents to be more valuable than families with one affected and one unaffected parent. This is not always true, as can be easily seen when a simple, fully penetrant dominant disease is considered. If the allele frequency is low, affected parents will be heterozygotes, and unaffected parents will be mm homozygotes. In a family with a single affected parent, this parent will transmit an M allele to the affected child. In a family with two affected parents, one parent will transmit an M allele to the affected child, but the other parent is equally likely to transmit an M or an m allele, thus weakening the evidence for association between the disease and the M allele. Similar arguments apply to other dominant models, when the locus has a considerable effect on disease susceptibility. Thus, for models 1 and 2, families with a single affected parent are optimal, whereas for model 3, in which the locus is less influential, two affected parents are optimal.

For recessive and additive models (models 4–7, 11, and 12), the position is less clear. Families with 0, 1, or 2 affected parents can be optimal. Sampling of families with a single affected parent often results in a worthwhile reduction in sample size, compared with sampling of families solely on the basis of an affected child. Note that families with two affected parents are always rare

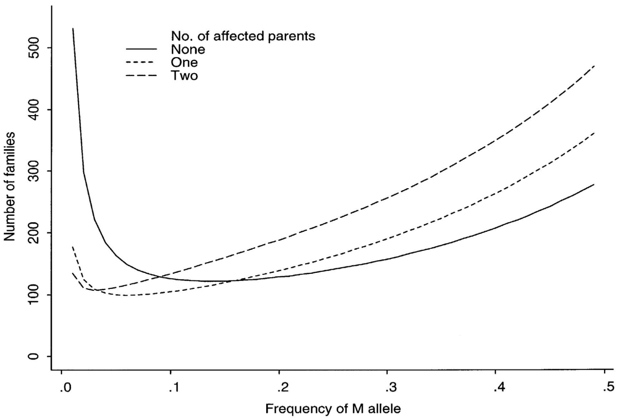

Figure 1 No. of families needed for power of 0.8 and size of 5×10^{-8}, plotted against frequency of the M allele, for families with 0, 1, and 2 affected parents and for disease model f_{MM} =, $f_{\mathrm{Mm}} = .07$, and $f_{\mathrm{mm}} = .01$.

but that families with a single affected parent are much more common, particularly for common diseases.

The results for sib-pair data also are reasonably easy to interpret (tables 4 and 5). The most obvious result is the high power, for nearly all the disease models, of affected–sib-pair families, compared with that for single-affected families. This already has been noted, by Risch and Merikangas (1996), for the special case of multiplicative disease models. Our results indicate that inclusion of unaffected offspring in the analysis can, but in general does not, result in extra power (models 1, 3, and 8) and never reduces the required sample size enough so that genotyping of the unaffected sib is worthwhile. We stress that this is true only if full parental genotypes are available. In the absence of parental genotypes, unaffected sibs indeed may be of value (Curtis 1997; Boehnke and Langefeld 1998; Spielman and Ewens 1998).

Of the test statistics used, T_{equal} gives the largest power increases when unaffected sibs are of value but, as expected, performs very badly in most cases. T_{sib} gives less dramatic results: usually, the power is comparable to that obtained when the unaffected child is ignored and, in a few cases, improves slightly (e.g., model 3). Again, this was expected, because T_{sib} tends to put little weight on transmissions to the unaffected child. Further work on alternative weightings of information from affected and unaffected sibs (Thompson 1997) has confirmed the lack of value of unaffected sibs when parental information is available. Intuitively, this is because the probability of a heterozygous parent transmitting an M allele to an unaffected child remains ~.5 for a broad class of disease models (Spielman and Ewens 1998), so that the number of M alleles transmitted to unaffected offspring varies little from its expected value under the null hypothesis.

For affected sib pairs, the relationship between parental disease status and power seems to be broadly similar to that for families with a single affected child, discussed above. Differences do arise in, for example, multiplicative models. The threshold gene frequency at which affected parents cease to be preferred is lower for pairs of affected sibs than for singletons, so that affected parents possibly may be preferred for families with a single affected child, and unaffected parents may be preferred for families with affected sib pairs. Similar results occur for other types of models (e.g., model 5).

Conclusions

We considered the influence of family structure on the power of the TDT, for a sample of affected children for whom full parental data were available. Theoretical models have been presented for the multiplicative model, and simulation has been used to cover a wide range of dominant, recessive, and additive models. The results given in this article reflect the full range of models tested, and, although the absolute numbers of families required varied greatly, some generalizations regarding the relative power of different family structures may be made. For many of the models tested, ascertainment of parent-offspring trios with one affected parent resulted in a substantial increase in power. Furthermore, for the models in which an affected parent was not advantageous, the frequency of affected parents often was so low that these types of families would be difficult to identify; therefore, sample availability could be used as an effective guide to the most powerful TDT families to collect. The availability of an affected parent will depend on the trait under consideration. For a late-onset or a highly lethal trait, these family structures will be rare. However, for common and less severely debilitating diseases, such as asthma or diabetes, these families will be easily accessible. Two affected siblings are clearly the most efficient family unit and provided dramatic reductions in sample sizes in many of the models considered. Within these families, ascertainment of affected parents can be valuable and follows the same pattern as that seen for families with a single affected offspring. However, given a choice between sampling families with two affected offspring and sampling those with an affected parent, the value of the additional affected sibling outweighs the value of the affected parent, in each case.

Acknowledgments

We are grateful to Robert Curnow, Andrew Morris, and Chris Mathew, for their helpful discussions of this work, and to two anonymous reviewers, whose comments improved the paper. This work was supported, in part, by grant HG00571 from the National Institutes of Heath.

References

Boehnke M, Langefeld CD (1997) A transmission/disequilibrium test that uses both affected and unaffected offspring. Am J Hum Genet Suppl 61:A269

——— (1998) Genetic association mapping based on discordant sib pairs: the discordant-alleles test. Am J Hum Genet 62:950–961

Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman & Hall, London

Curnow RN, Morris AP, Whittaker JC (1998) Locating genes involved in human diseases. Appl Stat 47:63–76

Curtis D (1997) Use of siblings as controls in case-control association studies. Ann Hum Genet 61:319–333

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Schaid DJ (1996) General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 13:423–449

Schaid DJ, Sommer SS (1994) Comparison of statistics for candidate-gene association studies using cases and parents. Am J Hum Genet 55:402–409

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

Sham PC, Curtis D (1995) An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. Ann Hum Genet 59:323–336

Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 59:983–989

——— (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450–458

Thompson DJ (1997) Using information from unaffected sibs in genetic association studies. MS dissertation, University of Reading, Reading, United Kingdom