

THE QUALITATIVE AND QUANTITATIVE
TRANSMISSION / DISEQUILIBRIUM TESTS (TDTs)

Warren J Ewens

Birmingham, AL

July 2008

These notes should be taken in conjunction with the slides presented at the TDT lecture. References to the slides (as [slide 1], [slide 2], ...) are given in these notes to show the relation between the material in the notes and that in the slides. The slides can thus be taken as giving the main points in the notes.

INTRODUCTION

The (*qualitative*) transmission/disequilibrium test (TDT) was introduced by Spielman et al. [ref. 1] as a test of linkage between a marker locus and a purported disease locus. (*Qualitative* – the data consist of the disease state, “affected” or “not affected”, as contrasted with a *quantitative* measurement.) The aim was to produce a test which is not affected by population stratification. A marker locus is one whose location in the genome is known, and where also the genetic constitution of any individual is also known, and today is often associated with a SNP (single nucleotide polymorphism). The TDT can also be used, with certain forms of data, as a test for association between the alleles the marker locus and those at the disease locus. To introduce the properties of the test when used for either purpose we initially assume [slide 2] that there are two possible alleles, M_1 and M_2 , at the marker locus and two possible alleles, D_1 and D_2 , at a purported disease locus.

In formal statistical terms, we denote the recombination fraction between disease and marker loci by the (unknown) parameter θ [slide 3], and then formally test the null hypothesis $\theta = \frac{1}{2}$ (implying that disease and marker loci are unlinked) against the alternative hypothesis $\theta < \frac{1}{2}$ (implying that disease and marker loci are linked) [slide 4].

Tests involving qualitative data (affected or not affected) can be carried out either by population-based tests or by family-based tests [slide 5]. The most frequently used population-based test is the case-control method, which is now described.

THE CASE-CONTROL TEST

The case-control test uses data in a two-by-two table [slide 6] as shown below.

	M ₁	M ₂	Total
Number of genes in affected individuals (cases)	n_{11}	n_{12}	$2R_1$
Number of genes in unaffected individuals (controls)	n_{21}	n_{22}	$2R_2$

The data consist of a sample of R_1 individuals affected by the disease of interest (cases) and also a sample of R_2 individuals not affected by the disease of interest (controls). Among the cases we count n_{11} M₁ genes and n_{12} M₂ genes, giving a total of $2R_1$ genes in all. Similarly among the controls we count n_{21} M₁ genes and n_{22} M₂ genes, giving a total of $2R_2$ genes in all. The case control test proceeds via a standard 2×2 table chi-square test, not described here. It is important to note that this test is a test of *association* between disease status and possession of M₁ or M₂ genes, and is thus not directly a test of *linkage* between marker and disease. It tests whether the population frequency of M₁ among cases differs significantly from its frequency among controls. Formally, [slide 7] if we define δ by

$$\delta = \text{freq}(D_1 M_1) - \text{freq}(D_1) \times \text{freq}(M_1), \quad (1)$$

where $\text{freq}(D_1 M_1)$ is the frequency of the gamete $D_1 M_1$ in the population, $\text{freq}(D_1)$ is the frequency of the allele D_1 in the population, and $\text{freq}(M_1)$ is the frequency of the of M_1 in the population, the case-control procedure tests the null hypothesis $\delta = 0$.

Association is essentially a statistical concept, relating to various frequencies in some population, and the value of δ might differ in magnitude from time to time even in the same population, or differ from one population to another. The event " δ not zero" implies that the population frequency of M₁ among individuals affected by the disease of interest is not equal to the frequency of M₁ among unaffected individuals. The inequality of these two frequencies is often a

more useful way of stating that there is association between the genes at the disease and marker loci, and testing for such an inequality is the basis of the case-control test for association.

If the case-control test is one of association, why is it used as a test of linkage? The reason is historical [slide 8]. Suppose that D_1 is the disease predisposing allele, and arose by a single mutation from D_2 some time in the past. Suppose also that this mutation happened to arise on an M_1 -bearing gamete. Then at the time of this initial mutation there was the largest possible degree of association between D_1 and M_1 . If the disease and marker loci are very closely linked, this association might persist to some extent at the present day, and be picked up by the two-by-two table test. Then the association test becomes in effect a surrogate test of linkage.

Unfortunately [slide 9], agencies other than linkage are known to cause association, in particular population stratification. The population sampled might consist of a mixture of two subpopulations, and M_1 and D_1 might occur at high frequency in one subpopulation but not the other. This will imply a non-zero δ that is not due to linkage of the marker to the disease. As a result, case-control methods are now used with great care. One approach is to attempt to ensure that the data come from a homogeneous population. Another is to attempt to estimate population stratification structure and allow for it in the testing procedure. A third is the method of “genomic control”, where data from the entire genome are used to assess the significance of any association found between the marker and the disease. A fourth approach, which we now consider, is to abandon *population-based* tests and to use instead *family-based* tests, which overcome the stratification problem.

THE TRANSMISSION / DISEQUILIBRIUM TEST (TDT).

The TDT is a family-based test whose main motivation at the time of its introduction was to avoid complications due to potential population stratification [slide 10]. To simplify the discussion we

consider only the case of family trios (father, mother, and child) [slide 11] where the child is affected by some disease whose genetic basis we wish to discover. Since the child is affected, he/she had at least one disease gene (D_1) transmitted from his/her parents. Only transmissions from heterozygous (M_1M_2) parents are informative, so we consider *only* transmissions from these parents. If the disease locus is not linked to the marker locus (that is, if the null hypothesis $\theta = \frac{1}{2}$ is true), the probability that a heterozygous parent transmits M_1 to an affected child is equal to the probability that such a parent transmits M_2 to an affected child, (both probabilities being $\frac{1}{2}$). However, if the null hypothesis is not true, these two probabilities are *not* necessarily equal. For example, if D_1 is positively associated with (say) the marker locus allele M_1 , there is an increased chance that a parent transmits the marker a gene M_1 along with a disease gene D_1 . It can be shown that when the null hypothesis is not true, the probabilities that a heterozygous parent transmits M_1 compared to M_2 to an affected child differ by an amount $K(1-2\theta)\delta$, [slide 12] where K is a complicated constant depending on marker and disease allele frequencies, δ is the coefficient of association between marker and disease loci and θ is the recombination fraction between marker and disease loci.

It follows that if, in a sample of n family trios where in all trios the child is affected by the disease of interest, the total number of transmissions of M_1 from heterozygous parents differs sufficiently from the total number of transmissions of M_2 from heterozygous parents, we have evidence that the null hypothesis is not true. How can this be quantified? Suppose that we define, in family trio i , the quantity w_i as the number of M_1 genes transmitted to the affected child minus the null hypothesis mean value of this number [slide 13]. For example, [slide 14] if both parents are heterozygotes, this mean is 1, and thus w_i is either +1 (child is M_1M_1), 0 (child is M_1M_2) or -1 (child is M_2M_2). The sum of the w_i values taken over all the n family trios thus gives us information about

whether we should accept or reject the null hypothesis. This idea leads to a test statistic defined in terms of the sum of the w_i values by $z = 2 \sum_i w_i / \sqrt{m}$. Since both large positive and large negative values of $\sum_i w_i$ are of interest, it is preferred to use the square of this statistic, namely

$$4 (\sum_i w_i)^2 / m. \quad (2)$$

This is the (qualitative) TDT statistic [slide 15]. The denominator term m is the number of heterozygous parents in the data set and is a “normalizing” factor having the property that, under the null hypothesis, the TDT statistic (2) has approximately a chi-square distribution with one degree of freedom. This leads to a formal statistical test of the null hypothesis. The statistic (2) is more frequently written [slide 16] as

$$(n_1 - n_2)^2 / m, \quad (2a)$$

where n_1 is the total number of transmissions of M_1 , and n_2 is the total number of transmissions of M_2 , from all heterozygous parents in the data set.

The TDT test is also valid as a test of linkage [slide 17] when the data come from families with two or more affected offspring. This is so because, under the null hypothesis that disease and marker loci are unlinked, transmissions to different children are independent.

The TDT test has no power unless there is association between the genes at the marker locus and those at the disease locus. This is because the term $(1-2\theta)\delta$ referred to above is 0 when $\delta = 0$, whatever the value of θ . Conversely [slide 18] the larger the association the higher the power of the TDT test (as a test of linkage).

THE TDT AS A TEST OF ASSOCIATION

It is possible that one might want to test for association [slide 19] rather than for linkage. The TDT is a valid test of this hypothesis also, even using data from subdivided populations, provided that only data from families with only one affected child are used in the test. It is not directly applicable

to families with more than one affected child. The reason for this 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 genes transmitted from a heterozygous parent to two affected children are not independent if marker and disease loci are linked. Martin et al. [ref. 2] have overcome this problem by devising a test for association which allows the use of data from several affected children within the same family by forming a "t"-like statistic instead of a z-like statistic.

GENERALIZATIONS 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. When there are k alleles M_1, M_2, \dots, M_k possible at the marker locus there is no clear-cut approach to testing for linkage (and association). One approach is to use the maxTDT statistic, computed as follows. For each i , ($i = 1, 2, \dots, k$) we group all alleles other than allele i as "non- i " and compute a "two-allele" TDT statistic as prescribed in (2) above. The maxTDT statistic is the largest of the k TDT statistics so formed. This test statistic reduces to the TDT statistic above when $k = 2$. One may not use chi-square tables to test for its significance, since such a deliberately chosen largest TDT statistic does not have a null hypothesis chi-square distribution. Approximate significance points of the maxTDT statistic are given in Ewens and Spielman [ref. 3].

GENERALIZATIONS OF THE TDT: THE SIB-TDT

The TDT as described above uses data from families where marker genotypes are available for father, mother, and affected offspring. When diseases with onset in adulthood or old age are studied,

it may be impossible to obtain genotypes for markers in the parents of the affected offspring. This difficulty has limited the applicability of the TDT. Several methods using genotype information from unaffected sibs are available for this situation, but we do not give the details here.

QUANTITATIVE TDTs

Quantitative trait transmission/disequilibrium tests (quantitative TDTs) are now commonly used [slide 20] in family-based genetic association studies of quantitative traits. There are many quantitative QTDT procedures available in the literature, and here we can only consider some of them. By “quantitative” we mean that we have some quantitative measurement (e.g. BMI) [slide 21] for the child in each of the n trios in the data, instead of the qualitative (affected / not affected) information on the child. In these notes the quantitative measurement is denoted by Y (when it is taken as a random variable) and by y (when it is taken as being non-random). In broad terms, quantitative TDTs consider the relation between the observed quantitative measurement y and the observed transmission information w , which has the same interpretation as in the qualitative TDT, namely the difference between the actual number of M_1 genes passed on to by the parents of each affected child and the null hypothesis mean number passed on (slide [22]). When this number is taken as a random variable it is written in upper case (W).

The quantitative TDTs discussed here are all available in the frequently used QTDT and FBAT packages. They can be divided into two main groups. The first group involves *regression* models [slide 23], in which the measurement Y is taken as a random variable, and the independent variable in the regression models is w . In all but one of the regression models discussed the parental mating type is also taken as an independent variable in the regression. The second group are *non-*

regression, and in contrast to the regression models, take W as a random variable and y as non-random. These are discussed later.

Regression models. For regression models it is assumed that the marker genotypes of all members in each of the n family trios in the data are known, as is the transmission quantity w and the quantitative measurement y in each child. In trio i , the observed value of this measurement is denoted by y_i [slide24], and is assumed to depend on the value w_i for that trio as well as the parental mating type for that trio. It is also taken to have some (unknown) variance σ^2 , the same for all trios. The five principle regression models [refs.4 and 5] in the QTDT package all take the measured quantities as dependent (i.e. random) variables, therefore denoted in upper case as Y , and are as follows.

Allison “linear” model.

$$\text{For } M_1M_1 \times M_1M_2 \text{ parental mating type: } Y = \mu + \beta w + E,$$

$$\text{For } M_1M_2 \times M_1M_2 \text{ parental mating type: } Y = \mu + \alpha_1 + \beta w + E,$$

$$\text{For } M_2M_2 \times M_1M_2 \text{ parental mating type: } Y = \mu + \alpha_2 + \beta w + E.$$

Allison “quadratic” model.

$$\text{For } M_1M_1 \times M_1M_2 \text{ parental mating type: } Y = \mu_1 + \beta_1 w + \beta_2 w^2 + E,$$

$$\text{For } M_1M_2 \times M_1M_2 \text{ parental mating type: } Y = \mu_2 + \beta_1 w + \beta_2 w^2 + E,$$

$$\text{For } M_2M_2 \times M_1M_2 \text{ parental mating type: } Y = \mu_3 + \beta_1 w + \beta_2 w^2 + E.$$

Abecasis “orthogonal” model.

$$\text{For } M_1M_1 \times M_1M_2 \text{ parental mating type: } Y = \mu + \beta w + E,$$

$$\text{For } M_1M_2 \times M_1M_2 \text{ parental mating type: } Y = \mu + \alpha + \beta w + E,$$

$$\text{For } M_2M_2 \times M_1M_2 \text{ parental mating type: } Y = \mu + 2\alpha + \beta w + E.$$

Abecasis “within only” model.

For all parental mating types: $Y = \mu + \beta w + E$.

Abecasis “dominance” model (Ab-Dom).

For $M_1M_1 \times M_1M_2$ parental mating type: $Y = \mu + \beta_1 w + \gamma d + E$,

For $M_1M_2 \times M_1M_2$ parental mating type: $Y = \mu + \alpha + \beta_1 w + \gamma d + E$,

For $M_2M_2 \times M_1M_2$ parental mating type: $Y = \mu + 2\alpha + \beta_1 w + \gamma d + E$.

(In this model, $d = -1$ for a homozygous child and $+1$ for a heterozygous child, and corresponds to W_d in the QTDT package documentation. For the data that we consider, the B_d term in QTDT package documentation for this model is a constant across the three mating types, and is thus absorbed into the constant μ .) In all five models Greek symbols describe unknown parameters and E is a random residual term having mean zero and (unknown) variance σ_Y^2 .

The Abecasis models have some deficiencies. The “orthogonal” model is aimed at separating (making orthogonal, or independent) the effect of parental mating type and the transmission information w . However it assumes a possibly unrealistic linear effect of parental mating type. The dominance model confounds a test of transmission with a test of dominance, so we therefore do not consider it further here. For simplicity, we consider here only the first of the two Allison models as well as the Abecasis “orthogonal” and “within” models. In all of these models [slide 25] the null hypothesis is $\beta = 0$, implying no effect of the transmitted gene at the marker locus on the measurement of interest. This is in effect a test of the null hypothesis that the marker locus is unlinked to any locus having any role in determining the measurement. These three models are nested [slide 26], with the Allison model being the most general and the Abecasis “within” model the most restrictive.

The testing procedures in all three models follow standard multiple regression methods. The null hypothesis model in each case removes a certain sum of squares for the measurement Y , and the full model removes a larger (or in rare cases an equal) sum of squares. The difference between these two sums of squares forms the key component of the numerator of the F statistic used in all testing methods [slide 27]. The use of the F distribution to determine the significance of the observed value of the F statistic is appropriate only if the data have a normal distribution. For cases where the data are taken from one extreme tail of some distribution, for example very large values of the quantitative measurement, this might be an unreasonable assumption.

The aim of a TDT procedure is to overcome problems arising from population stratification. It is thus important to ask: do the above models do this? Suppose that population stratification is completely confounded with mating type, so that all trios where the parental mating type is M_1M_1 come from one stratum (sub-population), all trios where the parental mating type is M_1M_2 come from a second stratum (sub-population), and all trios where the parental mating type is M_2M_2 come from a third stratum (sub-population). Suppose also that for reasons having nothing to do with the marker locus, the mean measurement differs in these three strata. Then the Abecasis models does not necessarily overcome stratification problems. The Allison model, however, will overcome these problems.

Non-regression models. The QTDT package includes both the Rabinowitz and the Monks-Kaplan approaches to quantitative TDT analysis. These procedures do not involve a regression model. They are quite similar to each other, so here we describe only the Rabinowitz [ref. 6] approach. The main property of this approach (and the Monks-Kaplan approach) is that [slide 28], in direct contrast to the regression approaches discussed above, here the transmission quantity W is taken as the random

variable and the quantitative measurement y is taken as non-random. Rabinowitz defines y_i^* by $y_i^* = y_i - \bar{y}$, where \bar{y} is the average of the y_i values taken over the n children in the data. His test statistic z is [slide 29]

$$z = \frac{\sum y_i^* w_i}{\sqrt{\sum (y_i^*)^2 \sigma_i^2}} \quad (3)$$

In this expression the sum is taken over $i = 1, 2, \dots, n$ and σ_i^2 is the null hypothesis variance of the random transmission value W_i in trio i . This variance depends on the parental mating type in that trio and is easily established.

Under the null hypothesis the mean of the numerator in this statistic is 0, and the denominator is the null hypothesis standard deviation of the numerator. This is why the statistic is written as a z . Central limit theorem arguments then show that if n exceeds about 20, the statistic has an approximate $N(0,1)$ distribution when the null hypothesis is true.

There is an interesting relationship between regression-based tests and the Rabinowitz test. This can be seen by considering the hypothesis testing procedure in a “role-reversal” regression model of the form

$$W_i = \alpha + \beta y_i^* + E_i. \quad (4)$$

The estimate of β in this regression is $\sum w_i y_i^* / \sum (y_i^*)^2$, and the statistic testing for departures of β

from zero is $t = \frac{\sum w_i y_i^*}{s}$, where s is the usual regression estimate of the standard deviation of

$\sum W_i y_i^*$. The Rabinowitz statistic (3) has the same numerator as that in t but has, in the denominator, the known null hypothesis standard deviation of $\sum W_i y_i^*$ rather than a regression-based estimate of this standard deviation.

There is an important difference between the hypothesis being tested by all the quantitative TDT procedures described above and the original qualitative TDT. The original TDT assesses whether the sum of the w_i values differs significantly from zero. By contrast [slide 30], none of the quantitative TDT procedures described above assess whether the w_i values (or their weighted sum in the Rabinowitz procedure) differ significantly from zero. This can be seen from the fact that they are all unchanged if an arbitrary constant is added to the w_i values. What they do test is whether there is significant *change* in the value of w as y changes (or of y as w changes). This is explicit in the regression procedures and also applies for the Rabinowitz procedure. It can be shown that it is the test of a non-zero *intercept* in the regression (4), namely \bar{w} , that is equivalent to the qualitative TDT test.

The Allison procedures [slide 31] always overcome potential population stratification problems. This is not necessarily true of the Abecasis “within” procedure. The Rabinowitz “non-regression” approach also overcomes stratification problems. An important question is the form of data for which QTDT procedures are advisable. “Extreme” values, such as high BMI measurements, are unlikely to have a normal distribution, and thus the use of F charts for the regression-based tests must be questioned. Since all QTDT methods concern changes in the w values as the measured quantity y changes, they are possibly best suited to random rather than selected data. There are many further issues relating to the use of QTDT procedures that are not taken up in these notes, and not even discussed in the literature, and the theory for QTDT processes is still not complete.

REFERENCES

- [1] 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.
- [2]. Martin, E., Kaplan N.L. and Weir, B.W. (1997) Tests for linkage and association in nuclear families. *American Journal of Human Genetics* 61:439-448.
- [3] Ewens, W.J. and Spielman, R.S. (1999) Disease associations and the transmission/disequilibrium test (TDT). *Current protocols in Genetics* 20, Pp. 1.12.1 – 1.12.15.\
- [4] Allison DB (1997) Transmission-disequilibrium tests for quantitative traits. *Am J Hum Genet* 60:676-690.
- [5] Abecasis GR, Cardon LR, Cookson WO (2000) A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* 66:279-292.
- [6] Rabinowitz D (1997) A transmission disequilibrium test for quantitative trait loci. *Hum Hered.* 47:342-350.

