Statistics in Molecular Biology
IMS Lecture Notes – Monograph Series (1999) Volume 33

STATISTICAL ASPECTS OF THE TRANSMISSION/DISEQUILIBRIUM TEST (TDT)

BY WARREN J. EWENS

University of Pennsylvania

The transmission/disequilibrium test (TDT) was introduced as a direct test of linkage which is not affected by the problem of population stratification. Such a test is needed since much of the data used currently for linkage tests does come, or might be suspected to come, from stratified populations. Also, the test is valid when the data include relatives, since it overcomes the dependence properties usually associated with such data. In practice, the main purpose of the procedure is to test for linkage between a marker locus and a purported disease locus - a link in the chain of activities whose ultimate aim is to locate disease loci. The test differs from frequently used tests based on sharing of marker alleles between affected relatives and, unlike sharing tests, is related to the population concept of association. These differences are discussed. Many interesting questions arise in the statistical theory of the TDT, some of which are still unresolved. One of the aims of this paper is to raise and discuss these.

1. Introduction. The aim of this paper is to discuss both genetical and statistical aspects of the so-called transmission/disequilibrium test (TDT) of Spielman et al. (1993). The TDT is a test for linkage between a marker locus and a disease locus, and may also, for some forms of data, be used as a test of association between these loci. However, the properties of the test when used for these two purposes are different. Although the test is naturally of more interest to geneticists than statisticians, there are several statistical aspects of the test that deserve attention, and also several for which the statistical theory is still not complete. A discussion of some of these will be given in this paper.

2. Genetical background. Since this presentation is intended for statisticians, we first give a brief definition of key genetical terms that will be used in the sequel, as well as some population genetics theory.

Many characteristics which we have are controlled by the genes that we carry. Genes may be thought of as beads on a string, the string in this case being the chromosome, or gamete. Just as specific beads have given locations on the string, so also genes occur at specific positions, or loci, on the chromosome. Thus we might say: "The genes controlling eye color occur on chromosome 16, at a locus in such and such a position on this chromosome."

AMS 1991 *subject classifications.* Primary 62G10, 92D10; secondary 62P10.
Key words and phrases. Linkage, association, transmission/disequilibrium test, genetic diseases, hypothesis testing, power.

A gene is of one or other type, and the possible gene types at a locus are called the alleles at this locus. We might thus define the allele for blue eyes and the allele for brown eyes. Thus an allele is a type name whereas a gene is an actual material object. Sometimes the words "gene" and "allele" are used loosely and interchangeably, in particular in human genetics, so that one talks about the gene for blue eyes, and talks about the transmission of an allele from parent to offspring. In conformity with this we will sometimes adopt the incorrect usage for these words.

Different allelic types are denoted by upper-case letters, often with suffixes, such as A_1, D_2, and so on. It is a convention to use the same letter, but different suffixes, for possible alleles at the same locus.

The chromosomes in any individual appear in matched pairs, one deriving from the mother of the individual and one from the father. Chromosomes are not necessarily passed on faithfully from parent to offspring - sometimes a "crossover" will occur whereby a parent passes on part of one of his/her chromosomes and the remaining part of the other. In such a case we say that a recombination event has occurred at the crossover point. In fact two or more crossovers can occur in the transmission of a chromosome. Genes at loci close to each other, that is closely linked genes, tend to be passed on together from parent to offspring, while genes far apart on the chromosome are passed on almost independently.

Marker genes and marker loci are central to locating disease genes by linkage analysis. These have two essential properties: (i) we know where the various marker loci are on the chromosomes, and (ii) we can tell the allelic type of the (two) marker genes any given individual has at a given marker locus. It is also important to point out a third property, that there are now thousands of marker loci scattered throughout the chromosomes we carry, so that the markermap is "dense".

The main result of population genetics theory that we need is that the allelic frequencies at closely linked loci tend to "co-evolve" - the evolutionary processes at two closely linked loci are not independent. The frequencies of alleles that are not closely linked evolve essentially independently. One outcome of this is that if "D_1" is an allele at some locus "D" and "M_1" is some allele at a closely linked locus "M", then often

$$freq(D_1M_1) \neq freq(D_1) \times freq(M_1)$$

When this inequality holds, we say that there is association between the alleles at the two loci. Association is thus an essentially statistical concept, relating to gene and chromosome frequencies in some population, and the coefficient

of association, defined below, is closely related to the statistical concept of a correlation coefficient.

If there are only two alleles (D_1 and D_2) at a locus D, and only two alleles (M_1 and M_2) at a locus M, we may define the "coefficient of association" δ between the alleles at the two loci by

$$\delta = freq(D_1M_1) - freq(D_1) \times freq(M_1)$$

(The replacement of D_1 by D_2, or M_1 by M_2, or both, might change the sign of δ but not its absolute value. Since only this absolute value is important, no loss of generality is implied by using M_1 and D_1 in the definition of δ.)

When δ is nonzero there is an association between the alleles at the two loci. In the co-evolutionary example just described, this association has arisen because of the linkage between the two loci. A better expression in this context is that there is "association due to linkage", or (in the frequently used expression), that the two loci are in "linkage disequilibrium".

If D_1 is an allele causing some disease, or contributing towards causing some disease, and M_1 is an allele at a closely linked marker locus, then the event "δ not zero" implies

$$freq(M_1 \text{ among those with the disease}) \neq freq(M_1 \text{ among those free of the disease})$$

The inequality of these two frequencies is often a more useful way of stating that there is association between the genes at the two loci, and testing for such an inequality can then be used as a way of testing for such an association.

3. Linkage tests: historical background. Over the last eighty years linkage analysis has gone through several phases. In the first of these, simple and obvious linkages were observed, with no statistical testing being involved. As an example, the location of a gene (more strictly, allele - here we adopt the loose terminology of human genetics) for hemophilia was found very early on, in part by using by simple association - hemophilia occurs much more often in men than in women. (Another important component to this conclusion is the pattern of inheritance of hemophilia.)

This form of argument that there is an association between gender and disease status led naturally to the next type of test for linkage, also based on the idea of association, namely case-control methods.

In case-control methods we compare the frequency of the gene M_1 in a sample of "cases" - individuals with the disease in question - with that in a

sample of "controls" - those free of the disease. If these two frequencies differ significantly, as judged by a contingency table chi-square calculation, we might use this fact to argue that the disease locus is closely linked to this marker.

However such an argument involves a logical error. It might be true that linkage between the two loci leads to association between them, but this does not imply that an observed association implies that the loci are linked, since agencies other than linkage are known to cause association. A case of particular importance is that of population stratification. The population sampled might consist of a mixture of two or more subpopulations, and some given marker gene as well as the disease gene might occur at high frequency in one subpopulation but not the other. This will imply a high value for the coefficient of association between the genes at the marker and the disease loci, but clearly this association does not necessarily imply linkage of the marker to the disease. This is a cause for concern for linkage analysis in stratified populations deriving from the admixture of recently arrived immigrant groups, and as a result, the case-control method has fallen out of favor as a method for linkage analysis. Interest moved, instead, to linkage analyses using marker gene sharing properties among affected sibpairs - the third broad method used, historically, to test for linkage between disease and marker loci.

As with case-control studies, sharing methods also focus on some marker locus "M", of known location, and some putative disease locus "D", of unknown location. In formal statistical language, we want to test the null hypothesis the marker locus "M" is unlinked to the disease locus "D" against the alternative hypothesis that the marker locus "M" is linked to the disease locus "D", the interesting case being that disease and marker loci are closely linked.

This method operates by considering sharing properties of marker genes among affected sib pairs. If the marker and disease loci are linked, we expect an excess over random expectation of sharing of marker genes by affected sibs, the argument being the following. Suppose as a simplifying example that the disease is rare and recessive. Then if the two sibs in a family are both affected by the disease, the most likely situation is that both affected sibs received the disease gene D_1 from their (heterozygous D_1D_2) parents, and thus they both tend to share the same marker genes at any marker locus closely linked to the disease locus. If marker and disease loci are unlinked, there will be no excess over the 50% random expectation of sharing of marker genes by affected sibs, apart from those caused by random statistical fluctuations. Note that by bringing the testing procedure "within families" sharing methods overcome the problem of population stratification.

Thus the "sharing" method of testing for linkage results in a simple "$p = \frac{1}{2}$" binomial test, and it has been used very frequently, with substantial success,

over the last thirty years or so for this purpose.

One problem, however, with this method is that there can be "too much" sharing among sibs - leaving aside the sex chromosomes, sibs on average share half their genetic material, and this creates significant "noise" in the sharing procedure. This problem is particularly acute for diseases of most current interest, namely complex diseases - for these, it is sometimes hard to pick out the often faint signals of linkage for complex diseases amidst this quite large noise.

This problem, and the problem of stratification for case-control studies, led my colleague R. S. Spielman and me to propose the transmission/disequilibrium test (TDT), and later the sib transmission/disequilibrium test (S-TDT), as a procedure that attempts to combine the beneficial features of the "sharing" and case-control methods, while at the same time overcome their disadvantages [Spielman et al. (1993), Spielman and Ewens (1998)]. The main aim of this paper is to describe these tests and to discuss statistical questions to which they give rise.

4. The TDT: introduction. The thinking behind the TDT goes back to the previously discarded idea of testing for linkage via the case-control concept of association. However, it uses case-control ideas in such a way as to overcome the stratification problem described above. This is done by using, as the "control" with whom we compare the marker genetic makeup of an affected individual, the "non-person" created by the genes "thrown away" by the two parents when the affected child was conceived. For example, suppose that the mother of a child affected by the disease is of genotype M_1M_3 at some marker locus and the father is of genotype M_2M_4. If the affected sib is of genotype M_1M_2, we know that this "non-person", never conceived, would have been of genotype M_3M_4. It is against this genotype of this "non-person" that we compare the genotype of the affected child actually conceived. This "within family" matching ensures that the TDT overcomes the case-control population stratification problem.

5. The TDT: details and properties. The introduction to the TDT procedure given above shows that the test reduces, in essence, to a comparison of what gene is transmitted and what gene is "thrown away", at a marker locus, by the parent of an affected child. If there are two possible alleles, M_1 and M_2, at the marker locus, this implies that the data matrix appropriate for the TDT will be as in Table 1, where the numbers in the table refer to the numbers of parents (of n affected children) in each of the four possible categories shown. We denote the respective probabilities that the parent of an affected child falls into one or other of the four cells of Table 1 by $P(1,1), P(1,2), P(2,1)$ and $P(2,2)$. When population stratification exists, then even if the mode of inheritance of the

TABLE 1

Combinations of transmitted and nontransmitted marker alleles M_1 and M_2 among the parents of n affected children.

		Non-transmitted allele		
		M_1	M_2	Total
Transmitted allele	M_1	n_{11}	n_{12}	n_1
	M_2	n_{21}	n_{22}	n_2
Total		n_1	n_2	$2n$

disease is known, there are many unknown parameters defining these four probabilities. For example, for a recessive disease the probabilities $P(1,1)$, $P(1,2)$, $P(2,1)$ and $P(2,2)$ are

$$
\begin{aligned}
P(1,1) &= [\sum_j \alpha_j\{p_jq_j^2 + \delta_jq_j\}p_j]/[\sum_j \alpha_jp_j^2] \\
P(1,2) &= [\sum_j \alpha_j\{p_jq_j(1q_j) + \delta_j(1\theta q_j)\}p_j]/[\sum_j \alpha_jp_j^2] \\
P(2,1) &= [\sum_j \alpha_j\{p_jq_j(1q_j) + \delta_j(\theta q_j)\}p_j]/[\sum_j \alpha_jp_j^2] \\
P(2,2) &= [\sum_j \alpha_j\{p_j(1q_j)^2 + \delta_j(1q_j)\}p_j]/[\sum_j \alpha_jp_j^2]
\end{aligned} \tag{5.1}
$$

In the expressions in (5.1), the summation is over different strata in the population, α_j is the proportion of the population in stratum j, q_j and p_j are the frequencies of M_1 and the disease gene respectively in stratum j, δ_j is the coefficient of association in stratum j, and θ is the recombination fraction between marker and disease loci, that is, the probability that the genes inherited from a given parent at the two loci come from different chromosomes in that parent. When the two loci are unlinked $\theta = \frac{1}{2}$, while when they are linked, $\theta < \frac{1}{2}$.

When the sample is taken from a population into which individuals from these strata have migrated, with subsequent intermarriage, the expressions for $P(1,1), P(1,2), P(2,1)$ and $P(2,2)$ become even more complicated. The above expressions, however, are sufficient to make the points at issue.

Our aim is to test the null hypothesis that disease and marker loci are unlinked, that is that the linkage parameter θ takes the value $\frac{1}{2}$. Initially it might seem, given the large number of unknown parameters, that it will be impossible to test this null hypothesis through a lack of degrees of freedom. This difficulty

is overcome by noting that $P(1,2) = P(2,1)$ when the null hypothesis is true, whatever the population stratification situation, whatever the mode of inheritance of the disease (that is, dominant, recessive, intermediate, etc.), indeed whatever the value of any parameter apart from θ might be. This comment is also true when the population sampled is one where migration from different strata, followed by intermarriage, has occurred. Further, in practice, this is the only constraint on the four cell probabilities that will occur when the null hypothesis is true.

Since there are many more parameters than degrees of freedom, the test of the null hypothesis that disease and marker loci are unlinked, using the data of Table 1, reduces to a test of the hypothesis $P(1,2) = P(2,1)$. Standard likelihood-ratio methods lead to a test statistic that depends only on n_{12} and n_{21}, and which is asymptotically given by the simple formula

$$\frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \tag{5.2}$$

Under the null hypothesis, and with the appropriate forms of data discussed below, this statistic has approximately a chi-square distribution with one degree of freedom. If one is not satisfied with this approximate test, one can test for significance using an exact binomial procedure (the McNemar test). The main statistical properties of the TDT test procedure are as follows.

(i) The most important question about the TDT concerns the circumstances under which it is a valid test (i.e. has the nominated Type I error). The motivation for developing the TDT was to eliminate the possibility of artifactual effects due to population structure. This aim is achieved: the TDT is valid as a test of linkage whatever the population structure might be. This conclusion follows from the fact that the TDT uses only data involving transmissions from heterozygous M_1M_2 parents, and when marker and disease loci are unlinked, the M_1 and M_2 genes are equally likely to be transmitted from such parents, whatever the population structure.

(ii) Data that may be used in the test. The TDT is valid as a test of linkage when the data come from families with one affected offspring, two or more affected offspring, a mixture of the two, or multigenerational, provided special cases such as identical twins are disallowed. This arises because, under the null hypothesis that disease and marker loci are unlinked, transmissions from one heterozygous parent to different affected children are independent, as are transmissions from two heterozygous parents to the same child, as well as multigenerational transmissions.

Combining (i) and (ii), the TDT may be used as a valid test of linkage, no matter what the structure of the population from which the sample was taken,

and no matter what the relationships between the individuals in the sample.

(iii) Unlike many test statistics in genetics using data in a 2×2 table, the TDT is not a 2×2 chi-square test statistic. In fact, of the data in Table 1, the TDT statistic uses only the data values n_{12} and n_{21}. While this conclusion derives from likelihood ratio theory, it has a natural genetical interpretation. It is a standard result of genetics that linkage information can be found only from heterozygous (here M_1M_2) parents, and it is only the data values n_{12} and n_{21} in Table 1 that correspond to such parents.

(iv) Power. There are two reasons why calculation of the power of the TDT test is difficult. First, as is seen in (5.1), there are many unknown parameters involved in the sampling procedure and the power of the test is a function of these. No simple power curve can be drawn describing the power of the test. Second, even the situation described in (5.1), with many parameters describing the structure of the population from which the sample was drawn, is oversimplified. Diseases of interest currently are "complex", that is are caused in a complex interactive way by the genes at many loci, and the description above relates to only one single disease gene locus. In such a case the calculation of a unique power of the test is, in practice, impossible.

From the statistical point of view, it is not known whether the TDT test is uniformly most powerful (in comparison with other tests using the data of Table 1). This is a matter that deserves investigation.

(v) Association. There are several properties of the TDT test which revolve around the concept of association, two of which are explored here.

First, the TDT test has no power unless there is association between the genes at the marker locus and those at the disease locus. This can be most easily be seen in the case where there is no population stratification, for which the probabilities in the four cells in Table 1 are of the form

$$
\begin{aligned}
P(1,1) &= q^2 + \frac{q\,\delta}{p} \\
P(1,2) &= q\,(1-q) + \frac{1-\theta-q\,\delta}{p} \\
P(2,1) &= q\,(1-q) + \frac{(\theta-q)\,\delta}{p} \\
P(2,2) &= (1-q)^2 - \frac{(1-q)\,\delta}{p}
\end{aligned}
\tag{5.3}
$$

In equations (5.3), p is the frequency of the disease gene in the population, q the frequency of the marker allele M_1, and δ is the coefficient of association between these alleles. Clearly, when $\delta = 0$, there can be no power of a test of the

hypothesis $\theta = \frac{1}{2}$, since in this case the four frequencies in 5.3 are independent of θ.

Second, there is a converse side of this coin. The more association there is, the higher the power of the TDT as a test of linkage. To the extent that structured populations sometimes create association the TDT will use this association and gain increased power. This will not, of course, automatically happen. If the data come from an area with recent immigration from different strata, followed by intermarriage, the association within the original strata together with the association caused by the admixture procedure itself will lead to an overall association that might or might not increase the power of the TDT.

Third, one might want to test the null hypothesis $\delta = 0$, i.e. that there is no association between disease and marker genes, rather than test the hypothesis that disease and marker loci are unlinked. The TDT is a valid test of this hypothesis also, even using data from subdivided populations, provided that only data from simplex families (that is, families with only one affected child) are used in the test. The reason for this requirement is that the null hypothesis distribution of the TDT statistic implicitly assumes independent transmissions from parent to affected offspring, and even when there is no population association, the alleles transmitted from a heterozygous parent to two affected children are not independent. Martin et al. (1997) have overcome this problem by devising a test for association which may use data from affected children within the same family. The comparative properties of this test, and a TDT test where only one affected child are used from each family, are explored and present much statistical interest.

It sometimes causes confusion that, with data from simplex families, the same statistic can be used for two purposes, i.e. as a test of linkage and as a test of association. However, an investigator is free to state up-front what his/her null hypothesis is, and indeed should make such a statement. Two investigators working together and always using the same (simplex) data, one using the TDT as a test of association and one using the TDT as a test of linkage, will accept or reject their respective null hypotheses together (assuming that they use the same Type I error). This is despite the fact that the two background populations are different under these two different null hypotheses, as may be verified, for the case of unstratified populations and a recessive disease, by reference to equations 5.3: when $\delta = 0$ (no association) the four probabilities listed are different from those arising when $\theta = \frac{1}{2}$ (no linkage).

Much the strangest situation arises when multiplex families are used to test for linkage (but not association). Suppose we reject the null hypothesis that disease and marker are unlinked. We may not use these (multiplex) data to test directly for association between disease and marker, but we nevertheless

know that the test for linkage, which rejected the hypothesis of no linkage, only has power if there is such an association. Can we say we have evidence for an association? Here is a paradox that deserves more investigation from statisticians.

(vi) Historical factors. Suppose that the original disease gene mutation occurred some 2500 years ago. This implies that it is about 100 generations since the initial mutation first appeared. In such a case it is interesting to assess what residual association might exist between disease and marker.

Suppose the disease mutation occurred on a M_1-bearing chromosome in a population where M_1 and M_2 were equally frequent, and that the recombination distance of the marker locus M from the disease locus is θ. Then the probability that a chromosome now carries M_1, given it now carries the disease gene D_1, is approximately $e^{-100\theta}$. If $\theta = .001$, this probability is approximately 90%. In other words, we can still expect a high degree of association with the disease for a closely linked marker. On the other hand, the passage of time quickly purifies out associations of marker loci which are not closely linked to the disease, through the recombination process. Since markers are now dense on the genome, we can then hope that the association-based TDT will pick out only those markers that are closely linked to the disease locus.

There is of course a converse side to this coin also, namely that a dense set of markers might be necessary for one marker to be sufficiently close to the disease locus for this historical association to be picked up, and problems of multiple testing, so far unresolved for the TDT, will arise with such a dense set of markers.

There are several population genetic comments of relevance to this discussion. First, if there were multiple origins of the disease mutation the above discussion of course needs some modification. Second, we are studying a *conditional* evolutionary process: since we observe the disease now, it has continued to exist since its mutational origin. The population genetics of conditional processes - here, the condition being that the disease gene has survived so far - is rather different from that of unconditional processes [Ewens (1979)], so that the evolutionary statistics of the disease are quite complicated. Finally, a crucial tool of modern evolutionary population genetics is that of the coalescent. Now the disease genes in present-day sufferers of the disease have these genes coalesce in the original mutation (or a more recent common ancestor), and the coalescent process at a marker closely linked to the disease locus will resemble that at the disease locus. Little population genetics theory is available to study the details of this marker gene coalescent process, although given the current interest in the coalescent process, we can expect that this situation will change rapidly.

6. Discussion points concerning the TDT. There are many interesting statistical discussion points surrounding the TDT, of which we mention here only a few.

First, the test uses aggregated data: the data entries in Table 1 give no indication of the number of families involved and the number of affected sibs in each. How much information is given up in this aggregation? What other test statistics, possibly having properties more desirable than the TDT, could be constructed taking family sizes, or other data, into account? To what extent is the fact that the same (TDT) test statistic (5.2) arises no matter what mode of inheritance is assumed, a matter discussed further below, dependent on this aggregation?

Second, the TDT procedure assumes that parental genotypes are known directly (since only heterozygous parents may be used in the test). Suppose however that parental genotypes are not known directly, as might well be the case for a disease of old age where both parents are likely to be dead. The data then consist only of the marker locus genotypes of the affected sibs. However both parental disease genotypes can be inferred unambiguously to be M_1M_2 if there is at least one M_1M_1 sib and at least one M_2M_2 sib. In such a case, can we proceed as though these genotypes were known directly? The answer is "no", since an ascertainment bias arises for such families. This is most easily seen if there are only two sibs in the family: here one sib must be M_1M_1 and the other M_2M_2. The binomial assumption implicit in the use of (5.2) as a chi-square does not hold for such families, since the variance of the number of M_1 genes among the two affected sibs is zero. Here is a curious case where unambiguous inference has a different consequence from that of direct knowledge.

If parental genotypes are unavailable, what use can be made of unaffected sibs? In the situation just discussed, we may certainly use the TDT if the inference about the unknown parental genotypes is derived from unaffected sibs. But can unaffected sibs be used in a more direct way? Spielman and Ewens (1998) introduce the sib-TDT (S-TDT) test as a means of testing for linkage (and in some cases for association) when parental genotype information is not available but unaffected sib genotype information is. The details of the S-TDT procedure are too long to discuss here, but some statistical points may be mentioned briefly. If both parental and unaffected sib genotype information are available, do we use the TDT or the S-TDT? How do we carry out a combined test when some families provide parental genotype data, some unaffected sib data, some both? These questions are related to a discussion of the relative powers of the tests, itself a complicated matter. If one affected sib per family is the minimal (and only) data allowed in a TDT when used to test for association, what is the corresponding requirement for the S-TDT? These and further questions are

still in the process of being answered.

The sometimes curious statistical questions surrounding the TDT have caused some controversial discussion among statistical geneticists. Here we mention two such points.

First, it is sometimes claimed, since $P(1,2)$ and $P(2,1)$ in 5.3 are equal when $(\theta - \frac{1}{2})\delta = 0$, that the *only* null hypothesis that may be tested by the TDT statistic (5.2) is the hypothesis $(\theta - \frac{1}{2})\delta = 0$. My own view is that this comment is incorrect. A researcher is not only entitled, he/she is indeed obliged, to state up front what his/her null hypothesis and alternative hypothesis are [Lehmann (1986)]. This statement is made on genetical grounds, and depends on the genetical question of interest. It is thus appropriate to state up front, for example, that one's null hypothesis is that disease and marker loci are unlinked, rather than to infer what the hypothesis might be from the form of the parameters involved in the distribution of the test statistic used to test this hypothesis.

Of course this test of hypothesis happens to have no power if $\delta = 0$. We will then reject the null hypothesis with the same (Type I error) probability, whether it be true or not. However this is a different matter from claiming that we are never allowed to choose to test this hypothesis.

The second controversial point relates to the broad question of the relative merits and disadvantages of parametric and non-parametric tests in human genetics, particularly in linkage analysis, a matter hotly debated in some sections of the human genetics community.

The claim has been made that if some parametric test which makes certain assumptions (often about mode of inheritance - whether the disease of interest is recessive or dominant, for example) is identical to some non-parametric test, then in using this non-parametric test one is implicitly making the same mode of inheritance assumption as is made in the parametric test. This claim has been made, for example, by Whittemore (1996) and Greenberg et al. (1996), and challenged among others by Kruglyak (1997).

The TDT procedure bears on this question. Any test that uses the statistic (5.2) can be thought of as a simple non-parametric "$p = \frac{1}{2}$ coin-tossing" test, as is indicated by the discussion below (5.2). However, as noted above, the statistic (5.2) can be derived, and initially was derived, using likelihood ratio theory within the context of a parametric test. Although the theory for a recessive disease only has been discussed above, the same test statistic (5.2) and test procedure result whatever the mode of inheritance of the disease might be. It follows from this that there can be no implication, when using the TDT, about the mode of inheritance of the disease. This argues in favor of Kruglyak's challenge of the Whittemore and Greenberg et al. claim described above. Of course

TABLE 2

Combinations of transmitted and nontransmitted marker alleles $M_1, M_2, \ldots, M_k$ among the parents of n affected children.

		Non-transmitted allele				
		M_1	M_2	...	M_k	Total
Transmitted allele	M_1	n_{11}	n_{12}		n_{1k}	n_1
	M_2	n_{21}	n_{22}		n_{2k}	n_2
	⋮					
	M_k	n_{21}	n_{22}		n_{kk}	n_k
Total		n_1	n_2		n_k	$2n$

the overall question of when one should use a non-parametric test in linkage analysis and when one should use a parametric test is extremely complicated, and the above comments are not intended to bear on this question in a general way. They are intended solely to discuss one aspect of this matter.

7. Generalizations and extensions of the TDT: many marker alleles. In the above discussion it has been assumed that there are only two possible alleles that can arise at the marker locus. Usually however there are more than two possible alleles at any marker locus, and we now discuss the theory applying to the general k-allele case. This generalization raises several questions of statistical interest, some of which are still unanswered.

We denote the marker alleles $M_1, M_2, \ldots, M_k$. Then the natural generalization of Table 1, applying in the k-allele case, is Table 2.

Consider first the question of using the data in this table in a test for linkage between disease and marker loci. What do we use as test statistic?

There will be a $k \times k$ table of probabilities defining the probabilities for the entries in this table, generalizing those given in (5.1), containing perhaps hundreds of parameters, especially in stratified populations. This table of probabilities is thus immensely complicated. However, as with the probabilities in (5.1), it is possible to make one clear statement when the null hypothesis, that disease and marker loci are unlinked, is true. This is that symmetrically opposite probabilities in this table will be equal: in other words, under the null hypothesis that disease and marker loci are unlinked, symmetrically opposite entries in Table 2 have the same mean value.

It is thus tempting to test this null hypothesis by testing for symmetry in the data matrix. The standard methods to do this uses a test with $k(k-1)/2$

df. However it is generally agreed in the genetic context that it is not wise to use this test. It has too many degrees of freedom, and therefore comparatively uninformative marker alleles might well swamp information about a real linkage. We are thus in a position analogous to that sometimes arising in complex ANOVA and MANOVA designs, that use of formal statistical testing theory is not useful and a more ad hoc approach is required.

One broad category of such ad hoc tests uses the marginal data in Table 2. We test the hypothesis that disease and marker loci are unlinked by testing whether the vector of row marginal totals $(n_{1.}, n_{2.}, \ldots, n_{k.})$ differs significantly from the vector $(n_{.1}, n_{.2}, \ldots, n_{.k})$ of column marginal totals. Such a comparison has the potential to lead to a useful test of the hypothesis that disease and marker loci are unlinked, since under this hypothesis the two vectors have the same mean vector. Any test based on these marginal totals has $k-1$ df and thus largely removes the swamping effect inherent in the $k(k-1)/2$ df "symmetry" test. Nevertheless, something in general must be lost in adopting a $k-1$ df test, since these two vectors might, in some hypothetical situation, have the same mean vector even when disease and marker loci are linked. Thus this new procedure seems to be testing a weaker hypothesis than that of symmetry. It would be important to know under what genetical circumstances the two forms of test are equivalent.

There are several tests available in the literature of the linkage hypothesis which use these marginal totals. Sham and Curtis (1995) and Duffy (1995) provide log-linear computer-intensive tests and Harley et al. (1995) provide a computer-intensive logistic regression model. We focus here however on three other test statistics that have appeared in the literature and which are direct analogues of the two-allele TDT in that they reduce to the two-allele TDT when $k = 2$.

The first of these is known in the statistical literature as the Stuart statistic [Agresti (1990)]. In the genetical literature this is known as the generalized TDT (or GTDT) test statistic of Schaid (1996).

Using this statistic one first calculates, for $i = 1, 2, \ldots, k$, the quantities d_i, defined by $d_i = n_{i.} - n_{.i}$. The sum of these quantities is necessarily zero, so that without loss of information one ignores one arbitrarily chosen marker allele (say allele k) and forms a vector $\mathbf{d}'$ defined by

$$\mathbf{d}' = (d_1, d_2, \ldots, d_{k-1})$$

If the null hypothesis that disease and marker loci are unlinked is true, the estimate of the variance of d_i is $n_{i.} + n_{.i} - 2n_{ii}$ and the estimate of the covariance between d_i and d_j is $-(n_{ij} + n_{ji})$. These variance and covariance estimates are

formed into a matrix $\mathbf{V}$ and the GTDT test statistic is then defined as

$$\text{GTDT} = \mathbf{d}'\mathbf{V}^{-1}\mathbf{d}$$

Note that, as required for a test of linkage, this statistic does not use the values $n_{11}, n_{22}, \ldots, n_{kk}$, since these values cancel out in the definition of $\mathbf{d}'$ and $\mathbf{V}$. Under the null hypothesis that disease and marker loci are unlinked, the GTDT statistic has asymptotically a chi-square distribution with $k-1$ df, and thus can be used as a test statistic for linkage by referring the observed value of the statistic to tables of significance points of this distribution. The data that may be used in this test are identical to those for which the TDT may be used in a test for linkage.

Although the GTDT statistic is possibly the most natural generalization of the two-allele TDT statistic, its calculation requires the inversion of a large and possibly sparse matrix, and indeed the inverse matrix might not exist for small data sets. It is thus useful to have a test statistic that is very similar to GTDT but which does not involve inversion of a matrix. Such a statistic was proposed by Spielman and Ewens (1996). This statistic is W, defined by

$$W = \{(k-1)/k\} \sum_i \left[(n_{i.} - n_{.i})^2/(n_{i.} + n_{.i} - 2n_{ii}\right]$$

As with the GTDT statistic, this statistic also does not use the data values $n_{11}, n_{22}, \ldots, n_{kk}$. It also has a distribution very close to chi-square with $k-1$ df under the null hypothesis of no linkage, and like the GTDT statistic, also reduces to the two-allele TDT statistic when $k = 2$. The data that may be used in this test are identical to those for which the TDT may be used in a test for linkage.

If $n_{ij} + n_{ji}$ is the same for each independent (i, j) pair, W and GTDT are identical. In this case W can be expressed as the sum of squares of $k-1$ random variables with null hypothesis mean 0 and variance 1, each a linear function of a binomial random variable, verifying the chi-square approximation.

Although the tests based on GTDT and W to some extent overcome the "swamping" problem associated with tests with a large number of degrees of freedom, they do not completely overcome this difficulty. This leads us to consider a third test statistic, also introduced by Schaid (1996), called maxTDT. This statistic is computed as follows. For each i, $(i = 1, 2, \ldots, k)$ we lump all alleles other than allele i as "non-i" and compute a "two-allele" TDT statistic as prescribed in (5.2). We then choose as test statistic the largest of the k TDT statistics so formed, denoting this statistic maxTDT. In terms of the entries in Table 2, this statistic is the largest, as i takes successively the values $1, 2, \ldots, k$,

of the quantities

$$\frac{(n_{\mathrm{i}\cdot} - n_{\cdot\mathrm{i}})^2}{n_{\mathrm{i}\cdot} + n_{\cdot\mathrm{i}} - 2n_{\mathrm{ii}}} \tag{7.4}$$

This test statistic also reduces to the TDT statistic when $k = 2$.

Use of the maxTDT statistic largely avoids the swamping effect mentioned above. However one may not use chi-square tables to test for its significance, since such a deliberately chosen largest TDT statistic does not have a chi-square distribution. Nor are simple Bonferroni corrections for the significance points completely accurate for this test, because there is a simple linear constraint between the terms that are squared in the numerator of each such statistic.

This is most easily seen in the case referred to above where $n_{ij} + n_{ji}$ is the same for all (i, j) pairs. In this case the denominator term in (7.4) is independent of i. Now we have noted that the sum of the terms which are squared in the numerator in (7.4) is identically zero, so that for the case we consider the statistics in (7.4) are the squares of terms which must add to zero. Approximate significance points of the maxTDT statistic can then be found by an adaptation of the significance points given in Table 25 in Pearson and Hartley (1965). These values have been confirmed by a binomial permutation procedure [Ewens and Spielman (1997)].

This comment leads us to another class of k-allele TDT tests, namely permutation tests. These been studied in the TDT context especially by Kaplan, Weir and Martin (1997).

In a permutation procedure we consider any "sensible" test statistic computed from the data of Table 2, for example one or other of the statistics GTDT, W and maxTDT. We then permute the data in some way a large number of times, compute the same statistic for each permutation, and then declare the data to be significant at (say) the 5% level if the observed value of the test statistic is within the most extreme 5% of values found under permutation. An equivalent procedure is to find empirical tables of significance points from a large number of permutations, and this is the procedure adopted by Ewens and Spielman (1997) for the statistic maxTDT.

This approach raises the question of what one permutes over. In the tables presented by Ewens and Spielman (1997), the value of $n_{ij}+n_{ji}$ was kept constant in each permutation. In other words, the permutation consisted of a collection of binomial permutations, where for each permutation and each (i, j) pair, the value of n_{ij} was found from a binomial distribution with parameters $n_{ij} + n_{ji}$ and $\frac{1}{2}$.

There were two reasons for adopting this approach, one genetical and one statistical. The genetical reason is that the investigator has the power to choose

the value of $n_{ij} + n_{ji}$, which is the number of heterozygous M_iM_j parents in the sample, and may well choose these numbers to be roughly equal in an attempt to maximize the power of the test used. In such a situation it is reasonable to regard $n_{ij} + n_{ji}$ as fixed and not subject to random variation, and thus not to bevaried in a permutation test.

The statistical reason for keeping $n_{ij} + n_{ji}$ constant under permutation is that this sum is a sufficient statistic, in a multinomial distribution relating to all k^2 entries in Table 2, for the probability that an entry falls into one or other of the (i, j) or the (j, i) cells. This probability is not dependent on θ, the parameter being tested, so conditioning on a sufficient statistic for this probability is presumably appropriate.

Several points of statistical interest arise from these considerations. What is the best test statistic, of those we have considered, for testing for linkage? As an associated question, given that several interesting genetical questions do not fall in the formalities of statistical theory, when should we ignore formal theory as not being, in practice, the most suitable ones to use? What is the distribution of maxTDT? What are the best permutation procedures for k-allele permutation tests of linkage?

Acknowledgements. I wish to acknowledge the many excellent suggestions of an anonymous referee.

REFERENCES

AGRESTI, A. (1990). *Categorical Data Analysis.* Wiley, New York.

DUFFY, D. L. (1995). Screening a 2 cM genetic map for allelic association: a simulated oligogenic trait. *Genetic Epidemiology* **12** 595–600.

EWENS, W. J. (1979). *Mathematical Population Genetics.* Springer-Verlag, Berlin.

EWENS, W. J. and SPIELMAN, R. S. (1997). Disease associations and the transmission/disequilibrium test (TDT). *Current Protocols in Human Genetics* **Supplement 15** 1.12.1–1.12.13.

GREENBERG, D. A., HODGE, S. E., VIELAND, V. J. and SPENCE, M. A. (1996). Reply to Farrall. *American Journal of Human Genetics* **60** 738.

HARLEY, J. B., MOSER, K. L. and NEAS, B. R. (1995). Logistic transmission modeling of simulated data. *Genetic Epidemiology* **12** 607–612.

KAPLAN, N. L., MARTIN, E. R. and WEIR, B. S. (1996). Power studies for the transmission/disequilibriumtest with multiple alleles. *American Journal of Human Genetics* **60** 691–702.

KRUGLYAK, L. (1997). Nonparametric linkage tests are model free. *American Journal of Human Genetics* **61** 254–255.

LEHMANN, E. L. (1986). *Testing Statistical Hypotheses.* Wiley, New York.

MARTIN, E. R., KAPLAN, N. L. and B. S. WEIR. (1997). Tests for linkage and association in nuclear families. *American Journal of Human Genetics* **61** 439–448.

PEARSON, E. S. and HARTLEY, H. O. (1965). *Biometrika Tables for Statisticians.* Cambridge University Press, Cambridge.

SCHAID, D. J. (1996). General score tests for associations of genetic markers with disease using cases and their parents. *Genetic Epidemiology* **13** 423–450.

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

SPIELMAN, R. S., MCGINNIS, R. E. and EWENS, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus. *American Journal of Human Genetics* **52** 506–516.

SPIELMAN, R. S. and EWENS, W. J. (1996). The TDT and other family-based tests for linkage disequilibrium and association. *American Journal of Human Genetics* **59** 983–989.

SPIELMAN, R. S. and EWENS, W. J. (1998). A sibship test for linkage in the presence of association: the sib-transmission/disequilibrium test. *American Journal of Human Genetics* **62** 450–459.

WHITTEMORE, A. S. (1996). Genome scanning for linkage: an overview. *American Journal of Human Genetics* **59** 704–716.

DEPARTMENT OF BIOLOGY
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA PA 19104-6018
WEWENS@SAS.UPENN.EDU