TDT with Covariates and Genomic Screens with Mod Scores: Their Behavior on Simulated Data

John P. Rice, Rosalind J. Neuman, Stacy L. Hoshaw, E. Warwick Daw, and Chi Gu

Department of Psychiatry, Washington University School of Medicine, St. Louis, Missouri

We describe an extension to the TDT (transmission/disequilibrium test) which allows for more than two marker alleles and for covariates measured on the parent or offspring. We also describe a systematic genomic search where the mod score (maximized lod score) is computed for each marker under constraints on the population prevalence or penetrances of a single locus. © 1995 Wiley-Liss, Inc.

Key words: transmission/disequilibrium test, mod score, linkage analysis

INTRODUCTION

Association between a particular genetic marker and disease offers an alternative to linkage analysis for the identification of disease genes—although the relative merits of association versus linkage studies have been controversial. Localization of disease genes once linkage is detected will be difficult for complex traits. The linkage for Huntington's disorder, a genetically simple disease, was reported in 1983, and the locus was identified 10 years later; the identification of susceptibility loci for oligogenic traits will be much more difficult. In association studies, mutations in candidate genes may be tested in cases versus controls; this approach is facilitated by advances in DNA sequencing technology. In contrast, systematic linkage screening is done with simple sequence repeat polymorphisms which *a priori* are not expected to be causative in the disease studied. Association studies using anonymous markers are limited since the disease and marker alleles may be in equilibrium even though they are linked.

A critical consideration for association studies is the choice of control groups. Since population stratification is well known to cause a (genetically) spurious association, the

Address reprint requests to Dr. John P. Rice, Washington University Medical School, Department of Psychiatry (Box 8134), 4940 Children's Place, St. Louis, MO 63110.

© 1995 Wiley-Liss, Inc.

haplotype relative risk method (HRR) has become a popular alternative using the untransmitted parental alleles as a control sample. In this design, the sampling unit consists of independent cases and their parents. Spielman and colleagues [1993] proposed the TDT (transmission/disequilibrium test) which is a McNemar's test for matched samples. The various HRR methods have recently been reviewed by Schaid and Sommer [1994].

Linkage analysis for oligogenic traits under the assumption of a single locus is problematic [Rice et al., 1993]. One approach is to maximize the lod score (mod), as suggested by Risch [1984], Clerget-Darpoux et al. [1986], and Greenberg [1989]. Elston [1989] noted that maximizing the lod score is equivalent to maximizing the likelihood of the marker data conditional on all phenotypic data. In cases where segregation analysis under the wrong model would lead to "meaningless" parameter estimates, conditioning on the phenotypic information has intuitive appeal.

The present GAW data provide an opportunity to evaluate the utility of the TDT and of genomic searches using the mod score. These are two distinct approaches which use different types of information in the data provided.

METHODS

Generalized TDT

For a marker M with n alleles 1,...,n consider the probabilities π_{ij} with corresponding observations n_{ij}, where i refers to the transmitted allele and j to the nontransmitted allele. Here, if N cases and parents are observed, there are 2N observations. The hypothesis of no association is then parameterized as equality of marginals: $\pi_{1.} = \pi_{.1}, \pi_{2.} = \pi_{.2},..., \pi_{n.} = \pi_{.n}$. This is the standard test for paired categorical data [Grizzle et al., 1969] and implemented in the CATMOD procedure of SAS [1989]. In the case of two alleles, this corresponds to the TDT statistic of Spielman and colleagues [1993].

The elements of the vector $F(\pi) = (\pi_{1.},...,\pi_{n.},\pi_{.1},...,\pi_{.n})$ are called response functions, and the model is given as $F(\pi) = X\beta$, where X is the design matrix containing fixed constants and β is a vector of parameters to be estimated. For example, with two alleles,

$$\begin{pmatrix} \pi_{1.} \\ \pi_{.1} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

yield

$$\begin{aligned} \pi_{1.} &= \beta_1 + \beta_2 \\ \pi_{.1} &= \beta_1 - \beta_2 \end{aligned}$$

If $\beta_2 = 0$, then

$$\pi_{1.} = \pi_{.1},$$

so that

$$\pi_{11} + \pi_{12} = \pi_{11} + \pi_{21},$$

or

$$\pi_{12} = \pi_{21}.$$

Thus testing $\beta_2 = 0$ is equivalent to the usual TDT statistic.

The HRR methods require collapsing alleles into the dichotomy m and $\bar{m}$ (not m). For a marker with n alleles, this involves n tests which are not independent. The generalized TDT (GTDT) first tests for overall significance of a χ^2 with n-1 degrees of freedom to avoid statistical complications associated with testing the alleles separately. Tests on individual β's can then be used to determine which alleles impart different risks.

The other advantage of the GTDT is that covariates such as the sex or disease status of the parent may be used in CATMOD. There is, however, one complication. If, for example, both parents and the child are heterozygotes for the same two alleles (e.g., the father, mother, and child are all genotype 1,3), then we know one parent transmitted the "1" and not the "3", and vice versa, but there is ambiguity as to which parent did which. However, when adjacent markers are available, this ambiguity may be resolved so that attributes of the parent may be considered even in these ambiguous cases. See the Appendix for SAS code for the GTDT.

Mod Scores

We have modified the program ILINK [Lathrop et al., 1984] to allow maximization of the lod score. In this program (MODLINK) the maximization may be performed under a fixed prevalence (where the gene frequency is calculated in terms of penetrances f_1, f_2, and f_3 and the fixed prevalence K_p).

RESULTS

Marker Associations

We first compared marker allele frequencies in the parents of controls to the frequencies in 200 cases (one affected child from each ascertained family). We found seven markers to be significant at the p = 0.01 level. We next applied the 2 × 2 TDT statistic to all markers (for n alleles, n nonindependent tests were done). This test generated 15 tests significant at the 0.01 level. As noted above, this approach is problematic due to the multiple tests which are not independent. The GTDT test found four markers significant at the 0.01 level. We restricted further analysis to the three markers in Table I (D5G23, D1G31, D5G10) which were found to be significant in both the case-control and GTDT comparisons.

We collapsed the number of alleles to two and considered covariates of sex of the parent, sex of the offspring, and whether or not the parent was affected. The matched odds ratios for the alleles at these three markers were 3.17, 4.14, and 0.42, respectively. If a disease is recessive, we would expect affected parents not to differentially transmit an associated allele, whereas if a disease is dominant, we expect the affected parent to give a higher odds ratio. For marker D5G23, we found an odds of 3.6 for unaffected parents to transmit allele 7, and of 1.4 for affected parents (χ^2_1 = 34.6), indicating recessive-like transmission. The only significant covariate was "affected parent" for marker D5G23.

Mod-score Analyses

We used the program MODLINK to perform a genomic screen for all 360 markers. mod scores greater than 2 in any setting are displayed in Table II.

Because of the strong association detected at D5G23 (allele 7), we also divided the data according to whether the first affected offspring (FAO) was homozygous 7,7 at

TABLE I. Tests of Association

	Gene frequency in cases vs. parents of controls		TDT[a]	GTDT		Covariate
Marker	χ^2(df)	Allele	χ_1^2	χ^2(df)	Allele	Affected parent χ_1^2
D5G23	58.0 (7)	7	50.5	65.6 (7)	7	34.6
D1G31	29.9 (7)	8	27.9	34.0 (7)	8	—
D5G10	19.4 (4)	1	9.6	11.62 (3)	1	—

[a] TDT test for allele in previous column.

TABLE II. Mod Scores from Genomic Screens

Marker	$K_p = 2.3\%$, $f_3 = 0$	K_p free, $f_3 = 0$
D1G48	2.1	2.3
D1G55	3.0	3.1
D3G16	2.1	3.0
D3G23		2.6
D3G39		2.3
D4G30		2.3
D5G41		2.5

D5G23. Among the 160 families in which the FAO was not 7,7 were 109 families in which the FAO did not have the 8 allele at D1G31. A further subset of 29 of the 109 families had no 7 or 8 allele. We conducted a genome search on each of these three subsets of families and their complements (i.e., the remaining families in the total data sets). Results are displayed in Table III.

Post Hoc Interpretation

The true generating model included the two associated loci (D5G23 and D1G31) detected in Table I. However, none of the four trait loci were detected using linkage methods. That is, the above positive mod scores represent Type I error, and underscore the impact of multiple hypothesis tests in genomic screens. The 360 tests done are, however, not independent, so the expected number of significant tests is not given by 360 times the chosen critical value. Asymptotically, one would expect the tests in the two columns of Table II to have 3 (estimating θ, f_1, f_2) and 4 (estimating θ, K_p, f_1, f_2) degrees of freedom, respectively. Using the number of mod scores above 2, this would indicate an equivalent number of independent tests given by 3/0.027 = 111 and 7/0.056 = 125, respectively, where 0.027 and 0.056 are the p-values associated with a lod score of 2.

Table III represents a further subdivision and explanation which resulted in a mod score of 4.4 for marker D1G55. It should be emphasized that evidence for linkage to these markers was present in the data provided, but not in the population from which they were sampled.

TABLE III. Mod Scores Conditioned on Genotypes of the First Affected Offspring (FAO) at Associated Markers

Marker	Not 7,7 (N = 160)	7,7 (N = 40)	Not 7,7 and Not 8 (N = 109)	7,7 or 8 (N = 91)	Not 7 and Not 8 (N = 29)	7 or 8 (N = 171)
D1G55	4.4	0.0	2.7	0.9	1.4	1.9
D1G57	0.2	0.0	0.5	0.0	3.9	0.0
D3G16	2.2	0.5	2.1	0.9	0.7	1.6
D4G38	0.3	0.2	0.0	0.3	2.8	0.1
D6G22	1.4	0.6	0.0	2.2	0.1	2.0

Not 7,7: FAO is not homozygous 7,7 at D5G23.
7,7: FAO is homozygous 7,7 at D5G23.
Not 7,7 and Not 8: FAO is not 7,7 at D5G23 and does not have the 8 allele at D1G31.
7,7 or 8: Complement of (not 7,7 and not 8).
Not 7 and Not 8: FAO does not have the 7 allele at D5G23 nor the 8 allele at D1G31.
7 or 8: Complement of (not 7 and not 8).

DISCUSSION

We found the two highly significant associations with markers D5G23 and D1G31, without clear evidence for linkage to these chromosomal regions. This in part reflects the information available. There were 155 simplex families out of the 200 families provided, and they would provide little information for linkage, although they contributed to the GTDT computations.

We calculated average mod scores on data simulated by Suarez et al. [this issue] which consisted of 200 families with at least two affected siblings out of four. (They used 159 in their analysis.) They generated phenotypes determined by the effect of four loci modeled as loci A, B, C, and D in the GAW9 simulations (but without associations). Our mean mod scores were 0.82, 2.39, 2.45, and 3.29, respectively. The mod score for a dummy unlinked locus was 0.12. Accordingly, our inability to detect the true linkages in the GAW9 data reflects the information available in the data provided, and not inherent difficulties in linkage analysis.

The false mod scores above 2 and 3 indicate a basic concern in genomic screens with no candidate genes or known mode of inheritance for the disease. In humans, a 10 cM map would require approximately 360 markers. Our analyses indicate that these tests may behave almost independently in widely spaced maps, so that false positive results from multiple tests are of concern.

The data do indicate the ability of the GTDT statistic to provide compelling evidence for association with matched odds ratios of 3 or 4. Thus the strategy of first testing for associations between complex traits and candidate loci may be a good one given our results. Even in the multiplex families simulated, the average mod scores for loci A and B were only 0.82 and 2.39. Our GTDT approach, allowing for an arbitrary number of alleles and the incorporation of covariates, should be useful in evaluating such associations.

ACKNOWLEDGMENTS

This study was supported in part by USPHS grants MH37685, HG00201, AA08401, MH49500, and MH31302 from the National Institute of Mental Health and training grant T32 DA07261 from NIDA for Dr. Daw.

REFERENCES

Clerget-Darpoux F, Bonaïti-Pellié C, Hochez J (1986): Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42:393-399.

Elston RC (1989): Man bites dog? The validity of maximizing lod scores to determine mode of inheritance. Am J Med Genet 34:487-488.

Greenberg DA (1989): Inferring mode of inheritance by comparison of lod scores. Am J Med Genet 34:480-486.

Grizzle JE, Starmer CF, Koch GG (1969): Analysis of categorical data by linear models. Biometrics 25:489-504.

Lathrop GM, Lalouel JM, Julier C, Ott J (1984): Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci USA 81:3443-3446.

Rice JP, Neuman RJ, Hampe CL, Daw EW, Suarez BK (1993): Linkage analysis for oligogenic traits. Am J Hum Genet 53(Suppl):66.

Risch N (1984): Segregation analysis incorporating linkage markers. I. Single locus models with an application to type I diabetes. Am J Hum Genet 36:363-386.

SAS Institute Inc. (1989): SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 1, Cary NC.

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

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.

APPENDIX

For variables T (transmitted allele) and NT (nontransmitted allele), the SAS code for the GTDT is

```
proc catmod;
response marginals;
model T*NT=_response_  {covariates}/freq;
repeated hap 2;
title 'test of catmod';
quit;
```