Exploiting Excess Sharing: A More Powerful Test of Linkage for Affected Sib Pairs than the Transmission/Disequilibrium Test

Jacqueline Wicks

Centre for Mathematics and its Applications, Australian National University, Canberra; and Department of Mathematics, University of Queensland, Brisbane

The transmission/disequilibrium test (TDT) is a popular, simple, and powerful test of linkage, which can be used to analyze data consisting of transmissions to the affected members of families with any kind pedigree structure, including affected sib pairs (ASPs). Although it is based on the preferential transmission of a particular marker allele across families, it is not a valid test of association for ASPs. Martin et al. devised a similar statistic for ASPs, T_{sp}, which is also based on preferential transmission of a marker allele but which is a valid test of both linkage and association for ASPs. It is, however, less powerful than the TDT as a test of linkage for ASPs. What I show is that the differences between the TDT and T_{sp} are due to the fact that, although both statistics are based on preferential transmission of a marker allele, the TDT also exploits excess sharing in identity-by-descent transmissions to ASPs. Furthermore, I show that both of these statistics are members of a family of "TDT-like" statistics for ASPs. The statistics in this family are based on preferential transmission but also, to varying extents, exploit excess sharing. From this family of statistics, we see that, although the TDT exploits excess sharing to some extent, it is possible to do so to a greater extent—and thus produce a more powerful test of linkage, for ASPs, than is provided by the TDT. Power simulations conducted under a number of disease models are used to verify that the most powerful member of this family of TDT-like statistics is more powerful than the TDT for ASPs.

The transmission/disequilibrium test (TDT) (Spielman et al. 1993) is a popular, simple and powerful test of linkage. It has the property that, for data consisting of a random sample of transmissions from parents to a single affected child, it is a valid test of both linkage and association. The fact that association is required in order for linkage to be detected is an apparent weakness—but also a strength, because it means that the TDT may be helpful in refining the localization of disease-susceptibility genes. The argument supporting this claim is that various methods of linkage analysis may indicate that there is a disease-susceptibility gene in what can be a fairly large region; however, linkage disequilibrium is maintained over time in populations only if linkage is very tight. Thus, significant results for the TDT may indicate that the marker in question is very tightly linked to the disease-susceptibility locus.

However, for data consisting of a random sample of affected sib pairs (ASPs), it is well known that significant results for the TDT provide evidence for linkage, but not for association (Spielman and Ewens 1996). To establish the presence of association with the TDT—and thus use it to refine the localization of disease-susceptibility genes—it is necessary to discard transmissions to one member of each ASP.

It is therefore important to develop for ASPs a statistic that would use the information on transmissions to both members of an ASP and that would provide a valid test of both linkage and association. Martin et al. (1997) have devised such a statistic, which they have called "T_{sp}." It is similar to the TDT and is easy to calculate. To compare T_{sp} and the TDT, I define the following counts for ASP data. For parents with heterozygous marker genotype (1,2), let n_{11} be the number who transmit allele 1 to both of their children, let n_{22} be the number who transmit allele 2 to both of their children, and let n_{12} be the number who transmit allele 1 to one child and transmit allele 2 to the other child.

Received January 11, 2000; accepted for publication April 10, 2000; electronically published April 28, 2000.

Address for correspondence and reprints: Jacki Wicks, Department of Mathematics, University of Queensland, Brisbane QLD 4072, Australia. E-mail: wicks@maths.anu.edu.au

0002-9297/2000/6606-0033$02.00

Table 1

Approximate False-Positive Error Rates under the Null Hypothesis of No Linkage

	Approximate False-Positive Error Rate[a] for		
n	$T_{\text{sp}} = T(0)$	$\text{TDT} = T(\frac{1}{2})$	$T(1)$
50	.0540	.0569	.0602
100	.0494	.0546	.0524
200	.0510	.0513	.0517

[a] At 5% nominal significance, based on 10,000 simulated data sets.

When this notation is used, the statistic T_{sp} and the TDT for ASPs are given by, respectively,

$$T_{\text{sp}} = \frac{(n_{11} - n_{22})^2}{n_{11} + n_{22}}$$

and

$$\text{TDT} = \frac{(n_{11} - n_{22})^2}{\frac{1}{2}(n_{11} + n_{22} + n_{12})} \ . \qquad (1)$$

Simulation can be used to verify that T_{sp} is a valid test of both linkage and association, whereas the TDT is a valid test of linkage but not of association, for ASPs (see table 3 below).

The TDT is, however, a more powerful test of linkage for ASP data than is T_{sp}. I argue that the reason for this is that, when the TDT is applied to ASP data, it utilizes *excess sharing*—that is, the tendency for $n_{11} + n_{22}$ to exceed n_{12} in the presence of linkage. To see this, I note that

$$\text{TDT} = T_{\text{sp}} \times \frac{n_{11} + n_{22}}{\frac{1}{2}(n_{11} + n_{22} + n_{12})} \ .$$

So the TDT when applied to ASP data can be written as a product of T_{sp} and a factor that is a measure of excess sharing. The presence of linkage alone, without association, results in a tendency toward excess sharing. Therefore, positive test results for the TDT will sometimes be attributable to the presence of excess sharing when T_{sp} alone is not large enough to provide significant evidence for the presence of association in addition to linkage.

I argue that, in choosing a test of *linkage* for ASPs, it is possible to exploit excess sharing to a greater extent than is possible with the TDT—and thereby produce a more powerful test. The key to this lies in considering the denominators of the TDT applied to ASP data and of T_{sp} given in equation (1). Under the null hypothesis of no linkage (i.e., a recombination fraction of $\frac{1}{2}$), the probabilities associated with the categories defined by n_{11}, n_{22}, and n_{12} are, respectively, $\frac{1}{4}$, $\frac{1}{4}$, and $\frac{1}{2}$. Thus, under the null hypothesis of no linkage, we would expect to see $n_{11} + n_{22} \approx n_{12}$ and, thus, roughly similar values for both the TDT and T_{sp}. With this as motivation, I define the family of TDT-like statistics for ASPs,

$$T(\alpha) = \frac{(n_{11} - n_{22})^2}{(1 - \alpha)(n_{11} + n_{22}) + \alpha\ n_{12}} \ ,$$

for $0 \leqslant \alpha \leqslant 1$. We can see that T_{sp} and the TDT for ASPs are special cases corresponding to $\alpha = 0$ and $\alpha = \frac{1}{2}$, respectively.

Furthermore, it can be shown by means of standard statistical theory that, under the null hypothesis of no linkage, $T(\alpha)$ is asymptotically distributed as a χ^2 random variable with 1 df, for all α ($0 \leqslant \alpha \leqslant 1$). To verify this, data were simulated under the null hypothesis of no linkage, to determine the false-positive error rates for $T(\alpha)$ with $\alpha = 0$, $\frac{1}{2}$, and 1. With $n = n_{11} + n_{22} + n_{12}$ used to denote the sample size (i.e., the number of independent heterozygous parents), 10,000 data sets for $n =$ 50, 100, and 200 were generated. The approximate false-positive error rates for a nominal significance level of 5% are given in table 1. These results confirm that, for the different sample sizes and values of α, the statistic $T(\alpha)$ has a false-positive error rate consistent with an asymptotic χ^2 distribution with 1 df.

Therefore, in the choice of a test of linkage for ASP data, it would seem most prudent to choose the value of α for which the statistic $T(\alpha)$ is most powerful. I have already argued that, because $\text{TDT} = T(\frac{1}{2})$ utilizes excess sharing, its power exceeds that of $T_{\text{sp}} = T(0)$. Furthermore, it can be seen that the α value that results in the greatest use of excess sharing is not $\alpha = \frac{1}{2}$ (i.e., the TDT) but, rather, $\alpha = 1$. Indeed, we can write

$$T(\alpha) = T(0) \times \frac{n_{11} + n_{22}}{(1 - \alpha)(n_{11} + n_{22}) + \alpha\ n_{12}} \ .$$

This demonstrates that $T(\alpha)$ equals $T(0)$ ($=T_{\text{sp}}$) multiplied by a factor that is a measure of excess sharing.

Table 2

Disease Models Used in Power Simulations

Model	Disease-Allele Frequency	Penetrances[a]
Classical dominant	.001	1, 1, 0
Common dominant	.1	1, 1, .1
Common recessive	.1	1, .1, .1
Multiplicative	.1	.9, .3, .1
Additive	.1	.3, .2, .1

[a] Probabilities that one will be affected, given homozygosity for the disease allele, heterozygosity for the disease allele, and absence of the disease allele, respectively.

Table 3

Approximate Power under Complete Linkage and Varying Degrees of Association

	Power[a] for			
Association Status and Model	$T_{sp} = T(0)$	TDT = $T(\frac{1}{2})$	$T(1)$	$T(1)$ without IBD Information[b]
Absent:				
Classical dominant	.0494	.1149	.2658	.2128
Common dominant	.0538	.0813	.1326	.1182
Common recessive	.0530	.0690	.1006	.0920
Additive	.0540	.0599	.0689	.0669
Multiplicative	.0554	.0672	.0919	.0842
Moderate:				
Classical dominant	.5396	.6835	.8244	.7856
Common dominant	.3510	.4268	.5174	.4962
Common recessive	.1856	.2197	.2700	.2595
Additive	.0844	.0872	.1001	.0970
Multiplicative	.2284	.2548	.2980	.2900
Maximum:				
Classical dominant	.9976	.9995	1.0000	.9998
Common dominant	.9236	.9454	.9636	.9583
Common recessive	.5561	.5935	.6395	.6400
Additive	.1862	.1873	.2078	.2057
Multiplicative	.6891	.7109	.7446	.7384

[a] At 5% nominal significance, based on 10,000 simulated data sets. Each data set consists of 50 independent nuclear families with two affected children each.

[b] Under the assumption that there is no additional IBD information available to resolve transmissions in ambiguous families (see the text).

This factor is largest when $\alpha = 1$. Thus, in this family of statistics,

$$T(1) = (n_{11} - n_{22})^2/n_{12}$$

is the most powerful test of linkage and, in particular, is more powerful than the TDT.

Simulation was used to verify that the power of $T(1)$ exceeds that of the TDT. Data were generated for a biallelic marker with equally frequent alleles completely linked (i.e., with a recombination fraction of 0) to a disease-susceptibility locus. Several disease models were used, with the parameter values given in table 2. In each case, 10,000 data sets consisting of 50 independent nuclear families each with two affected children were simulated. Table 3 contains the results for data generated under complete linkage and varying degrees of association. The level of association was controlled by setting different values for the linkage-disequilibrium parameter δ, as in the work by McGinnis (1998): for the simulations under which association was absent, it was set equal to 0; for the simulations under moderate association, it was set equal to half of its maximum possible value; and, for the simulations under maximum association, it was set equal to its maximum possible value.

An important practical issue in the analysis of nuclear families with $T(\alpha)$ concerns those families in which both parents and both affected children have heterozygous marker genotype (1,2). In the absence of additional identity-by-descent (IBD) information, it is not clear whether the contribution of such a family to the data counts n_{11}, n_{22}, and n_{12}, should be $n_{11} = n_{22} = 1$ or $n_{12} = 2$. Additional IBD information that will help to resolve transmissions in these families will be available either if the marker locus is a multiallelic one for which the alleles have been pooled or if other nearby marker loci have been typed. For families with any other genotype configuration, it is possible to determine the contribution to the data counts n_{11}, n_{22}, and n_{12} directly from the family members' genotypes at the marker locus. The families for which transmission is ambiguous do not affect calculation of TDT = $T(\frac{1}{2})$; however, for all other values of α, $T(\alpha)$ is affected. The results in table 3, both for $T_{sp} = T(0)$ and for the first column of results for $T(1)$, are calculated under the assumption that IBD transmission in such families can be determined.

Appropriate strategies can be devised to handle the ambiguous families when transmissions cannot be determined on the basis of additional IBD information. One possible strategy is to set the contribution of each ambiguous family equal to the contribution expected under the null hypothesis of no linkage. Thus, these families would be scored as $n_{11} = n_{22} = \frac{1}{2}$ and $n_{12} = 1$. Simulation was used to verify that $T(1)$ has the correct asymptotic false-positive error rate under the null hypothesis of no linkage when this strategy is used (results not shown). When transmissions cannot be resolved in ambiguous families, and it becomes necessary to use this strategy, $T(1)$ necessarily loses power. The question of interest, then, is how the power of $T(1)$ in this circumstance compares with that of TDT = $T(\frac{1}{2})$. Table 3 includes results for $T(1)$ calculated under the assumption that no additional IBD information is available and that families in which both parents and both affected children have marker genotype (1,2) are scored according to the strategy given above.

The results for the simulations in table 3 demonstrate a number of important features. First, the simulations in which association is absent confirm that $T_{sp} = T(0)$ has the correct false-positive error rate under no association and that it is thus a valid test of association. These simulations also confirm that, for $\alpha > 0$, $T(\alpha)$ is not a valid test of association. These results and those for the simulations in which association is present also demonstrate the increase, in the power to detect linkage, that is achieved by the utilization of excess sharing for $T(\alpha)$ when $\alpha > 0$. In particular, the use of $T(1)$ provides an increase in power over the TDT = $T(\frac{1}{2})$, as a test of linkage for ASPs. The extent of the increase in power varies according to the disease model.

The results in table 3 also confirm that, when transmission in ambiguous families cannot be resolved by additional IBD information, and when the aforementioned strategy for handling these families is employed, the power of $T(1)$ is still greater than that of the TDT. The reason for this is as follows. The ambiguous families affect both the TDT and $T(1)$ in the same way, because the scoring of these families does not change the numerator of either and increases the denominator of both by the same amount. Therefore, these families have the same effect on the TDT and $T(1)$, and, for these families, the latter loses the advantage that it has over the TDT. However, $T(1)$ still has an advantage over the TDT, because it is still able to exploit excess sharing in other families, to a greater extent than does the TDT. Thus, it will always be more powerful than the TDT, regardless of whether transmission in the ambiguous families can be resolved by additional IBD information.

Thus, we see that $T(1)$ is the most powerful test of linkage for ASPs from the family of TDT-like statistics given by $T(\alpha)$ for $0 \leqslant \alpha \leqslant 1$. The statistic $T_{sp} = T(0)$ provides the baseline and is a valid test of both linkage and association for ASPs. All members of the $T(\alpha)$ family are based on preferential transmission of a particular marker allele; however, increasing values of α also result in increasing utilization of excess sharing and thus more powerful tests of linkage for ASPs. The statistic TDT = $T(\frac{1}{2})$ goes only part way in the utilization of excess sharing; the statistic $T(1)$ utilizes it to the fullest extent possible.

Finally, I would note that the method used to obtain the power advantage of $T(1)$ can be applied to the TDT when data consist of a combination of ASPs and other family types. This means that it is possible to exploit the excess sharing among ASPs when one is analyzing data comprising ASPs and other family types. For example, consider a data set that is a combination of some ASPs and their parents and some affected singletons and their parents. For the families with affected singletons, I define the following data counts. For parents with heterozygous marker genotype (1,2), let n_1 be the number who transmit allele 1 to their affected child and let n_2 be the number who transmit allele 2 to their affected child.

The TDT can be used to combine different family types to give an overall test of linkage (Spielman et al. 1993). In particular, families with ASPs and families with affected singletons can be combined. In this case, the TDT is given by

$$\mathrm{TDT} = \frac{(n_1 + 2n_{11} - n_2 - 2n_{22})^2}{n_1 + n_2 + 2n_{11} + 2n_{22} + 2n_{12}} \ .$$

The power of this test can be improved by exploiting, in a way similar to that outlined above for $T(1)$, the excess sharing among the ASPs. This yields the statistic

$$\frac{(n_1 + 2n_{11} - n_2 - 2n_{22})^2}{n_1 + n_2 + 4n_{12}} \ ,$$

which will be a more powerful test of linkage than is the TDT. This approach can be used to increase the power of the TDT for any combination of family types that includes some ASPs and their parents. Furthermore, it is an open question whether, to obtain further increases in power, it is possible to exploit the sharing relationships in sibships with more than two affected children.

Acknowledgments

I am grateful to Warren Ewens, Eden Martin, Terry Speed, and Sue Wilson for interesting discussions as well as helpful comments on an earlier draft of this report. Thanks also are due to two anonymous reviewers whose suggestions improved this report significantly.

References

Martin ER, Kaplan NL, Weir BS (1997) Tests for linkage and association in nuclear families. Am J Hum Genet 61:439–448

McGinnis RE (1998) Hidden linkage: a comparison of the affected sib pair (ASP) test and transmission/disequilibrium test (TDT). Ann Hum Genet 62:159–179

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

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516