Genetic Epidemiology 21:53–67 (2001)

A General and Accurate Approach for Computing the Statistical Power of the Transmission Disequilibrium Test for Complex Disease Genes

Wei-Min Chen[1] and Hong-Wen Deng[1,2]*

[1]*Osteoporosis Research Center & Department of Biomedical Sciences, Creighton University, Nebraska*
[2]*Laboratory of Statistical and Molecular Genetics, College of Life Sciences, Hunan Normal University, ChangSha, Hunan, P. R. China*

Transmission disequilibrium test (TDT) is a nuclear family–based analysis that can test linkage in the presence of association. It has gained extensive attention in theoretical investigation and in practical application; in both cases, the accuracy and generality of the power computation of the TDT are crucial. Despite extensive investigations, previous approaches for computing the statistical power of the TDT are neither accurate nor general. In this paper, we develop a general and highly accurate approach to analytically compute the power of the TDT. We compare the results from our approach with those from several other recent papers, all against the results obtained from computer simulations. We show that the results computed from our approach are more accurate than or at least the same as those from other approaches. More importantly, our approach can handle various situations, which include (1) families that consist of one or more children and that have any configuration of affected and nonaffected sibs; (2) families ascertained through the affection status of parent(s); (3) any mixed sample with different types of families in (1) and (2); (4) the marker locus is not a disease susceptibility locus; and (5) existence of allelic heterogeneity. We implement this approach in a user-friendly computer program: *TDT Power Calculator*. Its applications are demonstrated. The approach and the program developed here should be significant for theoreticians

Contract grant sponsor: NIH; Contract grant number: R01 GM60402-01A1; Grant sponsor: Health Future Foundation; Grant sponsor: Hunan Normal University.

*Correspondence to: Hong-Wen Deng, Ph.D., Osteoporosis Research Center, Creighton University, 601 N. 30th St., Omaha, NE 68131. E-mail: deng@creighton.edu

Received for publication 1 May 2000; Accepted 13 September 2000

to accurately investigate the statistical power of the TDT in various situations, and for empirical geneticists to plan efficient studies using the TDT. Genet. Epidemiol. 21:53–67, 2001.

Key words: transmission disequilibrium test; linkage; association; power analysis

INTRODUCTION

The transmission disequilibrium test (TDT) [Spielman et al., 1993] was first developed to control for population admixture/stratification in testing for linkage in the presence of association between marker loci and disease susceptibility loci (DSL). It has been argued that the association studies that use the TDT will be the future major approach to search for genes underlying human complex diseases [Risch and Merikangas, 1996]. There have been extensive interests in theoretical development and/or extension of the TDT in recent years [Schaid, 1998]. The TDT has been extended to the situations with multiple sibs in the absence of parental information [Curtis, 1997; Spielman and Ewens, 1998; Boehnke and Langefeld, 1998; Monks et al, 1998; Schaid and Rowland, 1998] and to QTL identification [Allison, 1997; Rabinowitz, 1997; Xiong et al, 1998; Schaid and Rowland, 1999]. Approaches more robust and powerful than the TDT may be developed when the linkage disequilibrium between a marker and a DSL is not strong [e.g., Huang and Jiang, 1999]. When the marker locus is not a DSL, the power of the TDT is much reduced [Müller-Myhsok and Abel, 1997; Xiong and Guo, 1998; Tu and Whittemore, 1999]. The effects of special ascertainment of families through affection status of family members [Whittaker and Lewis, 1998] and allelic heterogeneity [Slager et al, 2000] on the power of the TDT have been investigated.

In each of the above studies, an accurate power computation is crucial in order to make correct comparisons of the powers of the TDT and other extended methods under various scenarios and in order to draw quantitatively (and sometimes qualitatively) correct conclusions. Although it is generally quite accurate to obtain the statistical powers of the TDT and various extended methods via the Monte Carlo simulation approach, the simulation is generally time-consuming, especially with large sample sizes and for some parameters. Hence, it is often not feasible to investigate extensively and systematically the statistical power of the TDT with the Monte Carlo simulations. Various analytical approaches for computing the statistical power of the TDT have been developed [Risch and Merikangas, 1996; Camp, 1997; Whittaker and Lewis, 1998; Knapp, 1999]. However, these power computation approaches suffer some limitations, and some of them are erroneous and misleading. For example, as pointed out by Knapp [1999] and Camp [1999], the approach of Camp [1997] is technically inadequate. Several other approaches may sometimes yield inaccurate analytical results of the power that can be as large as 10% different from the simulated power [Whittaker and Lewis, 1998]. It is noted that, for some of the approaches developed for the power computation, the analytical results were not compared and cross-checked with simulation results. An excellent exception is the approach developed by Knapp [1999]. As demonstrated by simulations [Knapp, 1999], one of his approximations is more accurate than any other then known approaches. However, Knapp's [1999] approach is limited because it requires that (1) the marker locus be a DSL, (2) only families with

single affected offspring (SAO) or an affected sib pair (ASP) be considered, and (3) the disease statuses of both parents be unknown and random.

In this paper, an extension of Knapp's [1999] first approximation is developed to compute the statistical power of the TDT. Our approach is not only highly accurate, but also quite flexible. Some of the situations that our method can handle include (1) families that consist of one or more children and that have any configuration of affected and nonaffected sibs; (2) families ascertained through the affection status of parent(s); (3) any mixed sample with different types of nuclear families in (1) and (2); (4) the marker locus is not a DSL; and (5) existence of allelic heterogeneity. We implement our approach in a user-friendly computer program, *TDT Power Calculator*.

METHODS

Throughout our investigation, as in "classical" TDT analysis [Spielman et al., 1993], we assume (1) a qualitative trait (disease), (2) the marker genotypes available from both parents, and (3) at least one affected child available from a nuclear family. In our power analysis, the study population is assumed to be in Hardy-Weinberg equilibrium. The null hypothesis of the TDT is no linkage, with the assumption of association.

Fundamentals of Our Analytical Approach of Power Computation for the TDT

The fundamentals of our approach were laid out by Knapp [1999]. In nuclear families, the specific marker genotypes of the parents together with those of the child(ren) constitute a specific *type* of family. Assume that there are $(k + 1)$ different *types* of families. For $1 \le i \le k + 1$, let s_i denote the probability that a family is of *type i*, u_i denote the number of the M alleles transmitted from heterozygous parents to their affected offspring for a family of *type i*, and v_i denote the number of other alleles transmitted. Let the vectors $u = (u_i)_{1 \le i \le k}$, $v = (v_i)_{1 \le i \le k,}$ and $s=(s_i)_{1 \le i \le k}$. Let χ^2_{TDT} denote the TDT statistic. Under the alternative hypothesis of linkage, $\sqrt{\chi^2_{TDT}}$ approximately follows a normal distribution, with mean μ_X and variance σ^2_{A1}, where

$$\mu_X = \sqrt{n}\,\frac{e_1 - e_2}{\sqrt{e_1 + e_2}}, \tag{1}$$

$$\sigma^2_{A1} = \frac{4d_1 - (e_1 - e_2)^2}{4(e_1 + e_2)} + \frac{d_{1,2}(e_1 - e_2)}{(e_1 + e_2)^2} + \frac{d_2(e_1 - e_2)^2}{4(e_1 + e_2)^3}, \tag{2}$$

e_1, e_2, d_1, d_2, $d_{1,2}$ are functions of u, v, and s, and n is the number of families [Knapp, 1999].

Now let $\chi^2_{1,x}$ denote the x-quantile of a χ^2 distribution with 1 degree of freedom and let Z_x denote the x-quantile of a standard normal distribution. α is the given significance level of the test. It can be proven that $Z^2_{1-a/2} = \chi^2_{1,1-a}$. Then,

$$\begin{aligned} \text{Power of the TDT} &= \Pr(X^2_{TDT} > \chi^2_{1,1-a}) \\ &= \Pr(X < -Z_{1-a/2}) + \Pr(X > Z_{1-a/2}) \\ &\approx \Phi\left(\frac{-Z_{1-a/2} - \mu_X}{\sigma_{A1}}\right) + \Phi\left(\frac{-Z_{1-a/2} + \mu_X}{\sigma_{A1}}\right), \end{aligned} \tag{3}$$

where $\Phi(x)$ is the cumulative distribution function of a standard normal variable. Obviously, the power of the TDT is a function of the vectors *u, v, s* and the sample size *n*; or, the sample size *n* is a function of the vectors *u, v, s* and the power. Numerical methods such as the bisection method to solve equations can be employed to compute the sample size *n* needed for a specified power.

Family Type Pooling

Here we define the family *class*, which is similar to the concept of *type* used by Knapp (1999). To compute the power correctly, the only requirement for a *class* is that, for each family belonging to *class i*, the numbers of the marker allele M transmitted and not transmitted from heterozygous parents to their affected offspring must be u_i and v_i respectively. Since for different *types* of families, it is possible that $(u_i, v_i) = (u_j, v_j)$, we can pool these two *type*s of families together as one *class*, with a new *class* probability $s_i + s_j$. The above combination of *type*s into *class*es will not affect the statistical power of the sample. To prove it, suppose that among SAO families, there are *k* informative (with at least one heterozygous parent) *type*s with parameters (s_i, u_i, v_i) $(i=1,2,\ldots,k)$. The parental mating type and the child marker genotype of families with the $(k\text{-}1)^{th}$ type are MM*Mm and MM, respectively, and those of families with the k^{th} type are Mm*mm and Mm, respectively. Obviously, $u_k = u_{k-1} = 1$ and $v_k = v_{k-1} = 0$. Therefore, these two *types* can be combined into one *class*. After pooling, there will be *k*-1 informative *class*es with parameters (s_i', u_i', v_i') $(i = 1,2,\ldots,k\text{-}1)$. Parameters for the first *k*-2 *class*es remain unchanged. However, now, $u'_{k-1} = u_{k-1}$, $v'_{k-1} = v_{k-1}$, and $s'_{k-1} = s_{k-1} + s_k$. According to the definition

$$e_1 = \sum_{i=1}^{k} u_i s_i$$

[Knapp, 1999], we have

$$e_1' = \sum_{i=1}^{k-1} u_i' s_i' = \sum_{i=1}^{k-2} u_i' s_i' + u_{k-1}' s_{k-1}' = \sum_{i=1}^{k-2} u_i s_i + u_{k-1}(s_{k-1} + s_k) = \sum_{i=1}^{k-2} u_i s_i + u_{k-1} s_{k-1} + u_k s_k = \sum_{i=1}^{k} u_i s_i$$

Therefore, $e_1 = e_1'$. Similarly, $e_2' = e_2$, $d_1' = d_1$, $d_2' = d_2$, and $d_{1,2}' = d_{1,2}$. According to Equations (1)–(3), the power remains the same before and after the combination of *types* into *classes*.

One *class* may also be divided into several different *class*es with families of different structures for computational ease. A family structure is defined by the number of children and disease status of individuals in a nuclear family. If sampled nuclear families are of different structure, the sample is said to be of a heterogeneous family structure. Suppose that the sample contains families having two different structures (those with SAO and those with ASP), and SAO families of *type i* and ASP families of *type j* have the same (u_j, v_j). Therefore, they may be combined as one *class.* However, for computational ease, this *class* may also be broken into two *class*es, one for the SAO family structure and the other for the ASP family structure. The flexibility of combination of *type*s and the division of *class*es is a powerful tool for our approach. It generalizes the approach of Knapp [1999] so that computation of the probabilities of any *class* in a general situation is relatively easy.

Probability of family class

For the families of the *i*th *type* or *class*, one can count u_i and v_i easily just like in the data analyses with the TDT. To compute s_i, the probability of a family being of the *i*th *class*, we numerate a detailed list of probabilities of the families of all *types*, then pool them into different *classes*.

In the computation, the order of parent is considered. If only one parent is affected, the affected one is called the first parent simply for notational convenience. Let D, M, G, and H denote the DSL allele, marker allele, two-locus genotype, and haplotype of the parents, respectively, and d, m, g, and h denote those of the children respectively. Assume that the genotype of the first parent is G^1, or $D_1^1D_2^1M_1^1M_2^1$, where the superscripts "1" denote the first parent and the subscripts denote the chromatid where the allele is located. The genotype of the second parent is G^2, or $D_1^2D_2^2M_1^2M_2^2$, and that of the s^{th} child is g^s, or $d_1^sd_2^sm_1^sm_2^s$. t children are assumed to be in a family. Disease status is denoted by C^s for the s^{th} child, and P^x for the x^{th} parent. Given the disease status (affected, A; nonaffected, N; unknown, X), the probability that the family members have a specific set of marker genotypes is:

$$\Pr(M_1^1M_2^1, M_1^2M_2^2, m_1^1m_2^1, \ldots m_1^tm_2^t \mid P^1, P^2, C^1, \ldots, C^t) = \frac{\Sigma \Pr(G^1, G^2, g^1, \ldots, g^t, P^1, P^2, C^1, \ldots, C^t)}{\Pr(P^1, P^2, C^1, \ldots, C^t)},$$

where the summation in the numerator is over all possible combinations of genotypes *at the DSL* in parents and children, and the denominator can be obtained from the summation of the numerator over all combinations of genotypes *at the marker* locus. The joint probability in the numerator is:

$$\begin{aligned}
&\Pr(G^1, G^2, g^1, \ldots, g^t, P^1, P^2, C^1, \ldots C^t) \\
&= \prod_{x=1}^{2} \Pr(G^x) \Pr(P^x \mid G^x) . \prod_{s=1}^{t} \Pr(g^s \mid G^1, G^2) \Pr(C^s \mid g^s) \\
&= \prod_{x=1}^{2} [\prod_{k=1}^{2} \Pr(H_k^x)] \Pr(P^x \mid D_1^x D_2^x) . \prod_{s=1}^{t} [\prod_{k=1}^{2} \Pr(h_k^s \mid G^k)] \Pr(C^s \mid d_1^s d_2^s)
\end{aligned}$$

The probabilities in the above equation are defined as follows. The frequency of haplotype $\Pr(H_k^x)$ can easily be computed according to the definition of the linkage disequilibrium coefficient $\delta = \Pr(AM) - \Pr(A)\Pr(M)$. Assume the penetrance of individuals with the genotype *ij* at the DSL is ϕ_{ij}. The probabilities of disease status conditional on the disease genotype of the parents are, respectively, $\Pr(P^x = A \mid D_1^xD_2^x) = \phi_{D_1^xD_2^x}$, $\Pr(P^x = N \mid D_1^xD_2^x) = 1 - \phi_{D_1^xD_2^x}$, and $\Pr(P^x = X \mid D_1^xD_2^x) = 1$. Similarly, for the s^{th} child, $\Pr(C^s \mid d_1^sd_2^s)$ can be expressed in terms of penetrances. The transmission probability for the first haplotype to a child given the first parent's two-locus genotype is

$$\Pr(h_1^s \mid G^1) = \frac{1-\theta}{2}[\Pr(h_1^s = H_1^1) + \Pr(h_1^s = H_2^1)] + \frac{\theta}{2}[\Pr(h_1^2 = D_1^1M_2^1) + \Pr(h_1^s = D_2^1M_1^1)]$$

If families of heterogeneous structures are considered for the power computation, e.g., *20%* of families with SAO and *80%* of families with ASP, one can first

compute s_i and hence e_1, e_2, d_1, d_2, and $d_{1,2}$ in families of each structure separately, then sum the corresponding values for different family structures, each weighted by their proportions in the sample. In this example, assume the e_1 in families with SAO and ASP is e_{11} and e_{12}, respectively. The e_1 in the combined families of these two heterogeneous structures will be $0.2e_{11} + 0.8e_{12}$. e_2, d_1, d_2, $d_{1,2}$ can be computed similarly. Therefore, the analytical power of the TDT can be obtained. Note that we do not assume the number of alleles at the DSL. Therefore, it can handle the case of allelic heterogeneity, which refers to the existence of multiple disease susceptibility alleles at a DSL.

Although the logic of the above approach is simple, its implementation is not efficient due to the extensive computation involved, particularly when the number of children from each nuclear family is large. Therefore, in the Appendix, we develop a more efficient algorithm, which is currently employed in our computer program.

Computer Program

A program "*TDT Power Calculator (PC)*" has been developed to implement our methods, (available at http://www.creighton.edu/~weimin). Not only can this program compute the statistical power or sample size analytically, it also has a built-in simulation module for optional use to obtain the simulated empirical power (to corroborate analytical power) and type I error rates. In this simulation module, nuclear families are simulated under assigned parameters and hypothesis tests are repeated for a given number of times. The simulated power is the proportion of repeated simulations that the null hypothesis is rejected.

To compute the power of the TDT, three kinds of parameters are necessary: (1) Population-wide parameters: disease allele frequencies and genotype penetrances. (If the marker locus is not a DSL, recombination fraction and linkage disequilibrium between the marker and the DSL, and marker allele frequencies should be considered). (2) Parameters for family structures: the number of different family structures in the sample; the total number of children within families, the number of affected children, parental disease status, and the number of families for each kind of family structure. (3) Statistical parameters: significance level and the number of repeated simulations.

RESULTS

Employing the program "*TDT Power Calculator (PC)*", we performed a number of studies, comparing our analytical computation results with simulation results, and our analytical results with those from previous approaches. In Tables I–IV, the numbers given are the sample sizes computed by our analytical approach under the given significance level $\alpha = 5*10^{-8}$ in order to achieve the 80% statistical power. One hundred thousand simulations were performed to test the accuracy of the required sample sizes. The standard error for the simulated power is about *0.0013*. Unless otherwise specified (e.g., Table II), the marker allele is regarded as the disease susceptibility allele. The numbers within parentheses are the simulated power achieved with the sample size computed by our analytical approach. The closer the simulated power to the specified power of 80%, the higher the accuracy. p is the frequency of the disease susceptibility allele. ϕ_{AA}, ϕ_{Aa}, and ϕ_{aa} are the genotype penetrances. ϕ is the disease prevalence. δ is the linkage disequilibrium coefficient. γ is the genotypic relative risk.

TABLE I. Sample Size Necessary to Gain 80% Power in the TDT With SAO And ASP Under a Multiplicative Genetic Model*

		SAO		ASP	
γ[a]	p[b]	R&M	PC	R&M	PC
4.0	0.01	1,098 (0.800)	1,100 (0.804)	235 (0.791)	239 (0.806)
	0.1	150 (0.787)	152 (0.802)	48 (0.777)	49 (0.799)
	0.5	103 (0.783)	105 (0.802)	61 (0.774)	63 (0.805)
	0.8	222 (0.790)	224 (0.800)	161 (0.783)	164 (0.798)
2.0	0.01	5,823 (0.773)	5,991 (0.801)	1,970 (0.766)	2,034 (0.795)
	0.1	695 (0.773)	717 (0.803)	264 (0.769)	273 (0.799)
	0.5	340 (0.767)	352 (0.803)	180 (0.770)	186 (0.800)
	0.8	640 (0.773)	660 (0.802)	394 (0.771)	407 (0.800)
1.5	0.01	19,320 (0.767)	20,019 (0.800)	7,776 (0.771)	8,068 (0.804)
	0.1	2,218 (0.768)	2,300 (0.799)	941 (0.765)	977 (0.800)
	0.5	949 (0.767)	985 (0.798)	484 (0.767)	503 (0.801)
	0.8	1,663 (0.768)	1,725 (0.800)	941 (0.764)	977 (0.799)

*R&M represents the analytical power computation approach of Risch and Merikangas [1996]. PC represents our approach.
[a]γ is the genotypic relative risk.
[b]p is the frequency of the disease susceptibility allele at the DSL.
The numbers within parentheses are the simulated power with the sample sizes computed by our approach.

SAO and ASP

Table I contains the results for families with either SAO or ASP. Method "R&M" was developed by Risch and Merikangas [1996], which was later followed and employed by several others [e.g., Xiong and Guo, 1998]. Method "*PC*" is ours. The results show that our approach is the most accurate one. As correctly pointed out by Knapp [1999], Risch and Merikangas [1996] only considered one-sided normal distri-

TABLE II. Sample Size Necessary to Gain 80% Power in the TDT With SAO When the Marker Allele Is Not a Disease Susceptibility Allele*

	Proportion of maximum δ					
		p	Multiplicative	Recessive	Additive	Dominant
ϕ_{AA}			0.8	0.8	0.7	0.5
ϕ_{Aa}			0.2	0.1	0.37	0.5
ϕ_{aa}			0.05	0.1	0.04	0.05
	0.1	1.0	520 (0.800)	6,302 (0.800)	286 (0.800)	309 (0.799)
		0.8	806 (0.801)	9,797 (0.799)	445 (0.800)	481 (0.800)
		0.6	1,420 (0.800)	17,336 (0.800)	785 (0.801)	849 (0.800)
		0.4	3,165 (0.802)	38,839 (0.800)	1,748 (0.801)	1,894 (0.799)
	0.3	1.0	127 (0.799)	193 (0.800)	161 (0.800)	240 (0.802)
		0.8	197 (0.804)	297 (0.800)	251 (0.801)	375 (0.804)
		0.6	345 (0.798)	519 (0.799)	443 (0.800)	663 (0.802)
		0.4	766 (0.800)	1,147 (0.800)	986 (0.801)	1,482 (0.799)

*The frequency of marker allele is 0.4, and the recombination fraction between the marker locus and DSL is 0. p is the frequency of the disease susceptibility allele. ϕ_{AA}, ϕ_{Aa}, and ϕ_{aa} are the genotype penetrances. ϕ is the disease prevalence. δ is the linkage disequilibrium coefficient.
The numbers within parentheses are the simulated power with the sample sizes computed by our approach.

TABLE III. Sample Size Necessary to Achieve 80% Power in the TDT With SAO, ASP, and DSP With Parents Having Difference Disease Status*

		NN		AN		AA		XX	
		PC	W & L	PC	W & L	PC	W & L	PC	W & L
1									
	SAO	100 (0.798)	108 (0.891)	45 (0.820)	45 (0.819)	126 (0.798)	126 (0.798)	61 (0.799)	64 (0.903)
	DSP	133 (0.795)	146 (0.903)	43 (0.818)	43 (0.821)	107 (0.801)	107 (0.801)	66 (0.797)	73 (0.912)
	ASP	25 (0.813)	26 (0.862)	24 (0.807)	24 (0.809)	74 (0.804)	74 (0.806)	27 (0.798)	26 (0.750)
2									
	SAO	254 (0.802)	258 (0.817)	193 (0.799)	196 (0.813)	156 (0.801)	158 (0.813)	235 (0.800)	240 (0.818)
	DSP	270 (0.800)	275 (0.816)	202 (0.804)	205 (0.817)	160 (0.807)	162 (0.818)	250 (0.801)	255 (0.820)
	ASP	88 (0.805)	90 (0.825)	79 (0.795)	81 (0.818)	72 (0.794)	74 (0.819)	85 (0.805)	87 (0.827)

*W&L denotes the results obtained by the analytical approach of Whittaker and Lewis [1998] and PC denotes those by our approach. The disease status of parents of each nuclear family is denoted as follows: NN, two unaffected parents; AN, one parent affected and the other unaffected; AA, two parents affected; XX, disease status of the two parents is not considered when a family is ascertained.

The numbers within parentheses are the simulated power with the sample sizes computed by our approach.

TABLE IV. Sample Size Necessary to Achieve 80% Power in the TDT With Different Family Structures Under Different Models of Inheritance (MOI)*

	MOI								
	ϕ	p	ϕ_{AA}	ϕ_{Aa}	ϕ_{aa}	S+D	S+A	D+A	S+D+A
Multiplicative	0.1	0.125	0.55	0.19	0.065	242	126	128	150
						(0.799)	(0.814)	(0.813)	(0.807)
Recessive	0.107	0.1	0.8	0.1	0.1	1,568	308	324	441
						(0.801)	(0.803)	(0.802)	(0.805)
Additive	0.14	0.1	0.5	0.3	0.1	338	186	188	219
						(0.801)	(0.807)	(0.801)	(0.801)
Dominant	0.1	0.1	0.13	0.13	0.09	4,080	2,524	2,534	2,895
						(0.802)	(0.805)	(0.791)	(0.805)

*The simulation parameters under different MOI are specified in columns 2–6. In this table, S, D and A denote the types of children in nuclear families: S = SAO, D = DSP, and A = ASP. S+D, S+A, and D+A denote mixture of families with different structures of different types of children, with each family structure composing the same proportion of 50% in the mixture. S+D+A denote the mixture of families with three different structures, with each composing the same proportion of 1/3 in the mixture. The numbers within parentheses are the simulated power with the sample sizes computed by our approach.

bution for approximating the distribution of the TDT statistic. Even if a two-sided normal distribution is considered, the accuracy of the method of Risch and Merikangas [1996] is the same as the second approximation (the less accurate one) of Knapp [1999] for SAO. Serious errors in the paper of Camp [1997] were found by Knapp [1999] and Camp (1999). In some cases, the simulated power based on the calculated sample size with Camp's (1997) approach is less than 0.002 although it is expected to be 0.80. Slager et al. (2000) unfortunately made the same mistake as Camp (1997) in their analytical power computation. Even with the correction and improvement of Camp [1999], her approach is still less accurate than Knapp's [1999] or ours.

Marker Allele Is Not a Disease Allele

The sample size needed increases at an accelerating rate in two situations (Table II): (1) when allele frequency of the marker deviates increasingly from that of the DSL; and (2) when the degree of linkage disequilibrium deviates increasingly from its maximum. This is true even if the marker locus and the DSL are so closely linked that the recombination between them is essentially zero. The case when there is recombination between the marker locus and the DSL shows qualitatively the same conclusion (data not shown). Muller-Myhsok and Abel [1997] and Xiong and Guo [1998] pointed out this phenomenon, and Tu and Whittemore [1999] elaborated on it with detailed analyses. Table II also shows that our approach is extremely accurate and robust.

Parental Disease Status Is Considered

In Table III, in the first situation, the penetrances for AA, Aa, aa are 0.77, 0.77, and 0.028, respectively. P is 0.05, and the disease prevalence is 0.1. In the second situation, the penetrances for AA, Aa, and aa at the DSL are 0.55, 0.19, and 0.07, respectively. P is 0.125, and the disease prevalence is 0.10. They correspond to the first and eighth situations respectively considered in Whittaker and Lewis [1998].

Table III shows the comparison of two different approaches (ours and that of Whittaker and Lewis [1998]) to compute the sample size when parental disease sta-

tus is (1) unknown (XX), (2) neither parent affected (NN), (3) only one parent affected (AN), or (4) two parents affected (AA). In the situation investigated by Whittaker and Lewis [1998], nuclear families have (1) SAO, (2) DSP (discordant sib pairs, i.e., one affected and the other unaffected), or (3) ASP. Our results show that there are significant differences between the accuracy of the two approaches. As is confirmed by the simulations, all the results on the sample sizes computed using our *TDT Power Calculator* when the parental disease status is considered are much more accurate. For the first case of dominant inheritance considered in Whittaker and Lewis [1998], the sample sizes given by their analytical approach for 80% power can often yield 90% power in simulations. These results indicate that Whittaker and Lewis's analytical approach [1998] is inaccurate and overestimates the required sample sizes.

Multiplex Families

TDT Power Calculator can be employed to demonstrate the effect of the number of children on the power of the TDT (Fig. 1). The larger number of affected children from each family can help increase the power, particularly when the required sample size is huge and the number of affected children in each family is small. However, the increasing rate is not linear. For example, in the recessive situation in Figure 1, when families with three affected children are recruited, the sample size is reduced more than 150 times comparing with the required number of families with SAO. Our observations show that, usually, if unaffected children exist in the nuclear families, the power will be lower than when there are less or no unaffected children.

Families with Different Structure

In practice, families of different structures can be recruited. Table IV shows when there are several family structures to be considered in the TDT power computation, our approach is still quite accurate. Our study developed the first approach for computing the power of the TDT in practice with nuclear families of mixed structures.

DISCUSSION

In this paper, we develop a general and highly accurate approach to compute analytically the power of the TDT. We compare the results computed from our approach with those from several other recent papers, both against the results obtained by computer simulations. We show that the results computed with our approach are more accurate than (sometimes significantly) or at least the same as those derived from other approaches. More importantly, our approach can handle various situations, which include (1) families that consist of any number of children with any configuration of affected and nonaffected sibs; (2) families ascertained through the affection status of parent(s); (3) any mixed sample with different types of families in (1) and (2); (4) the marker locus is not a DSL; and (5) existence of allelic heterogeneity. To date, our approach is the most accurate and most general one for the analytical power computation of the TDT. We implement this approach in a user-friendly computer program, *TDT Power Calculator*.

Many alternative approaches to mapping disease genes are available for empirical geneticists. They include the model-based linkage analyses, model-free linkage analyses (the allele sharing approach), variance component analyses, case-control

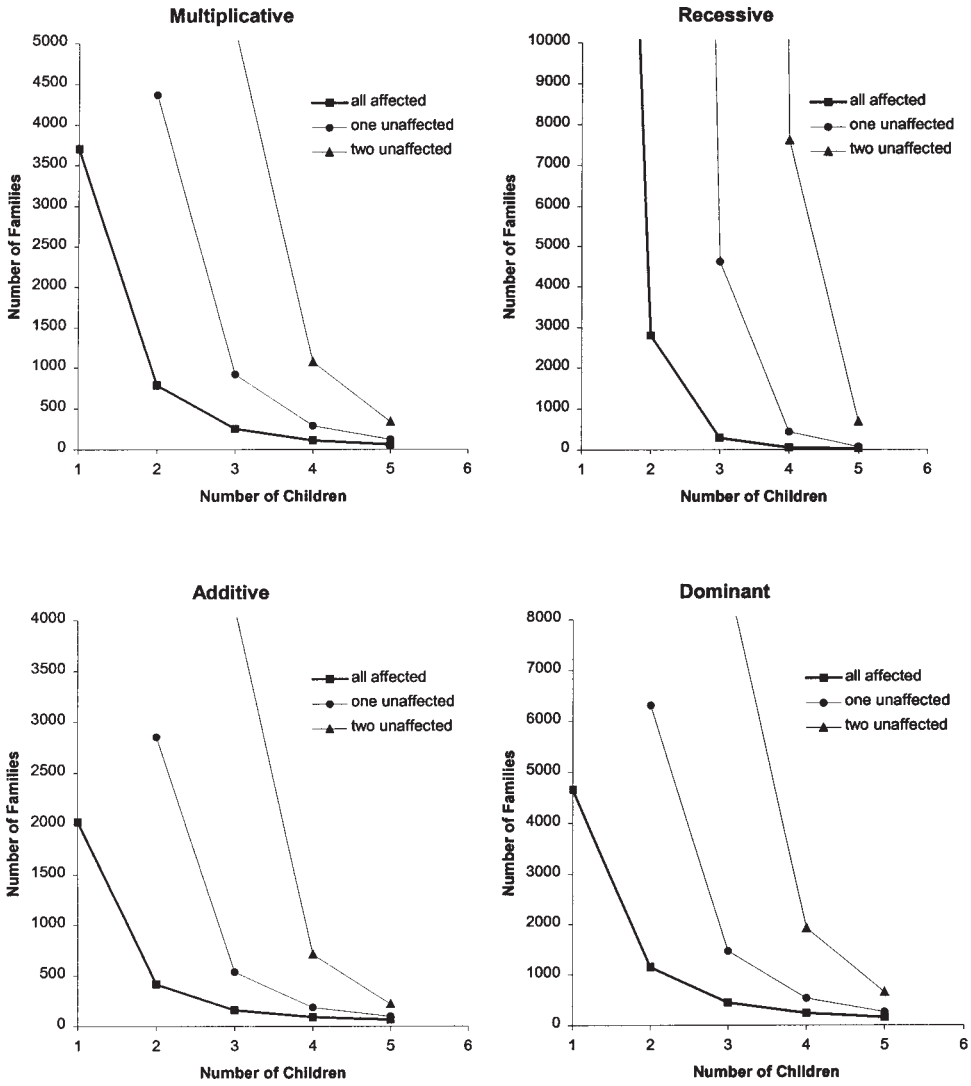

Fig. 1. The effect of the number of children on the power of the TDT. The allele frequency of the marker and the disease susceptibility is 0.4 and 0.05, respectively. No recombination occurs between the marker and the DSL. Linkage disequilibrium is 0.8 of its maximum. The penetrances under various genetic models are, respectively, as following: multiplicative model, 0.625, 0.25, and 0.1; recessive model, 0.9, 0.1, and 0.1; additive model, 0.85, 0.5, and 0.15; and dominant model, 0.5, 0.5, and 0.2. The disease prevalences are 0.12, 0.10, 0.19, and 0.23, respectively. The significance level is 0.001 and the power is 0.8. The total number of children per nuclear family is given on the X-axis. Families with none, one, and two unaffected children are indicated by different dots and lines. The number of affected children in each type of nuclear families can be inferred from the data on X-axis and the number of unaffected children in nuclear families as indicated by different types of dots.

studies, TDT and model-based association studies, etc. [Elston, 1998]. Which method to choose in a specific genetic study is a practical question that empirical geneticists have to face. To answer this kind of practical questions, accurate and general methods for power computation for different DSL searching approaches should be developed and made user-friendly. The approach and the program developed here should be significant for theoreticians to investigate accurately the statistical power of the TDT in various situations, and for empirical geneticists to efficiently plan studies using the TDT.

Usually, for an inaccurate approximation approach, the sample size computed analytically may significantly deviate from true sample size, especially under the following situations: (1) low disease prevalence; (2) low frequency of the disease susceptibility allele(s); (3) high genotypic relative risk; and/or (4) small sample sizes. Tables I–IV show that our approximation is quite robust and accurate even in the above situations. Further investigations at a more standard significance level (larger than $5*10^{-8}$) show that the approximation in our methods works well across the whole distribution of the TDT statistic. For example, in the second situation of the multiplicative model in Table III (where the required sample size is 806), we investigated the significance levels of 0.1, 0.05, 0.005, 0.001, 0.0001, and $5*10^{-8}$, respectively. With sample sizes of 125, 159, 270, 346, 455, and 806, the simulated powers with 100,000 simulations are 0.800, 0.800, 0.800, 0.801, 0.802, and 0.801 respectively, and the analytical powers are 0.801, 0.801, 0.801, 0.800, 0.799, and 0.800, respectively. It shows that, under different significance levels, the analytical powers are very close to the simulated power and hence highly reliable.

TDT Power Calculator has many other potential utilities that can be explored. For example, employing *TDT Power Calculator*, it is easy to find that the TDT picks up additive effects in the penetrances and hence the TDT study under additive model of inheritance (MOI) has greater power than that in other MOIs. This phenomenon is also apparent in Tables II and IV, and Figure 1. With the *TDT Power Calculator*, by examing the simulated type I errors, one can easily verify that (1) the TDT is always a valid test for linkage in the presence of linkage disequilibrium; (2) the TDT is a test for both linkage and linkage disequilibrium if there is only one affected child in each nuclear family; and (3) the TDT is not a valid test for association if multiple affected children are recruited in a family.

ACKNOWLEDGMENTS

This study was supported by grants from NIH, Health Future Foundation and HuNan Normal University, and by a graduate student tuition waiver (to W.-M.C.) from Creighton University. We thank Hai-Yan Wang for her helpful discussions on several statistical issues at the initial stage of this work. We are grateful to Dr. Daniel Schaid and two anonymous reviewers for helpful comments that helped to improve the manuscript.

REFERENCES

Allison DB. 1997. Transmission-disequilibrium tests for quantitative traits. Am J Hum Genet 60:676–90.

Boehnke M, Langefeld CD. 1998. Genetic association mapping based on discordant sib pairs: the discordant-alleles test. Am J Hum Genet 62:950–61.

Camp NJ. 1997. Genomewide transmission/disequilibrium testing: consideration of the genotypic relative risks at disease loci. Am J Hum Genet 61:1424–30.
Camp NJ. 1999. Genomewide transmission/disequilibrium testing: a correction. Am J Hum Genet 64:1485–7.
Curtis D. 1997. Use of siblings as controls in case-control association studies. Ann Hum Genet 61:319–33.
Elston RC. 1998. Linkage and association. Genet Epidemiol 15:565–76.
Huang J, Jiang Y. 1999. Linkage detection adaptive to linkage disequilibrium: the disequilibrium maximum-likelihood-binomial test for affected-sibship data. Am J Hum Genet 65:1741–59.
Knapp M. 1999. A note on power approximations for the transmission/disequilibrium test. Am J Hum Genet 64:1177–85.
Monks SA, Kaplan NL, Weir BS. 1998. A comparative study of sibship tests of linkage and/or association. Am J Hum Genet 63:1507–16.
Müller-Myhsok B, Abel L. 1997. Genetic analysis of complex diseases. Science 275:1328–9.
Rabinowitz D. 1997. A transmission disequilibrium test for quantitative trait loci. Hum Hered 47:342–50.
Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273:1516–7.
Schaid DJ. 1998. Transmission disequilibrium, family controls, and great expectations. Am J Hum Genet 63:935–41.
Schaid DJ, Rowland CR. 1998. The use of parents, sibs, and unrelated controls to detection of associations between genetic markers and disease. Am J Hum Genet 63:1492–1506.
Schaid DJ, Rowland CM. 1999. Quantitative trait transmission disequilibrium test: allowance for missing parents. Genet Epidemiol 17:S307–12.
Slager SL, Huang J, Vieland VJ. 2000. Effect of allelic heterogeneity on the power of the transmission disequilibrium test. Genet Epidemiol 18:143–56.
Spielman RS, Ewens WJ. 1998. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450–8.
Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–16.
Tu IP, Whittemore AS. 1999. Power of association and linkage tests when the disease alleles are unobserved. Am J Hum Genet 64:641–9.
Whittaker JC, Lewis CM. 1998. The effect of family structure on linkage tests using allelic association. Am J Hum Genet 63:889–97.
Xiong MM, Guo SW. 1998. The power of linkage detection by the transmission/disequilibrium tests. Hum Hered 48:295–312.
Xiong MM, Krushkal J, Boerwinkle E. 1998. TDT statistics for mapping quantitative trait loci. Ann Hum Genet 62:431–52.

APPENDIX: IMPROVED ALGORITHM WITH FAMILY CLASSIFICATION

*Class*es are defined by (u_i, v_i). For simple demonstration, we consider a two-allele marker locus. Denote one marker allele as *M* (or *1* in Tables AI–AII), the other alleles as *m* (or *0* in tables). There are a total of three informative parental mating types: *Mm*MM*, *Mm*mm*, and *Mm*Mm*.

First, let's consider a simple situation (Table AI) that the disease status of both parents in each nuclear family is the same, i.e., AA, NN, or XX. Focus on nuclear families with the parental mating type *Mm*MM*. Among *t* children, there are *k* affected with the marker genotype MM (denoted as *11*) and *(a-k)* affected children with Mm genotype (denoted as *10*). In our notation, orders of alleles and individuals are considered so that the first allele of a genotype comes from the first parent. Here is a specific example in Table AII: $1011(11)_k(01)_{a\text{-}k}$, which denotes that the marker genotype(s) of the first parent, the second parent, the first *k* affected children and the last affected *a-k* children are *Mm*, *MM*, *MM* and *mM*, respectively. There are $4C_a^k$ such kind of families. This is because there are C_a^k combinations of children's geno-

Table AI. Classification of Families With the Same Parental Affection Status

G_M	u_i	v_i	Freq	Range
$1011(11)_k(01)_{a-k}$	k	a-k	$4C_a^k$	$0 = k = a$
$1000(10)_k(00)_{a-k}$	k	a-k	$4C_a^k$	$0 = k = a$
$1010(11)_{k1}(10)_{k2}(01)_{k3}(00)_{k4}$	$2k_1+k_2+k_3$	$k_2+k_3+2k_4$	$\frac{4a!}{k_1!\,k_2!\,k_3!\,k_4}$	$0 = k_1, k_2, k_3, k_4 = a$ $k_1+k_2+k_3+k_4 = a$

types and disease status given a specific order of the parents, and there are 4 possible configurations (10*11, 01*11, 11*10, 11*01) of the parents with the genotype Mm*MM.

The probability of the set of marker genotypes in a nuclear family provided the specific family disease status is denoted as $Pr(G_M|\text{family disease status})$. In the above example, the probabilities among the kinds of families are the same. Therefore, the total probability that this type of families contributes to a *class* with (u_i, v_i) being $(k, a\text{-}k)$ is

$$Pr(G_M|\text{family disease status})*4C_a^k \tag{A1}$$

The following is a general approach to compute the probabilities of the set of marker genotypes given specific family disease statuses. We do not consider the marker genotypes of unaffected sibs here because they do not contribute to u_i and v_i.

$$\begin{aligned}
&\Pr(G_M \mid \text{family disease status}) \\
&= \Pr(M_1^1 M_2^1, M_1^2 M_2^2, m_1^1 m_2^1, \ldots m_1^a m_2^a \mid P^1, P^2, C^1, \ldots, C^t) \\
&= \frac{\sum_{D_1^1, D_2^1, D_1^2, d_1^1, d_2^1, \ldots, d_1^t, d_2^t} \Pr(G^1, G^2, g^1, \ldots, g^a, d_1^{a+1} d_2^{a+1}, \ldots, d_1^t d_2^t, P^1, P^2, C^1, \ldots, C^t)}{\Pr(P^1, P^2, C^1, \ldots, C^t)} \\
&= \frac{\sum_{D_1^1, D_2^1, D_2^2} \{\prod_{x=1}^{2} \Pr(G^x) \Pr(P^x \mid G^x) \sum_{d_1^1, d_2^1, \ldots d_1^t d_2^t} \Pr(g^1, \ldots, g^a, d_1^{a+1} d_2^{a+1}, \ldots, d_1^t d_2^t, C^1, \ldots, C^t \mid G^1, G^2)\}}{\Pr(P^1, P^2, C^1, \ldots, C^t)}
\end{aligned}$$

where in the numerator,

$$\begin{aligned}
&\sum_{d_1^1, d_2^1, \ldots, d_1^t, d_2^t} \Pr(g^1, \ldots, g^a, d_1^{a+1} d_2^{a+1}, \ldots, d_1^t d_2^t, C^1, \ldots, C^t \mid G^1, G^2) \\
&= \sum_{d_1^1, d_2^1, \ldots, d_1^t, d_2^t} \{\prod_{s=1}^{a} \Pr(C^s, g^s \mid G^1, G^2) . \prod_{s-a+1}^{t} \Pr(C^s, d_1^s d_2^s \mid G^1, G^2)\} \\
&= \{\prod_{s=1}^{a} \sum_{d_1^s} \sum_{d_2^s} \prod_{k=1}^{2} \Pr(h_k^s \mid G^k) \Pr(C^s \mid d_1^s d_2^s)\} \cdot \{\prod_{s=a+1}^{t} \sum_{d_1^s} \sum_{d_2^s} \prod_{k=1}^{2} \Pr(d_k^s \mid D_1^k D_2^k) \Pr(C^s \mid d_1^s d_2^s)\}
\end{aligned}$$

The denominator can be computed as the summation of the numerators for all possible marker genotypes.

Thus, the probability of a *class* can be obtained. u_i and v_i, the count that M alleles are transmitted or not transmitted from marker heterozygous parents to affected children, are also given in Table AI. Following this improved algorithm with family classification, the computational time is reduced substantially, from an exponential function to a linear function of the number of affected children.

If the two parental affection statuses are different, enumeration can be performed as in Table AII. Minor changes need to be made in Formula (A1) only.

Table AII. Classification of Families With Different Parental Affection Status

G_M	u	v	Freq	Range
$1011(11)_k(01)_{a-k}$	k	a-k	$2C_a^k$	$0 = k = a$
$1110(11)_k(10)_{a-k}$	k	a-k	$2C_a^k$	$0 = k = a$
$1000(10)_k(00)_{a-k}$	k	a-k	$2C_a^k$	$0 = k = a$
$0010(01)_k(00)_{a-k}$	k	a-k	$2C_a^k$	$0 = k = a$
$1010(11)_{k1}(10)_{k2}(01)_{k3}(00)_{k4}$	$2k_1 + k_2 + k_3$	$k_2 + k_3 + 2k_4$	$\frac{4a!}{k_1!\,k_2!\,k_3!\,k_4}$	$0 = k_1, k_2, k_3, k_4 = a$ $k_1 + k_2 + k_3 + k_4 = a$