Biometrical Journal **45** (2003) 3, 349–366

Estimation of Linkage and Association from Allele Transmission Data

James D. Malley[1, *] Richard A. Redner[2], Thomas A. Severini[3], Judith A. Badner[4], Sinisa Pajevic[1], and Joan E. Bailey-Wilson[5]

[1] Center for Information Technology, National Institutes of Health
[2] Department of Mathematical and Computer Sciences, University of Tulsa, OK
[3] Department of Statistics, Northwestern University, IL
[4] Department of Psychiatry, University of Chicago, IL, and
[5] National Human Genome Research Institute, National Institutes of Health USA

Summary

The *TDT* provides a hypothesis test for the presence of linkage or association (linkage disequilibrium). However, since the *TDT* is a single test statistic, it cannot be used to separate association and linkage. The importance of this difficulty, following a significant *TDT* result, has been recently emphasized by Whittaker, Denham and Morris (2000), who alert the community to the possibility that a significant *TDT* may result from loose linkage and strong association, *or* from tight linkage and weak association.

To attack this problem we start with the parametric model for family-based allele transmission data of Sham and Curtis (1995) (or Sham (1998)) and find that the parameters in the model are not always identifiable. So we introduce a reparameterization that resolves the identifiability issues and leads to a valid likelihood ratio (*LR*) test for linkage.

Since the linkage and association parameters are both of interest, we next introduce and apply an integrated likelihood (*IL*) approach to provide separate point estimates and confidence intervals for these parameters. The estimates are shown to have generally small bias and mean square error, while the confidence intervals have good average length and coverage probabilities. We compare the power of the *IL* approach for testing linkage and, separately association, with the *TDT* and *LR*.

Key words: TDT; Parameter estimation; Association; Linkage; Integrated likelihood.

1. Introduction

Since its introduction, the transmission/disequilibrium test (*TDT*) of Spielman, McGinnis and Ewens (1993, 1994) has achieved widespread use in genetics for testing for association and/or linkage of a marker gene with a putative disease gene. There are now many variants and extensions of the *TDT*. Many of these are discussed in the informative series by Schaid (1996, 1998, 1999), which is also especially helpful in understanding the comparison of the *TDT* with case-control methods for measuring presence of association. See also Knapp (1999), and Wilson (1997) for further discussion and extensions of the *TDT*.

[*] Corresponding author: e-mail: jmalley@helix.nih.gov

The essential purpose in the original *TDT*, and evidently in all later extensions, is that of a pure significance test of the null model, which states that there is no association and/or no linkage of a marker gene with the disease gene. But one might also be interested in estimating these model parameters and deriving associated confidence intervals, as this may provide much more insight to the researcher than simple tests of a null point hypothesis. Hence we present methods for obtaining parameter estimates and confidence intervals for linkage and association from allele transmission data as is collected for use in the *TDT*. The separate estimates of association and linkage derived here enable one to declare which part of the null model is being rejected by the *TDT*: no linkage, or no association, or both. In addition we use the reparameterization introduced here to allow us to comment on the performance of the *TDT* under varying conditions. Other researchers have achieved similar success with this type of approach, that is, estimation and generation of confidence intervals for important genetics parameters; see for example CORDELL and ELSTON (1999), CORDELL and CARPENTER (2000).

In the next section we present the SHAM (1998) probability model for family-based allele transmission data. In Section 3 we discuss the identifiability problem and consider the likelihood ratio test. Section 4 contains a discussion of the integrated likelihood approach for confidence intervals.

We close the paper with the presentation of Monte Carlo simulations and a discussion of our results. In general, our procedure has greater power than the *TDT* for detecting association, especially when linkage is only weak or moderate. Moreover, when association is at least moderate, the confidence intervals for linkage successfully distinguish weak from moderate or strong linkage.

A version of our C code, **ELAAT** (Estimation of Linkage and Association from Allele Transmission) is freely available on-line at: http://mscl.cit.nih.gov/spaj/elaat

2. The Probability Model for Allele Transmissions

Transmission/disequilibrium data consist of a square table of counts. Classically, they are obtained by genotyping an affected child and both parents. From this data one can, for each parent, specify which allele was transmitted to the child, except in the ambiguous case of doubly heterozygous parents with identical alleles. The basic assumption is that the parent-offspring pairs have been ascertained through a random sample of affected children from a randomly mating population. This basic probability model also assumes that all allele transmissions are unambiguously known, and uses transmission data from just a single parent. Of course, in general, genotype data from both parents is required to identify alleles transmission in the chosen parent. In a later paper we will examine robustness of our estimation methods when this latter condition is relaxed and allele transmission data from both parents is included, as well as when multiple affecteds from a single family are included. The allele transmission model used here is the simplest useful model:

allele transmission data from just one parent per family trio, under the assumption of random mating.

Our notation follows that of Sham and Curtis (1995). A basic reference is Sham (1998) in which the p's and q's of Sham and Curtis (1995) are reversed. The model, in full generality has $k \geq 2$ marker alleles so let:

$p_i =$ frequency of marker allele $i, 1 \leq i \leq k$
$q_1 =$ frequency of susceptibility allele at disease locus
$q_2 =$ frequency of normal allele at disease locus
$h_{1i} =$ frequency of the haplotype with disease allele and marker allele i
$e_{1i} = h_{1i}/q_1 p_i$ linkage disequilibrium between disease allele and marker allele i
$f_{rs} =$ penetrance given genotype (r, s) at disease locus
$\theta =$ recombination frequency.

These basic variables are then used to define

$$K = q_1^2 f_{11} + 2q_1 q_2 f_{12} + q_2^2 f_{22} = \text{population prevalence of disease}$$

and

$$B = q_1[q_1(f_{11} - f_{12}) - q_2(f_{11} - f_{22})]/K = \text{the mode of inheritance parameter.}$$

We note that other definitions for linkage disequilibrium (LD) are present in the literature; see for example the discussion in Cordell and Elston (1999).

As Sham and Curtis (1995) observe, the parameter B is "one minus the ratio of the conditional probability of transmitting the normal allele, given that the offspring is affected, to the unconditional probability of transmitting the normal allele." In still other words, it expresses the degree to which selection through affected offspring has diminished the transmission of the normal allele.

Continuing, the probability π_{ij} that a parent having genotype (i, j) transmits allele i and does not transmit allele j to an affected is given by:

$$\pi_{ij} = p_i p_j (1 + B[(e_{1i} - 1) + \theta(e_{1j} - e_{1i})]) \quad \text{for} \quad 1 \leq i, j \leq k\,.$$

Note first that if the parental genotype is (i, i), then one of the i alleles must be transmitted, and the other allele, also of type i, is not transmitted. Second, in this basic form of the allele transmission model only the contribution of one parent is counted. If both are counted it must be assumed that the parental contributions are independent. This is in general only valid under the null model, as discussed by Sham and Curtis (1995). However, it is also the case that more than one affected can be counted in a single family, as can be seen from the derivation given in Sham and Curtis (1995), assuming that the meiosis events leading to allele transmissions are independent across the affecteds.

A crucial fact about the probability model for allele transmission data is that the model is not identifiable if the recombination frequency $\theta = 0.5$ or if the mode of inheritance term B is unknown. That is, distinct parameter values can generate the same cell probabilities, for all observed data values. The following result, whose proof is given in Appendix I, clarifies this issue:

Theorem: Let $B > 0$ be given. Then the model is identifiable if and only if $\theta \neq 0.5$ and $e_{1i} \neq 1$ for some index i.

So we reparameterize the model in order to solve this identifiability problem. To simplify we focus on the case for just two marker alleles, and introduce a new parameter for the linkage disequilibrium term $d = aB$, where $a = e_{11} - 1$. We then propose the model:

$$\begin{aligned}
\pi_{11} &= p^2[1 + d]\,, \\
\pi_{12} &= p(1-p)\,[1 + d - \mathrm{d}\theta/(1-p)]\,, \\
\pi_{21} &= p(1-p)\,[1 + \mathrm{d}p/(1-p) + \mathrm{d}\theta/(1-p)]\,, \\
\pi_{22} &= (1-p)^2\,[1 - \mathrm{d}p/(1-p)]\,.
\end{aligned}$$

In this model we observe that (i) d is free of confounding with recombination frequency θ; (ii) for all values of θ and p, d is bounded by $-1 \leq -B \leq d \leq B(1-p)/p$ and (iii) when d is zero the linkage term θ drops out of the model.

The null hypothesis in which we are interested in is that we have no association ($d = 0$) or that we have no linkage ($\theta = 0.5$). But the model is nonidentifiable exactly when $\theta = 0.5$ or $d = 0$. Hence nonidentifiability occurs for any parameters that satisfy the null hypothesis. This of course generates interesting technical difficulties; these are discussed in the next section.

Assume now that data has been collected, $x = (x_{11}, x_{12}, x_{21}, x_{22})$, where the components correspond to counts for the cells with probabilities $(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$. Then for the model above the log likelihood is (apart from an additive constant) given by:

$$l(d, B, \theta, p \mid x) = \sum_{i,j} x_{ij} \log\,[\pi_{ij}(d, \theta, p)]\,.$$

Before discussing our integrated likelihood approach to estimation and intervals we use our reparameterization to first investigate the structure of the *TDT* and its connection to a likelihood ratio test for linkage.

3. The Likelihood Ratio Test and the *TDT* as a test for linkage

As we have noted, at the null value $\theta = 0.5$, the likelihood is not identifiable, and the null value is itself on the boundary of the parameter space. These two facts together invalidate the derivation of the distribution of the likelihood ratio (*LR*) test, as given in SELF and LIANG (1987); see GOLDSTEIN (1995) for details on how the boundary correction given in SELF and LIANG is often misused in the statistical genetics literature.

We resolve these two problems in one stroke using another parameterization for the basic allele transmission probability model. We begin by noting that if $d \neq 0$, $p \in (0, 1)$ and $\theta = 0.5$, then the likelihood function is not identifiable in that there are exactly two points in the parameter space for which the likelihood function has the same values (see Appendix I for a proof of this result). Specifically given one point $(d > 0, p, \theta = 0.5)$, then the other point is $(d^0 = -d/(d+1) < 0, p^0 = p(d+1), \theta = 0.5)$. So define

$$\Theta = \{-1 < d \leq (1-p)/p\} \times \{0 \leq \theta \leq 0.5\} \times \{0 < p < 1\},$$
$$\Theta_2 = \{0 \leq d \leq (1-p)/p\} \times \{0 \leq \theta \leq 1\} \times \{0 < p < 1\}$$

and let

$$\tau(d, p, \theta) = \begin{cases} (d, p, \theta) & \text{if} \quad d \geq 0, \\ (-d/(d+1), p(d+1), 1-\theta) & \text{if} \quad d < 0. \end{cases}$$

Then the mapping $\tau : \Theta \to \Theta_2$ is onto and, outside of the plane $\theta = 0.5$, the mapping is one to one as well. When $\theta = 0.5$ the function τ also maps points with negative values of d to points with positive values of d.

Finally, the likelihood function $L(\tau(d, p, \theta)) = L((d, p, \theta))$, is smooth over Θ_2, and the information matrix can be shown to be invertible when $\theta = 0.5$. Given all these facts, the standard asymptotic result is seen to hold for the likelihood ratio test (see for example FERGUSON (1996)) and we have the following theorem:

Theorem: Assume that a given marker allele has association $d \neq 0$. Let

$$LR = \max_{\Theta} l(d, \theta, p) - \max_{\Omega} l(d, \theta, p)$$

where

$$\Theta = \{-1 < d \leq B(1-p)/p\} \times \{0 \leq \theta \leq 0.5\} \times \{0 < p < 1\},$$
$$\Omega = \{-1 < d \leq B(1-p)/p\} \times \{\theta = 0.5\} \times \{0 < p < 1\}.$$

Then for testing $\theta = 0.5$, LR has an asymptotic χ^2 distribution, under the null hypothesis, with one degree of freedom:

$$P(2LR \geq c) = \alpha,$$

where c is such that $P(\chi_1^2 \geq c) = \alpha$, for χ_1^2 a chi-square with 1 d.f.

Next, using a Taylor series expansion for the *LR* test, one can show that it is asymptotically equivalent to the *TDT* (details not given, available from the corresponding author upon request). This equivalence can also be seen from the simulations in Table 3a and 3b. Thus, for both analytic and computational reasons, we see that the *TDT* is for all practical purposes equivalent to a likelihood ratio test for linkage, given that association is nonzero. It is, therefore, not useful to view or apply the *TDT* as a test of association in the presence of linkage. In fact, as might be expected, the *TDT* has little power for detecting association when linkage is weak or only moderate, and our simulations verify this point. For example, the integrated likelihood test (*IL*, see below) for detecting association, when linkage $\theta = 0.45$, has power 0.994, while the *TDT* has power 0.248. (Additional results of these simulations are given in Table 3a.)

Finally, we observe that ABEL and MULLER-MYHSOK (1998) discuss the *TDT* in relation to a likelihood test they introduce. However, their likelihood uses only data from the parents for which the transmitted and not transmitted alleles are different: they use the off-diagonal elements of the 2 × 2 data table. Our basic allele transmission likelihood, discussed above, uses all the data from the 2 × 2 data table, and data is collected under the same assumptions as that in ABEL and MULLER-MYHSOK (1998), e.g. random mating, Hardy-Weinberg equilibrium, etc. We note that the diagonal elements of the complete 2 × 2 table contain information about the association parameter, *d*, so for this reason alone we could expect our estimation and testing methods to have greater power.

This concludes our discussion of the identifiability issues and the relation of the *TDT* to the method of maximum likelihood. We next turn to our main procedure for estimating association and linkage.

4. Integrated Likelihood Estimation

4.1. *A general introduction to integrated likelihood estimation*

Good discussions of the integrated likelihood (*IL*) approach can be found in BERGER, LISEO, and WOLPERT (1999) (see also GELMAN et al. (1995), CARLIN and LOUIS (2000), or SEVERINI (2000)).

To assist the reader we present a brief outline of the method and consider a statistical model with parameters ψ and λ, where ψ is the parameter of interest and λ is a nuisance parameter, and let $L(\psi, \lambda)$ denote the likelihood function for a given data set. In general, likelihood-based inference for ψ is complicated by the presence of λ in the likelihood function.

One approach (see SEVERINI (2000) for references to many other approaches) is that of an *integrated likelihood function* of the form

$$L_{\mathrm{I}}(\psi) = \int_{\Lambda} L(\psi, \lambda)\, \pi(\lambda \mid \psi)\, \mathrm{d}\lambda$$

where $\pi(\lambda \mid \psi)$ represents a nonnegative weight function on Λ, the space of possible λ. Note that this space may depend on the value of ψ under consideration. Inference for ψ may then proceed by treating $L_{\mathrm{I}}(\psi)$ as a likelihood function for ψ. Thus, for example, an estimate of ψ may be obtained by maximizing $L_{\mathrm{I}}(\psi)$, to find $\hat{\psi}_{\mathrm{I}}$ say.

In order to carry out this approach it is necessary to choose the weight function $\pi(\lambda \mid \psi)$ and different choices of this function will lead to different forms for $L_{\mathrm{I}}(\psi)$. One choice is the uniform weight function that is constant on Λ. BERGER, Liseo and Wolpert (1999) note that this is an attractive choice when nothing else is suggested. Note that if Λ does not depend on the value of ψ under consideration, then $\pi(\lambda \mid \psi)$ may be taken to be 1. Otherwise, $\pi(\lambda \mid \psi)$ is a function of ψ but not of λ. The integrated likelihood functions used in this paper are all based on uniform weights.

The integrated likelihood function can be used as one would normally use a likelihood function. In particular, the value of ψ that maximizes $L_{\mathrm{I}}(\psi)$can be used as a point estimate of ψ. A frequentist confidence interval for the parameter of interest, ψ, may be obtained by solving for all ψ such that:

$$\{\psi : -2[L_{\mathrm{I}}(\psi) - L_{\mathrm{I}}(\hat{\psi}_{\mathrm{I}})] \le \chi_1^2\}$$

where χ_1^2 is a χ^2 variable on one d.f., and $\hat{\psi}_{\mathrm{I}}$ is the maximizer of $L_{\mathrm{I}}(\psi)$. Such intervals may also be used to obtain frequentist hypothesis tests.

Note that the integrated likelihood approach only requires a weight function for the nuisance parameter of the model. In contrast, a full Bayesian analysis would require a prior for the parameter of interest as well and the results of such as an analysis may depend heavily on the prior distribution used.

4.2. *Application of integrated likelihood to the* Sham *and* Curtis *(1995) model*

Recall the basic allele transmission probability model presented in Section 2, and consider inference for d, where we assume that d is not zero. We note that the set of allowed values for p depends on the value of d and B, since we have the upper bound

$$p \le B/(d+B)\,.$$

As B is not identifiable we set it to its maximum, $B = 1$. Simulations show that this arbitrary choice has little practical effect on inference for d.

Using the upper bounds $p = 1/(d+1)$, and $0 \le \theta \le 0.5$ we see that for a given value of d, the uniform weight on the space of possible (θ, p) is given by

$$\begin{aligned} \pi(\theta, p \mid d) &= 2/(1+d) \quad \text{for} \quad 0 \le \theta \le 0.5\,, \qquad 0 < p \le 1/(d+1) \quad \text{if} \quad d > 0 \\ &= 2 \qquad\qquad\quad \text{for} \quad 0 \le \theta \le 0.5\,, \qquad 0 < p \le 1/(d+1) \quad \text{if} \quad d < 0\,. \end{aligned}$$

For inference about θ, $\pi(d, p \mid \theta)$ is taken to be constant on the set

$$\{(d, p) \colon 0 < p < 1, -1 < d < \infty, p < 1/(d+1)\}\,,$$

which is the space of possible (d, p). Note that this space does not depend on the value of θ under consideration, so that $\pi(d, p \mid \theta)$ does not depend on θ.

To obtain accurate nominal levels for tests based on the *IL* approach, calibration of the interval above may be needed, in order to correct for any discrepancy between the observed and the stated nominal level of the test. Our Monte Carlo studies reveal that such calibration is needed for inference regarding θ, but not for d.

In our use of *IL* we emphasize that the introduction of flat priors is largely irrelevant to our final results, since the marginal likelihoods produced and the intervals and estimates derived from them, are all then examined in a conventional frequentist manner. The *IL* approach is simply one way among many for generating estimates and tests of the model parameters all within the frequentist context. Finally we note that when the intervals are considered as tests of the null model (no association and/or no linkage) we can see that they have other satisfactory frequentist properties: good nominal level and good power.

5. Monte-Carlo Simulations

To evaluate the performance of the *IL* methods, and to explore the sensitivity of the *TDT* or *IL* to population stratification, we generated a variety of simulated data. In particular we study five models, those considered in SHAM and CURTIS (1995): classical, and common, with both recessive and dominant genes, and Alzheimer disease (see POST and WHITEHOUSE (1998), for example, for background on Alzheimer disease). We also study three admixture models and three linkage heterogeneity models, and in all six of these mixture models one of the subpopulations is chosen to satisfy the null model. For the three population admixture models we set $\theta = 0.5$ and use $d = 0.057$, $d = 0.20$ and $d = 0.40$ with various disease and marker allele frequencies for the subpopulations. For the linkage heterogeneity models we mix together a population with $d = 0.0$ and $\theta = 0.5$ with a population with $d > 0$ and $\theta = 0.001$ using various disease and marker allele frequencies for the subpopulations. In the case that $d = 0.0$, we recall that the parameter θ is not identifiable and hence this variable can be chosen to have an arbitrary value.

In each case the Monte-Carlo simulations are based on using tables with 50 parent pairs (which generates 100 values in each table); this number of parent pairs was chosen as being of practical interest in current routine applications of the *TDT.* For each model we then generate 1000 of these tables. For each of these 1000 tables we generated an integrated likelihood point estimate and computed the average bias, root mean squared error and relative bias of these estimates. In addition, for each of these 1000 tables we computed a confidence interval based on 1000 additional replications and computed the probability that the point estimate was contained in the confidence interval: this is the coverage probability. The average length of the confidence intervals is also computed since coverage probabilities may be inflated if the average confidence interval length is large.

In Table 1 we present results for the five classical models and the six mixture models. The table contains the results of the integrated likelihood procedure and presents coverage probabilities, average interval length, average bias, root mean square error and the relative bias for the estimation of the parameters θ and d.

Table 2 contains the same types of values as in Table 1 except in this table we allow the parameters θ and d to vary in a regular way to explore the behavior of

Table 1

Monte-Carlo Simulations with Classical and Mixture Models Notation:

Classical Models

R1 = Classical recessive, $q = 0.0316$, $p = 0.25$, $\theta = 0.02$, $d = 2.00$
D1 = Classical dominant, $q = 0.0005$, $p = 0.25$, $\theta = 0.02$, $d = 1.25$
R2 = Common recessive, $q = 0.1000$, $p = 0.25$, $\theta = 0.02$, $d = 1.20$
D2 = Common dominant, $q = 0.0050$, $p = 0.25$, $\theta = 0.02$, $d = 0.60$
AD = Alzheimer disease, $q = 0.1300$, $p = 0.25$, $\theta = 0.02$, $d = 0.75$

Population Admixture Models

M1: $q = 0.25$, $q^* = 0.5$, $p = 0.3$, $p^* = 0.7$, $\theta = 0.50$, $d = 0.057$
M2: $q = 0.01$, $q^* = 0.03$, $p = 0.3$, $p^* = 0.7$, $\theta = 0.50$, $d = 0.20$
M3: $q = 0.01$, $q^* = 0.06$, $p = 0.2$, $p^* = 0.7$, $\theta = 0.50$, $d = 0.40$

Linkage Heterogeneity Models

M4: $r = 0.25$, $q = 0.01$, $p = 0.25$, $\theta = 0.50$, $\theta^* = 0.01$, $d = 0.00$, $d^* = 0.12$
M5: $r = 0.05$, $q = 0.01$, $p = 0.25$, $\theta = 0.50$, $\theta^* = 0.01$, $d = 0.00$, $d^* = 0.12$
M6: $r = 0.25$, $q = 0.01$, $p = 0.25$, $\theta = 0.50$, $\theta^* = 0.01$, $d = 0.00$, $d^* = 2.0$

Integrated Likelihood (*IL*) Estimation Results with Classical and Mixture Models

	cover θ	cover d	int len θ	int len d	av bias θ	av bias d	rms θ	rms d	rel bias θ	rel bias d
R1	0.984	0.947	0.1513	3.2872	−0.0188	−0.4016	0.0401	0.6899	−0.4682	−0.5821
D1	0.951	0.958	0.2831	3.6640	−0.0755	−0.5396	0.0801	0.7397	−0.9424	−0.7295
R2	0.937	0.906	0.2809	3.9446	−0.0779	−0.6790	0.0834	0.8589	−0.9342	−0.7906
D2	0.937	0.919	0.4445	4.6208	−0.2099	−0.6240	0.1218	0.9471	−1.7234	−0.6589
AD	0.958	0.953	0.4031	4.1269	−0.1533	−0.5774	0.1095	0.7697	−1.3999	−0.7501
M1	0.926	0.942	0.4681	0.6785	+0.0887	−0.0578	0.0951	0.1521	+0.9320	−0.3796
M2	0.372	0.910	0.4239	0.7606	+0.2540	−0.1251	0.1127	0.1688	+2.2537	−0.7415
M3	0.981	0.733	0.4155	1.0408	+0.0381	+0.2786	0.0652	0.3168	+0.5845	+0.8795
M4	0.732	0.946	0.4471	2.3580	−0.4180	−0.2870	0.0874	0.7435	−4.7833	−0.3861
M5	0.779	0.945	0.4559	2.2327	−0.4047	−0.2512	0.0907	0.5793	−4.4636	−0.4337
M6	0.938	0.909	0.1236	1.4904	−0.0264	+0.2237	0.0406	0.3782	−0.6494	+0.5914

(a) 1000 tables generated in all simulations (using SLINK).

(b) All based on 50 parent pairs (N = 100), except for M3 and M6 which had 1000 parent pairs (N = 2000).

(c) the bias and coverage probabilities for d in all M models were taken relative to the nonzero values for d given above.

(d) using q = disease allele frequency, p = marker allele frequency: in M1 20% of the affecteds come from population (q, p) in M2 20% of the affecteds come from population (q, p), in M3 5% of the affecteds come from population (q, p).

(e) in M4, M5, M6, r is the mixing rate such that proportion r comes from the population with values (θ, d).

Table 2
Monte-Carlo Simulations using Primary Models
Integrated Likelihood (*IL*) Estimation Results for the Primary Models

(d, θ)	cover	cover	int len	int len	avg bias	avg bias	rms	rms	rel bias	rel bias
	θ	*d*	θ	*d*	θ	*d*	θ	*d*	θ	*d*
0.0, 0.5	0.979	0.950	0.4566	2.1451	+0.0438	−0.1688	0.0746	0.5730	+0.5870	−0.2946
1.0, 0.5	0.987	0.770	0.3292	2.2401	+0.0315	+0.3571	0.0545	0.7485	+0.5784	+0.4771
2.0, 0.5	0.979	0.950	0.1050	1.6636	+0.0178	+0.0088	0.0314	0.4263	+0.5653	+0.0206
2.5, 0.5	0.976	0.938	0.0840	1.5002	+0.0148	+0.0011	0.0275	0.3988	+0.5381	+0.0028
1.0, 0.45	0.986	0.847	0.3407	2.2343	+0.0024	+0.2895	0.0624	0.6957	+0.0384	+0.4161
2.0, 0.45	0.975	0.950	0.1282	1.6604	+0.0003	+0.0088	0.0412	0.4263	+0.0065	+0.0206
2.5, 0.45	0.963	0.938	0.1094	1.5002	+0.0009	+0.0011	0.0359	0.3988	+0.0255	+0.0028
1.0, 0.35	0.967	0.936	0.3813	2.0204	−0.0059	+0.1378	0.0809	0.5779	−0.0725	+0.2385
2.0, 0.35	0.942	0.950	0.1616	1.6560	+0.0011	+0.0088	0.0424	0.4263	+0.0250	+0.0206
2.5, 0.35	0.934	0.938	0.1266	1.5002	+0.0011	+0.0011	0.0351	0.3988	+0.0319	+0.0028
1.0, 0.25	0.968	0.969	0.3587	1.8461	−0.0091	+0.0292	0.0837	0.4889	−0.1086	+0.0598
2.0, 0.25	0.938	0.950	0.1627	1.6510	+0.0025	+0.0086	0.0415	0.4257	+0.0596	+0.0201
2.5, 0.25	0.952	0.938	0.1193	1.5002	+0.0016	+0.0011	0.0318	0.3989	+0.0499	+0.0028
1.0, 0.15	0.966	0.980	0.3014	1.7155	−0.0199	−0.0821	0.0824	0.4140	−0.2414	−0.1984
2.0, 0.15	0.948	0.953	0.1574	1.6209	+0.0025	+0.0045	0.0394	0.4192	+0.0634	+0.0108
2.5, 0.15	0.947	0.938	0.1073	1.4997	+0.0024	+0.0011	0.0283	0.3987	+0.0852	+0.0027
1.0, 0.05	0.957	0.963	0.2352	1.6022	−0.0447	−0.2244	0.0732	0.3539	−0.6098	−0.6341
2.0, 0.05	0.956	0.968	0.1118	1.4934	−0.0009	−0.0737	0.0330	0.3678	−0.0273	−0.2005
2.5, 0.05	0.948	0.943	0.0803	1.4648	+0.0021	−0.0178	0.0216	0.3848	+0.0962	−0.0462
1.0, 0.001	0.969	0.885	0.1140	0.6492	−0.0283	−0.1420	0.0341	0.1340	−0.8307	−1.060
2.0, 0.001	0.980	0.910	0.0333	0.6033	−0.0061	−0.1035	0.0009	0.1532	−0.6480	−0.6756
2.5, 0.001	0.973	0.932	0.0176	0.6063	−0.0025	−0.0721	0.0050	0.1690	−0.5033	−0.0427

(a) at each setting of d and θ, 1000 tables were generated, marker allele frequency fixed at $p = 0.25$, data simulated using SLINK.

(b) all tables based on 200 parent pairs (N = 400), except for (d, θ) = (1.0, 0.001), (2.0, .001), (2.5, 0.001) which had 1000 parent pairs (N = 2000).

(c) notation: *cover* = coverage probability using 95% confidence interval for that parameter; int len = average interval length for 95% confidence interval; *avg bias* = average bias of parameter estimate; *rms* = root mean square error for parameter, taken with respect to known value for that parameter; *rel bias* = (*avg bias*)/(*rms*).

(d) when $d = 0$, the parameter θ drops out of the allele transmission model, and the results regarding estimation for d and θ are essentially the same for all values of θ. Hence the results for $d = 0$ are not duplicated for the other values of θ.

the *IL* approach over the parameter space. Again when $d = 0$, θ is not identifiable and can be chosen to have any value.

In Tables 3a and 3b we study the relationships between the Likelihood ratio test and the *TDT* for both the classical models of Table 1 and the models of Table 2.

Full details of the Monte Carlo simulations and results are given in Tables 1, 2, and 3.

Table 3a

Power results for the *TDT*, *IL*, and the Likelihood Ratio (*LR*) procedures on the Primary Models

(d, θ)	*IL* test for $d = 0$	*IL* test for θ = 0.5	*LR* test for θ = 0.5	*TDT*
0, 0.5	0.049	0.029	0.052	0.052
1.0, 0.5	0.158	0.012	0.053	0.053
2.0, 0.5	0.987	0.020	0.047	0.047
2.5, 0.5	0.999	0.023	0.048	0.048
1.0, 0.45	0.248	0.045	0.101	0.101
2.0, 0.45	0.994	0.194	0.248	0.248
2.5, 0.45	0.999	0.285	0.331	0.331
1.0, 0.35	0.649	0.394	0.535	0.535
2.0, 0.35	0.999	0.949	0.960	0.960
2.5, 0.35	0.999	0.993	0.994	0.994
1.0, 0.25	0.961	0.899	0.951	0.951
2.0, 0.25	0.999	0.999	0.999	0.999
2.5, 0.25	0.999	0.999	0.999	0.999
1.0, 0.15	0.999	0.995	0.998	0.998
2.0, 0.15	0.999	0.999	0.999	0.999
2.5, 0.15	0.999	0.999	0.999	0.999

(a) for all models results are formed using 1000 tables each with 200 parent pairs (N = 400).

(b) for the models (d, θ) = (1.0, 0.05), (2.0, 0.05), (2.5, 0.05), (1.0, 0.001), (2.0, 0.001), (2.5, 0.001), the power sof all tests were 0.999 and are hence omitted from this table.

(c) the *IL* tests are based on the 95% integrated likelihood confidence intervals for each parameter.

6. Results and Discussion

As indicated in the Introduction, and as discussed in Whittaker, Denham and Morris (2000), separation of association and linkage following a significant *TDT* result is important, since a significant *TDT* may result from loose linkage and strong association, *or* from tight linkage and weak association. Often only the latter possibility is considered as the alternative to the rejection of the null model when using the *TDT*. Moreover, as we have shown above, the TDT is not a useful test of association, having very little power when linkage is weak: it returns little information about the amount of association present in these cases. Also, while Whittaker, Denham and Morris (2000) suggest the possibility of obtaining maximum likelihood estimates for association and linkage, they argue that it would be "virtually impossible to distinguish, solely on the basis of [family-based data], between tight and loose linkage." Our results tell a rather different, more complex story.

Table 3b
Power Results for Classical and Mixture Models

Model	*IL* test for $d = 0$	*IL* test for $\theta = 0.5$	*LR* test for $\theta = 0.5$	*TDT*
R1	0.999	0.999	0.999	0.999
D1	0.989	0.972	0.988	0.983
R2	0.989	0.979	0.989	0.989
D2	0.600	0.412	0.603	0.593
AD	0.778	0.686	0.814	0.809
M1	0.145	0.073	0.153	0.148
M2	0.751	0.627	0.781	0.786
M3	0.072	0.018	0.050	0.049
M4	0.118	0.054	0.101	0.101
M5	0.132	0.065	0.129	0.129
M6	0.999	0.999	0.999	0.999

(a) the Classical models (R1, D1, R2, D2, AD) assume $\theta = 0.02$, and the values given above for the *IL* and the *LR* test for $\theta = 0.5$, and the TDT, are the observed powers of these tests using a stated level of 0.05.

(b) the Mixture models M1, M2, M3, M4, M5, and M6 assume a variety of values for *d* and θ in one of the two subpopulations, but the null values for the tests in all cases are $d = 0$ or $\theta = 0.5$. Details for these models are given in the Notes for Table 2.

The power calculations given in Tables 1, 2, and 3 tell part of the story for testing of association or linkage, when using the *IL* method to generate tests of the null models.

First of all, we see that the *IL* method for testing either association or linkage can yield significant improvements in power against the *TDT* (or, equivalently, *LR*, as we have seen), but lose some power in other cases. A representative case in point: for marker allele frequency $p = 0.25$, association $d = 2.5$, and recombination frequency $\theta = 0.45$, the power of the *TDT* is 0.331, while the *IL* approach (when used for testing $d = 0$) has power 0.999; see Table 3. Note that d is bounded above by $B(1-p)/p = 3.0$, so $d = 2.5$ in this instance represents strong association. Note also that when testing $\theta = 0.5$ (when $d = 2.5$, marker frequency $p = 0.25$, and true $\theta = 0.45$) the *IL* approach has power only 0.285, while the *TDT* and *LR* have power 0.331. Of course, a power of 0.331 is not especially good either: even when association is strong, and given the sample size of 50 parent pairs, no method presented here is especially good at detecting small changes from the null value of $\theta = 0.5$.

Secondly, we observe that, when association is at least moderately strong, the *IL* confidence intervals for linkage can effectively distinguish between tight and loose linkage. Thus consider model ($d = 2.5$, $\theta = 0.05$), and note that the 95% confidence interval for θ has average interval length 0.0803, while the average bias in estimating θ is only very slightly positive: +0.0021. Similar results were obtained for the models in which strong association is present: $d = 2.5$ with $d_{\max} = 3.0$. Slightly wider intervals for linkage are obtained when association is only moderate ($d = 2.0$). Those models with weak association ($d = 1.0$) show an expected corresponding increase in our inability to estimate recombination frequency. On the other hand, by substantially increasing the sample size (from 200 to 1000 parent pairs), we obtain short length intervals for recombination frequency at all levels of association.

As noted earlier our likelihood ratio test for linkage appears to have virtually the same power and rejection rates as the *TDT*, at all levels of association, even though the *TDT* is at least formally a test of the compound null hypothesis of no association **and/or** no linkage. In other words, inherent in the performance of the *TDT*, and in our result showing that it is analytically equivalent to *LR*, we see that the *TDT* cannot be usefully considered as a test of association. However, the *IL* procedure that detects association does so at all levels of linkage, and the *IL* procedures for association or linkage both lead to valid confidence intervals for those parameters.

Summarizing the power results, we find that the *IL* test for linkage lags somewhat behind that of the *TDT* (and *LR*). As a practical matter therefore we suggest using the *TDT* as a test for linkage (in the presence of strong association), and following that with the confidence interval for θ when the *TDT* (or *LR*) is found significant. We have not, however, investigated the statistical properties (e.g. observed level under the alternative hypothesis) of this two-stage, conditional procedure.

To be complete, we note that it is generally unlikely that a human population would have strong association and yet have weak linkage. However we could expect this to be more likely in certain isolated populations, such as the Amish or the Finns. We also could expect to see strong association with weak linkage in nonhuman studies such as those involving fruit flies, mice, yeast, or test crosses with animals.

Before taking up the mixture models and the performance of the *IL* procedure considered here, let us begin by noting that the basic Mendelian models as considered by Sham and Curtis (1995), for example, in fact have rather complex expressions in terms of the original allele transmission model parameters. Thus, even these ostensibly simple genetics models are in fact quite intricate when expressed in fully parameterized form. In this regard, we observed that our reparameterization of the transmission model, specifically our introduction of *d*, the generalized association term, allows us to simulate and interrogate models with well-calibrated levels of association and linkage. This then permits a methodical inquiry of any test or estimation approach based on allele transmission data with respect to performance under departures from the homogeneous population model.

We now observe that, from the Table 3b, the tests for linkage and association as derived from our intervals are basically as robust on the mixture models as the *TDT*. The observed levels for the *TDT* in the population admixture models M1, M2, and M3 are close to the nominal level of $p = 0.05$ in these cases, since by construction there is no linkage for these data ($\theta = 0.50$). Continuing, consider models M4 and M5, where 25% of the total population has no linkage or association, while the remaining 75% has strong linkage. We might expect therefore that the *TDT* had good power in these cases, but find that the powers are 0.101 and 0.129 (for M4 and M5, respectively). That is, power is rather close to the assigned nominal level of the test, $= 0.05$. The power values for *LR* test are virtually the same as that for the *TDT*, and neither are appreciable better than the *IL* test for linkage (0.54 in M4 and 0.65 in M5). The performance of the *TDT* in these cases would suggest no linkage or association, contrary to the true situation, and the performances of the other procedures are equally unrevealing. On the other hand, in model M6, having both strong linkage and association in 75% of the population, the power for all the procedures, including the *TDT*, is essentially 1.00.

Summarizing, we see inconsistent behavior of the *TDT* in these complex linkage heterogeneity models, and this is matched unfortunately by the behavior of the separate tests for linkage and association as we present here. This should not be so surprising, since separate estimates and intervals for linkage and association are ineffective when the population being sampled does not have the assumed likelihood function, in particular, when subpopulations have distinct model parameters.

An objection might be raised concerning the wide intervals for linkage that result in the mixture models, as shown in Table 1. We have seen that in a subpopulation with zero association, linkage is not well defined in the basic allele transmission model, and in particular the "true" value might as well be $\theta = 0.5$. When mixed with a subpopulation that has nontrivial association and tight linkage we found that wide intervals for θ often result. To argue from a simple point estimate of θ in this case to conclude that linkage was present in the whole population would be an error. We caution, therefore, that a wide interval for θ should, at any time, only be considered as very weakly informative of the presence or strength of any linkage. Of course, as we emphasized above, estimation in any mixture problem is not well posed unless the likelihood accounts for the mixing, and the standard allele transmission model used here and in the genetics literature does not do this.

Currently an extremely wide range of possible stratification and mixture populations are being contemplated in genetics, especially when studying complex traits, which may include features such as multiple susceptibility alleles, multiple linked and unlinked susceptibility loci and multiple types of disease etiology. Given this large universe of possible departures from a homogeneous population model, it is difficult to organize a testing plan for any test procedure. Our results are therefore but a step in this direction, with these first steps now possible using our properly parameterized allele transmission model.

Finally, it has been observed that model identifiability issues also arise with association studies based on case-control data (e.g., SHAM (1998), p. 165). The success of the techniques described here suggests the feasibility of parameter estimation and confidence intervals in these other experimental designs, for which only frequentist tests of null models are currently available.

Appendix 1

Identifiability

We note that the SHAM (1998) model given in terms of $\phi = \{p, e, \theta, B\}$ is not identifiable. To see this consider the likelihood function

$$L(x \mid \varphi) = \prod_{i=1}^{m} \prod_{j=1}^{m} (p_i p_j (1 + B((e_i - 1) + \theta(e_j - e_i))))^{x_{ij}} .$$

We define the parameter space to be the set of values of ϕ for which $(1 + B((e_i - 1) + \theta(e_j - e_i)))) > 0$ for each index i and j and $\sum_{i=1}^{m} p_i = 1$, $\sum_{i=1}^{m} p_i e_i = 1$, $p_i > 0$ and $e_i > 0$ for $1 \leq i \leq m$.

To help with the analysis also define $a_i = e_i - 1$and then

$$L(x \mid \varphi) = \prod_{i=1}^{m} \prod_{j=1}^{m} (p_i p_j (1 + B(a_i + \theta(a_j - a_i))))^{x_{ij}} .$$

Since $\sum_{i=1}^{m} p_i = 1$ and $\sum_{i=1}^{m} p_i e_i = 1$, then $\sum_{i=1}^{m} p_i a_i = \sum_{i=1}^{m} p_i (e_i - 1) = 0$.

We first note that in all cases, the parameter B cannot be determined independently of the other parameters. To see this let c be a non zero constant and define $B' = cB$ and $a'_i = \frac{1}{c} a_i$ for $i = 1, \ldots, m$. Then $\sum_{i=1}^{m} p_i a'_i = 0$ and the value of the likelihood function will not change since $B' a'_i = cB(1/c) a_i = Ba_i$ for each index i. Hence you cannot determine both B and the a_i's independently. From this point on we will assume that the value of B is positive and known.

The model is also not identifiable if all of the a_i's are equal for in this case θ drops out of the equations altogether. This is equivalent to the case that $e_i = e$ for each i which implies that $e_i = 1$ for each i since $1 = \sum_{i=1}^{m} p_i e_i = e \sum_{i=1}^{m} p_i = e$.

Finally the model is not identifiable in the case that $\theta = 1/2$ and $e_i \neq 1$ for at least one index i. To see this suppose that p and e are given. Then define

$$p'_i = p_i (1 + B(e_i - 1)) \quad \text{and} \quad e'_i = 1 - \left(\frac{1}{B}\right) + \left(\frac{p'_i}{Bp_i}\right) .$$

Since $\sum_{i=1}^{m} p_i = 1$ and $\sum_{i=1}^{m} p_i e_i = 1$, then $\sum_{i=1}^{m} p'_i = 1$. Hence

$$\sum_{i=1}^{m} p'_i e'_i = \sum_{i=1}^{m} p_i \left(1 - \left(\frac{1}{B}\right) + \left(\frac{p'_i}{Bp_i}\right)\right) = \sum_{i=1}^{m} \left(p_i - \frac{1}{B} (p_i - p'_i)\right) = 1$$

as required. It is then easy to see that $L(x, \theta = 1/2, p, e) = L(x, \theta = 1/2, p', e')$for every value of x. Hence the model is not identifiable.

Theorem A1: If $B > 0$ is known, $e_i \neq 1$ for some index i, and $\theta \neq 1/2$, then the model parameters are identifiable.

Proof: Assuming that B is know and let $\phi = \{\theta, p, e\}$ and $\phi' = \{\theta', p', e'\}$. Then suppose that $L(x, \phi) = L(x, \phi')$ for all values of x that satisfy $\sum_{i=1}^{m} \sum_{j=1}^{m} x_{ij} = n$ so that

$$\frac{L(x, \phi)}{L(x, \phi')} = \prod_{i=1}^{m} \prod_{j=1}^{m} \left(\frac{p_i p_j}{p'_i p'_j}\right)^{x_{ij}} \left(\frac{1 + B((e_i - 1) + \theta(e_j - e_i))}{1 + B((e'_i - 1) + \theta(e'_j - e'_i))}\right)^{x_{ij}} = 1\,.$$

Suppose that $x_{kl} = 0$ for all choices of k and l except for i and j and that $x_{ij} \neq 0$. Then for each i and j the expression above simplifies to

$$p_i p_j (1 + B((e_i - 1) + \theta(e_j - e_i))) = p'_i p'_j (1 + B((e'_i - 1) + \theta(e'_j - e'_i)))\,.$$

Once again to help with the analysis we can define $w_i = 1 + B(e_i - 1)$, then $B(e_j - e_i) = w_j - w_i$ and $\sum_{i=1}^{m} p_i w_i = 1$. We similarly define $w'_i = 1 + B(e'_i - 1)$. Finally let $r_i = p_i / p'_i$. Then we have

$$r_i r_j (w_i + \theta(w_j - w_i)) = (w'_i + \theta'(w'_j - w'_i)) \qquad \text{Eq A1}$$

for each i and j.

Letting $i = j$ this gives $r_i^2 w_i = w'_i$ which can be substituted to yield

$$r_i r_j (w_i + \theta(w_j - w_i)) = (r_i^2 w_i + \theta'(r_j^2 w_j - r_i^2 w_i))\,.$$

If we rewrite this expression as a linear combination of w_i and w_j we get

$$(r_i r_j (1 - \theta) - r_i^2 (1 - \theta'))\, w_i + (r_i r_j \theta - r_i^2 \theta')\, w_j = 0\,.$$

If we exchange i for j in this expression we get

$$(r_i r_j (1 - \theta) - r_j^2 (1 - \theta'))\, w_j + (r_i r_j \theta - r_j^2 \theta')\, w_i = 0\,.$$

Since $\sum_{i=1}^{m} p_i w_i = 1$, then $w_i \neq 0$ for at least one index i. Without loss of generality assume that $w_1 \neq 0$. Then, in order for the equation to have a solution for any index j, the determinant of the coefficients must vanish. Using Mathematica the value of this determinant is $r_1 r_j (r_1 - r_j)^2\ (\theta + \theta - 1) = 0$. But $r_1 \neq 0$ and $r_j \neq 0$ and since $\theta < 1/2$, then $\theta + \theta' - 1 \neq 0$. Hence $r_1 = r_j$ for each index j and hence $r_j = c$ for some constant c. But since $r_i = p_i / p'_i$, then this implies that $p_i = c p'_i$. However $1 = \sum_{i=1}^{m} p_i = \sum_{i=1}^{m} c p'_i = c$ and so $p_i = p'_i$ for each i.

Since $r_i^2 w_i = w'_i$, then we also have that $w_i = w'_i$ for each index i. This immediately implies that $e_i = e'_i$ for each index i. Finally if $e_i \neq e_j$ for some indices i and j, then $w_i \neq w_j$ for some indices i and j which implies that $\theta = \theta'$.

The result in the next theorem gives details of the nonidentifiability problem when $\theta = 1/2$. We continue to use the expressions $r_i = p_i/p'_i$, $w_i = 1 + B(e_i - 1)$ and $w'_i = 1 + B(e'_i - 1)$.

Theorem A2: Let $B > 0$ be given and let $\phi = \{p, w, \theta = 1/2\}$ a be point in the parameter space. If $w_i \neq 1$ for some index i, then for there is exactly one other point in the parameter space at which the likelihood function always has exactly the same value. In particular, given $\{p_1, p_2, \ldots, p_m\}$ and $\{w_1, w_2, \ldots, w_m\}$, the other point is $p'_i = p_i w_i$ with $w'_i = 1/w_i$ for $i = 1, 2, \ldots, m$.

Proof: Given Theorem A1, we conclude that $\theta' = \theta = 1/2$ and note that since $B \leq 1$ and $e_i > 0$ for each index i, then $w_i > 0$ for each index i. So it is easy to show that the constructed point in the parameter space. So by equation Eq. A1, we have that

$$r_i r_j \left(\frac{w_i + w_j}{2}\right) = r_i^2 \,\frac{w_i}{2} + r_j^2 \,\frac{w_j}{2}$$

or equivalently

$$(r_i - r_j)\,(r_i w_i - r_j w_j) = 0\,.$$

This expression will be zero if either the first or second factors are zero. So we consider the following two cases.

Case 1: For all i and j, $r_i = r_j$. Then as argued above, $p_i = p'_i$ and $w_i = w'_i$ for each index i.

Case 2: So we now suppose that $r_i \neq r_j$ for some indices i and j. So let $S_1 = \{i \mid r_i = r_1\}$ and let $S_2 = \{1, 2, \ldots, m\} - S_1$. Then since $r_i \neq r_j$ when $i \in S_1$ and $j \in S_2$, then $r_i w_i = r_j w_j$ for each $i \in S_1$ and $j \in S_2$. But since equality is transitive, it must be the case that $r_i w_i = r_j w_j$ for all indices i and j. So then $\frac{p_i}{p_j}\, w_i = r_i w = c$ for some constant c. But then $1 = \sum\limits_{i=1}^{m} p_i w_i = \sum\limits_{i=1}^{m} c p'_i = c \sum\limits_{i=1}^{m} p'_i = c$. So $c = 1$, $r_i w_i = 1$ and $p_i w_i = p'_i$ for each index i. So

$$w'_i = r_i^2 w_i = r_i (r_i w_i) = r_i = \frac{p_i}{p'_i} = \frac{p_i}{p_i w_i} = \frac{1}{w_i}\,.$$

Note that the condition $w_i \neq 0$ for some index i is required to conclude that $\theta' = \theta$ and guarantees that $w'_i \neq w$ for some index i so that the two points are distinct.

References

ABEL, L. and MÜLLER-MYHSOK, B., 1998: Maximum-likelihood expression for the transmission/disequilibrium test and power considerations. *American Journal of Human Genetics*, **63**, 664–667.

BERGER, J., LISEO, B., and WOLPERT, R., 1999: Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, **14**, 1–28.

CARLIN, B. and LOUIS, T., 2000: *Bayes and Empirical Bayes Methods for Data Analysis*. 2nd Edition. Chapman & Hall.

CORDELL, H. and CARPENTER, J., 2000: Bootstrap confidence intervals for relative risk parameters in affected-sib-pair data. *Genetic Epidemiology*, **18**, 157–172.

CORDELL, H., ELSTON, R., 1999: Fieller's theorem and linkage disequilibrium mapping. *Genetic Epidemiology*, **17**, 237–252.

GELMAN, A., CARLIN, J., STERN, and H., RUBIN, D., 1995: *Bayesian Data Analysis*. Chapman & Hall.

GOLDSTEIN, D., 1995: Asymptotic distributions of polylocus test statistics. *Genetic Epidemiology*, **12**, 195–202.

KNAPP, M., 1999: The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test. *American Journal of Human Genetics*, **64**, 861–870.

POST, S. and WHITEHOUSE, P. (editors), 1998. *Genetic Testing for Alzheimer Disease; Ethical and Clinical Issues.* Johns Hopkins Press.

SCHAID, D., 1996: General score tests for associations of genetic markers with disease using cases and their parents. *Genetic Epidemiology*, **13**, 423–449.

SCHAID, D., 1998: Transmission disequilibrium, family controls, and great expectations. *American Journal of Human Genetics*, **63**, 935–941.

SCHAID, D., 1999: Case-Parent design for gene-environment interaction. *Genetic Epidemiology*, **16**, 261–273.

SELF, S. and LIANG, K., 1987: Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605–610.

SEVERINI, T., 2000: *Likelihood Methods in Statistics.* Oxford University Press.

SHAM, P., 1998: *Statistics in Human Genetics*. Wiley & Sons.

SHAM, P. and CURTIS, D., 1995: An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Annals of Human Genetics*, **59**, 323–336.

SPIELMAN, R., MCGINNIS, R., and EWENS, W., 1993: Transmission test for diabetes mellitus (IDDM). *American Journal of Human Genetics*, **52**, 506–516.

SPIELMAN, R., MCGINNIS, R., and EWENS, W., 1994: The transmission/disequilibrium test detects cosegregation and linkage. *American Journal of Human Genetics*, **54**, 559–560.

WHITTAKER, J., DENHAM, M., and MORRIS, A., 2000: The problems of using the transmission/disequilibrium test to infer tight linkage. *American Journal of Human Genetics,* **67**, 523–526.

WILSON, S., 1997: On extending the transmission/disequilibrium test (TDT). *Annals of Human Genetics,* **61**, 151–161.

Received January 2002
Revised October 2002
Accepted December 2002