Genet. Res., Camb. (2004), **83**, *pp.* 133–141. © 2004 Cambridge University Press
DOI: 10.1017/S0016672304006743 Printed in the United Kingdom

Power calculations for the transmission/disequilibrium and affected sib pair tests using elementary probability methods

BARRY W. BROWN*
Department of Biostatistics and Applied Mathematics, Unit 237, University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA

(*Received 21 October 2003 and in revised form 5 January 2004*)

Summary

The transmission/disequilibrium test (TDT) and the affected sib pair test (ASP) both test for the association of a marker allele with some conditions. Here, we present methods for calculating the probability of detecting the association (power) for a study examining a fixed number of families for suitability for the study and for calculating the number of such families to be examined. Both calculations use a genetic model for the association. The model considered posits a bi-allelic marker locus that is linked to a bi-allelic disease locus with a possibly nonzero recombination fraction between the loci. The penetrance of the disease is an increasing function of the number of disease alleles. The TDT tests whether the transmission by a heterozygous parent of a particular allele at a marker locus to an affected offspring occurs with probability greater than 0·5. The ASP tests whether transmission of the same allele to two affected sibs occurs with probability greater than 0·5. In either case, evidence that the probability is greater than 0·5 is evidence for association between the marker and the disease. Study inclusion criteria (IC) can greatly affect the necessary sample size of a TDT or ASP study. IC considered by us include a randomly selected parent at least one parent or both parents required to be heterozygous. It also allows a specified minimum number of affected offspring to be required (TDT only). We use elementary probability calculations rather than complex mathematical manipulations or asymptotic methods (large sample size approximations) to compute power and requisite sample size for a proposed study. The advantages of these methods are simplicity and generality.

1. Introduction

Both the transmission/disequilibrium test (TDT) and the affected sib pair test (ASP) assess the evidence linking a bi-allelic marker at a known locus with a possibly unknown bi-allelic disease locus. Similar tests use a quantitative phenotype as the outcome (see, for example, Zhu & Elston, 2001). We do not consider these tests here because our methods do not apply. We describe methods for proceeding from a genetic model to calculations of the power of a test of fixed size and of the sample size needed to achieve a given power.

We consider the distribution of genotypes of affected offspring for the different genotypes of parents. Basic probability calculations, particularly conditional probability evaluations, are used to transform the genetic parameters of the model into (1) the probability of transmission by a heterozygous parent of a particular allele at a marker locus (TDT) to an affected offspring or the same marker allele to two affected offspring (ASP), and (2) the probability that an examined family is included in the study. These calculations reduce the problem of power or sample size for the test to the same calculation for the one-sample binomial with random sample size.

The primary advantages of our methods are their elementary nature and their generality of application. The methods apply to both the TDT and ASP, they use a general genetic model, and they handle arbitrary study inclusion criteria (IC). However, these methods do not provide analytic formulae for statistical power. Even were it possible to concatenate the equations presented here, the results would be too complex to enhance understanding. Other investigators provide

* Corresponding author. e-mail: bwb@mdanderson.org

formulae and, even when the formulae are derived with simplifying assumptions, they provide insight into the comparative properties of the TDT and ASP; our methods would require numeric assessments.

The one simplifying assumption used in our methods is that all families in a study contribute the same sample size to the study and that this sample size is the expected sample size averaged over family types. The actual contribution depends on the number of parents heterozygous at the marker locus and (for the TDT) the random number of affected offspring. Accounting for this variability would require large sample methods not used in the remainder of this work. We justify our simplification by two observations. The law of large numbers assures us that the mean familial contribution to the sample size tends to the population value, and there are typically many families in TDT and ASP studies. We are also reassured by the agreement of our method (without any correction for variable familial contribution) with Knapp's (1999) power calculated via simulation in a wide variety of cases as described below.

Spielman *et al.* (1993) developed the statistical aspects of the TDT to its current state. Knapp (1999), Tu & Whittemore (1999), and Chen & Deng (2001) use asymptotic methods to derive analytic formulae for power calculations.

Asymptotic methods provide large sample size approximations and are often available when exact methods are not. However, mathematical expertise is required to use and understand them. Additionally, there is usually no estimate of the precision of these approximations. Our only use of asymptotics is a computational convenience: the replacement of the binomial distribution with the approximating normal distribution when averaging power over various sample sizes for large studies. Were the user willing to have a computer program run for minutes instead of a few seconds, this approximation could be eliminated.

McGinnis (1998, 2000) derived analytic formulae for allele transmission probabilities and probability of a heterozygous parent. The difficulty with the analytic method lies in deriving the formulae and, for the user, following the derivation. In some cases, problem simplifications are necessary to obtain closed form mathematical results. Our methods follow those of McGinnis but we resort to computation rather than problem simplification and so arrive at algorithms instead of formulae.

2. Model

(i) *TDT and ASP*

(a) *Background*

A bi-allelic marker locus (A/B) is suspected to be linked to and in possible linkage disequilibrium with a bi-allelic disease locus (D/d) with disease-predisposing allele D and non-predisposing allele d. Parents and affected offspring in each family in the study are genotyped at the A/B locus.

(b) *TDT*

The number of times (0, 1, 2) that alleles A and B are transmitted by A/B heterozygous parents to an affected offspring is counted. The total number of transmissions of A is n_a; similarly, n_b is the total number of transmissions of B; and $n=n_a+n_b$. The statistic measuring linkage of the marker locus with disease is $\hat{p}_t=n_a/n$.

The TDT test can be either one- or two-sided depending on whether allele A has been identified in advance as being associated with the condition studied or whether either allele A or B might be implicated.

(c) *ASP*

One pair of affected sibs and each heterozygous parent is genotyped. The number of times that either the A or the B allele is transmitted by the A/B heterozygous parent to both affected sibs is denoted by n_s; the number of times that different alleles are transmitted is n_u; $n=n_s+n_u$. The statistic measuring linkage is $\hat{p}_s=n_s/n$. The ASP test is inherently two-sided.

(d) *Notation*

For the TDT, p_t is the true (alternative hypothesis) probability of transmission of A to an affected offspring by an A/B heterozygous parent. For the ASP test, p_s is the alternative hypothesis probability of transmission of the same allele to two randomly selected affected offspring. The variables p_s or p_t are generically denoted (depending on the test being considered) by p_a.

(ii) *Study design*

The design of a study can have a large effect on the size and cost of a study. Components of the design include the inclusion criteria (IC) that must be met for a family to enter the study and the sample size. The IC include parental A/B heterozygousity conditions and the minimum number of affected offspring in a family. The sample size is the number of families to be examined for possible inclusion in the study. The sample size required for a particular power of detection (and the number of genotypings to be performed) differ according to the IC. The cost of a study includes the cost of finding possibly eligible families and the cost of genotyping. The cost will also be considerably affected by the IC. The IC considered include the following.

https://doi.org/10.1017/S0016672304006743 Published online by Cambridge University Press

1. *Parental A/B heterozygousity conditions*. One of the following three criteria is used to determine the families eligible for the study.

   (1) Random parent. A randomly chosen parent is genotyped and the family is included in the study if this parent is A/B heterozygous.

   (2) One parent A/B heterozygous. A random parent is genotyped at the marker locus; if the parent is A/B heterozygous the family is included. If not, the other parent is genotyped and the family is included if this other parent is A/B heterozygous.

   (3) Both parents A/B heterozygous. Both parents are genotyped and both must be A/B heterozygous for inclusion.

2. *Minimal number of affected offspring*. For the TDT, the investigator has the option of using only one randomly chosen offspring per family in the study or of using all such. In the latter case, the investigator can specify k, the minimum number of affected offspring required for a family to be eligible for the study; k can be, and frequently is, 1. The use of a value greater than 1 may, in some cases, lower the requisite number of families examined or the number of genotypings to be performed to achieve a given statistical power.

(iii) *List from which families are selected*

Selection of families for the study occurs via random sampling from a list, which may consist of either affected offspring or families with affected offspring. The type of list to be considered in the calculation is the one that best mimics the ascertainment scheme to be used in the study. The list contains only families or members of families with at least k affected offspring.

The list is considered to be large enough that there is little chance of identifying the same family through different randomly chosen members. This condition eliminates the need to consider the complexities of sampling without replacement. With a list that is not much larger than the sample size, the sample size should be considered to be the number of different families to be identified.

3. Methods

Distinguishing indistinguishable cases is standard in probability derivations; its use here avoids the occurrence of powers of 2 whose use might not be obvious.

We consider haplotypes to consist of two ordered pairs of alleles, a first pair and a second pair. The first pair can be considered to be those inherited from the father and the second to be from the mother (even though phase is usually not determinable). By this convention, the haplotype AD/BD differs from BD/AD because the order of the alleles differs: the first pair of alleles are {AD} and {BD}, respectively.

There are four possible first or second pairs of a haplotype, {AD, Ad, BD, Bd}, so there are 16 ordered haplotypes for a parent in the study: {AD/AD, AD/Ad, AD/BD, AD/Bd, Ad/AD, Ad/Ad, Ad/BD, Ad/Bd, BD/AD, BD/Ad, BD/BD, BD/Bd, Bd/AD, Bd/Ad, Bd/BD, Bd/Bd}. An offspring type is one of these ordered haplotypes. A family type is the ordered haplotype of the father followed by the ordered haplotype of the mother (i.e. the ordered genotype of the parents). There are $16 \times 16 = 256$ family types.

Penetrances depend only on the number of D alleles and not on the ordering of the haplotypes within the genotype. Reordered haplotypes are combined in the summations leading to the results.

(i) *Parameters of the genetic model*

- Population frequencies of Bd, BD, Ad, AD. These values can optionally be calculated from the population frequencies of alleles A and D and the coefficient of disequilibrium,

$$\mathcal{F}_{AD} \times \mathcal{F}_{BD} - \mathcal{F}_{Ad} \times \mathcal{F}_{Bd} \tag{1}$$

  where $\mathcal{F}_{AD}$, for example, is the population frequency of the haplotype AD (McGinnis, 1998).
- The recombination fraction between the marker and the disease locus, θ ($0 \leqslant \theta \leqslant 0{\cdot}5$).
- The penetrance of a genotype as a function of the number of D alleles (0, 1, 2). By appropriate choice of penetrance values, an arbitrary mode of inheritance can be specified including additive, dominant, recessive, and others.
- The mean number of offspring per family, λ, including both affected and unaffected offspring. The number of offspring is Poisson distributed with mean λ for all families.

(ii) *Sample size and power calculations*

The sample size, N, is the number of individuals or families on the list that must be examined to assure that the study has a specified power of detecting linkage at some significance level. The sample size is calculated by varying the sample size, N, and examining the resulting power of the study, PowStudy(N). Because PowStudy(N) is monotone increasing with N, finding the value of N that yields the desired power is a solved problem (e.g. a bisection algorithm could be used).

The power of the study is calculated from three quantities: (1) p_I, the probability that a random member of the list leads to a family suitable for inclusion; (2) the expected sample size per family, $\mathcal{N}$; and (3) the alternative hypothesis value, p_a.

https://doi.org/10.1017/S0016672304006743 Published online by Cambridge University Press

The TDT can be either a one-or a two-sided test; the ASP is inherently two-sided. The significance level for a two-sided test can be transformed into an equivalent level for a one-sided test by halving it. Hence, we consider only one-sided tests.

Let PowBin(p_a, N) be the power of a one-sided test of the null hypothesis that the probability of an event is $p=0{\cdot}5$ against the alternative that $p>0{\cdot}5$ for a sample size of N when the true probability of an event is p_a and the significance level is α.

Because the number of families in the study is random, the power of a proposed study with N families examined is:

$$\text{PowStudy}(N)=\sum_{i=0}^{N}\text{Bin}(i,\,N,\,p_I)\text{PowBin}(p_a,\,i\times\mathcal{N}) \tag{2}$$

where Bin(i, N, p_I) is the probability that i families will be included in the study when N members of the list are examined and the probability of study inclusion of each is p_I.

For large N, this summation could require a great number of calculations of PowBin; consequently, when N is large, the binomial distribution is approximated by the normal and a Hermite formula is used to integrate the density of this normal distribution times PowBin. The Hermite formula approximates the integral by the sum of a fixed number of terms, each of which is a weight times PowBin evaluated at a specified quantile of the normal distribution. See Table 25.10 of Abramowitz & Stegun (1964) for details.

(iii) *Calculating probability of family inclusion, the sample size per family and the alternative hypothesis*

These quantities depend on the family type composition of the study, which in turn depends on the composition of the list. The family type composition of the list is dependent in turn on the expected number of affected offspring in each family type.

- Distribution of the number of affected offspring. The mean total number of offspring per family is λ; let the probability that an offspring of family type f is affected be $\text{P}_\text{A}(f)$. Then the number of affected offspring is distributed Poisson with mean $\mu(f)=\lambda\text{P}_\text{A}(f)$.
- Composition of the list. Whether the list is constituted of families or affected offspring, $\mathcal{L}(f)$, the proportion of the list consisting of family type f, is proportional to the population frequency of families of type f and to the probability that the family has at least k affected offspring. Thus, for a list of families, the expected frequency of family type f is

$$\mathcal{L}(f)=\frac{\mathcal{F}_f S(k,\mu(f))}{\sum_g \mathcal{F}_g S(k,\mu(g))} \tag{3}$$

where $S(k,\mu(f))$ is the probability of k or more affected offspring given that the mean number of affected offspring is $\mu(f)$ (i.e. S is the tail of the Poisson distribution).

$\mathcal{F}_f$, the frequency of family type f in the population, is the product of the population frequencies of two haplotypes: that of the father and that of the mother in family type f.

For a list of affected offspring from families with at least k affected offspring, the representation on the list of family type f is also proportional to the expected number of affected offspring of the family type, $\mathcal{N}_k(f)$. This is the mean number of affected offspring given that there are at least k such and is:

$$\mathcal{N}_k(f)=\mu(f)\frac{S(k-1,\mu(f))}{S(k,\mu(f))}. \tag{4}$$

The frequency of family type f in the list of affected offspring is

$$\mathcal{L}(f)=\frac{\mathcal{F}_f S(k,\mu(f))\mathcal{N}_k(f)}{\sum_g \mathcal{F}_g S(k,\mu(g))\mathcal{N}_k(g)} \tag{5}$$

- Frequency of family types in the study. Families on the list meeting the parental heterogeneity conditions will be eligible for the study. The probability that a random family from a list of families is eligible for the study is

$$p_I=\sum_f I_H(f)\mathcal{L}(f) \tag{6}$$

where $I_H(f)$ is 1 if family type f meets the parental heterogeneity requirements of the study and 0 otherwise.

The expected frequency of family type f in the study is

$$S_f=\frac{I_H(f)\mathcal{L}(f)}{\sum_g I_H(g)\mathcal{L}(g)} \tag{7}$$

for a list either of families or of offspring.

$\mathcal{N}(f)$ is the average sample size contributed to the study by a family of type f. For an ASP study or TDT with one affected offspring per family used, $\mathcal{N}(f)$ is the number of A/B heterozygous parents. For the TDT using all affected offspring, $\mathcal{N}(f)$ is the number of heterozygous parents multiplied by $\mathcal{N}_k(g)$. $\mathcal{N}$ is the average (over family types) contribution to the sample size of the study:

$$\mathcal{N}=\sum_f S_f\mathcal{N}(f) \tag{8}$$

Similarly, p_a, the alternative hypothesis probability of specific allele transmission (TDT) or transmission of the same marker allele to affected sibs (ASP), is the weighted average of the same value

https://doi.org/10.1017/S0016672304006743 Published online by Cambridge University Press

over the family types:

$$p_a = \frac{\sum_f S_f \mathcal{N}(f) p_a(f)}{\mathcal{N}}. \quad (9)$$

(iv) *Calculation of the probability that an offspring is affected and the alternative hypothesis probability for each family type*

The discussion in this section is limited to one family type, f, defined by the ordered haplotypes of the father and of the mother.

Denote the haplotype of one parent by wx/yz where w and y are one of {A, B} and x and z are one of {D, d}. As an example, consider AD/Bd. Four pairs of alleles can be transmitted to an offspring with probabilities given below. Without recombination, the allele pairs wx and yz can be transmitted, each with probability $(1-\theta)/2$ (AD and Bd in our example). With recombination, the pairs wz and xy can be transmitted with probability $\theta/2$. (Ad and BD in the example).

Unless the parent is heterozygous at both the marker and disease sites, there will be duplicate unordered allele pairs in the described calculations. These duplicates could be combined and the corresponding probabilities added but this is unnecessary, because the combination is logically performed in summations later.

From the previous results applied to the father and to the mother, we take all combinations of one allele pair from the haplotype of the father with one from the mother. The probabilities of the two allele pairs are multiplied to obtain the probability of an offspring with the specified genotype. Again, there are unordered duplicated accounted for in the summations.

We denote the frequency of offspring haplotype o by Fr(o) o, o_i, and o_j range over all possible offspring types for family type f.

The probability that an offspring of haplotype o is affected is the penetrance, Pen(o). The probability that a random offspring is affected is

$$\mathrm{P_A}(f) = \sum_o \mathrm{Fr}(o)\mathrm{Pen}(o), \quad (10)$$

where Fr(o) is the proportion of offspring type o in the family type and Pen(o) is the penetrance of this offspring type.

(a) *Probability of transmission of marker A (TDT only)*

We know the allele pair transmitted by each parent to o and so we can count the number of transmissions of A by a heterozygous parent, $C_A(o)$, which can take the values 0 to $n_H(f)$, the number of A/B heterozygous parents in family type f. The proportion of transmissions of A to an affected offspring by an A/B heterozygous parent is thus

$$p_t(f) = \frac{\sum_o \mathrm{Fr}(o)\mathrm{Pen}(o)C_A(o)}{\mathrm{P_A}(f)n_H(f)} \quad (11)$$

(b) *Probability of transmission of the same marker allele (ASP only)*

For the ASP, we examine all pairs of offspring types and count the number of transmissions of A or B to both offspring types by a marker heterozygous parent. Let the offspring types be o_i and o_j; the count of transmissions of the same allele is denoted by $C_S(o_i, o_j)$ and can take values 0 to $n_H(f)$. Hence

$$p_s(f) = \frac{\sum_{o_i}\sum_{o_j} \mathrm{Fr}(o_i)\mathrm{Pen}(o_i)\mathrm{Fr}(o_j)\mathrm{Pen}(o_j)C_S(o_i, o_j)}{\mathrm{P_A}(f)^2 n_H(f)}. \quad (12)$$

4. Results

(i) *Examples: TDT*

Our example is the first case in Table 1 of McGinnis (1998); the results are shown in Table 1. There is a considerable decrease in the requisite sample size and in the number of genotypings required if the study admits only families with at least two affected offspring. This might be counter to intuition but it is explained by the fact that, with the genetic parameters used, affected offspring are very common: 48% of the offspring in the population are affected. The table also shows that requiring both parents to be A/B heterozygous is unwise, because it results in 90% of examined families being excluded from the study, which greatly increases the number of families to be examined to achieve the specified power.

Random parent selection requiring two affected offspring is the most efficient IC in terms of the number of genotypes that must be determined; next best is requiring one parent to be A/B heterozygous with at least two affected offspring. In terms of the number of families included in the study, these two cases are again the best but their order is reversed.

(ii) *Examples: ASP*

Table 2 (top) shows results for ASP applied to the same genetic parameters as above. The genetic parameters used makes the ASP test much more demanding in sample size. However, requiring that one parent be A/B heterozygous is the most efficient IC in terms of the number of genotypes to be determined; random parent selection is the optimum in terms of the number of families examined.

https://doi.org/10.1017/S0016672304006743 Published online by Cambridge University Press

Table 1. *Sample size required for 80% power – TDT*

Design[a]	p_t	p_I	N fix	n_H	N offspring included	N families examined	Exp N genotypes
R/1/1	0·567	0·351	342	1·350	1	739	999
R/A/1	0·568	0·351	342	1·35	1·89	382	636
R/A/2	0·568	0·341	338	1·33	2·61	286	541
O/1/1	0·566	0·580	359	1·21	1	514	1326
O/A/1	0·567	0·580	354	1·21	1·89	268	830
O/A/2	0·567	0·567	354	1·20	2·61	199	692
B/1/1	0·571	0·123	317	2·00	1	1305	982
B/A/1	0·573	0·123	315	2·00	1·89	666	1055
B/A/2	0·572	0·115	306	2·00	2·61	507	833

A one-sided test with significance level 0·05 is used. The population frequency of D is 0·60 and of A is 0·75; the disequilibruim coefficient is the maximum possible at 0·015. The recombination fraction between the marker and disease loci is 0; the penetrance is 0·3, 0·45 and 0·6 for 0, 1 and 2 disease alleles, D. The mean number of offspring per family is three. Ascertainment is from a list of affected offspring.

[a] Family inclusion criteria. 'R', 'O', 'B' for random parent, one parent heterozygous, and both parents heterozygous at the marker locus. '1' following the first '/' indicates one affected offspring per family included; 'A' indicates the inclusion of all affected offspring. The final '1' or '2' is the minimum number of affected offspring per family, k. p_t is the probability that a heterozygous parent transmits A to an affected offspring. p_I is the probability that a family examined will meet the parental heterozygousity condition. 'N fix' is the fixed sample size, n, necessary for an 80% power. n_H is the average number of A/B heterozygous parents per family. 'N offspring included' means the average number of affected offspring per family. 'N families examined' means the average number of families to be examined to achieve 0·8 power. 'Exp N genotypes' means the expected number of (A/B) genotypings performed including those that exclude families who do not meet the heterozygousity requirement used.

Table 2. *Sample size required for 80% power – ASP. Sample-size requirements for the ASP test using the same parameter values as Table 1. Precisely two affected offspring are used for the ASP so the second and third entries of the inclusion criterion for TDT do not apply here.* (*Other columns as Table 1.*) *Case 1 uses the same parameter values as the previous table. In Case 2, the penetrances are changed to 0, 0·3 and 0·6 for 0, 1 and 2 D alleles*

Ascertainment	p_s	p_h	n	N H parents	N families examined	N genotypes
Case 1: penetrances 0·3, 0·45 and 0·6 for 0, 1 and 2 D alleles						
R	0·5113	0·3412	12107	1·33	26613	53854
O	0·5109	0·5670	13065	1·20	19153	38307
B	0·5125	0·1154	9921	2·00	43073	67711
Case 2: penetrances 0, 0·3 and 0·6 for 0, 1 and 2 D alleles						
R	0·5882	0·2355	207	1·19	749	1278
O	0·5791	0·4241	256	1·11	552	1572
B	0·6339	0·0468	89	2·00	993	2078

Table 2 (bottom) shows an example that is more favorable to the ASP test. It differs from the previous one only in the penetrances. In this example, random parent selection is most efficient in the number of genotypes to be performed and requiring one parent to be heterozygous is most efficient in the number of families required.

(iii) *Comparison with simulations*

Knapp (1999) presents the results of 5000 simulations of the power of the TDT for a study using a single affected offspring from families with at least one A/B heterozygous parent. The sample sizes of his simulated studies were chosen to produce 80% power for

https://doi.org/10.1017/S0016672304006743 Published online by Cambridge University Press

Table 3. *Power by our method minus power from simulation. Distribution of power calculated by our methods minus that obtained by Knapp (1999) from 5000 simulations, each in 96 sets of parameters*

Method	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	% below 0·01	% below 0·02
Fix n[a]	−0·0361	−0·0024	0·0033	0·0036	0·0135	0·0307	60	87
Var n[b]	−0·0190	−0·0022	0·0020	0·0011	0·0069	0·0262	76	94

[a] Fix n: the distribution of differences where the study sample size is fixed at its expectation.
[b] Var n: the distribution when power is averaged over possible sample sizes.
Successive columns show the minimum, first quartile, median, mean, third quartile and maximum value of the difference. The final two columns show the percentage of cases that are less than 0·01 and 0·02 in absolute value.

a two-sided significance level of $1{\cdot}0 \times 10^{-7}$. The cases simulated include four modes of disease gene inheritance (multiplicative, additive, recessive and dominant), three penetrance ratios $(1{\cdot}5, 2, 4)$, and four values for the population proportion of disease gene $(0{\cdot}01, 0{\cdot}1, 0{\cdot}5, 0{\cdot}8)$; this results in 48 cases. The number of cases is doubled to 96 by simulating sample sizes arising from two different approximations. The distribution of differences between the power from our method and that reported by Knapp is shown in Table 3. Notice that Knapp reports power to only two decimal places – this could increase or decrease the difference between his power and ours by up to 0·005.

The differences between the power from simulation and our methods were small. Allowing for a variable n slightly improves the agreement of simulation with our methods.

5. Discussion: possible extensions to the methods

We have used probability calculations to determine: (1) p_a, the alternative probability hypothesis; (2) p_1, the probability that a random family chosen from the list will meet the parental heterozygousity conditions of the study; and (3) $\mathcal{N}$, the average contribution to the study sample size of an included family.

Power is calculated separately for each possible number of families, N, in the study. (N can range from 0 to the number of families examined.) The computations assume that each family contributes the same sample size, $\mathcal{N}$, and the study power is the average of the powers at each N weighted by the probability of having N families in the study.

We describe two possible extensions to this method; neither is implemented in the available computer program.

(i) *Variable contribution to study sample size*

The actual contribution per family to the study sample size varies systematically by family type owing to the different numbers of marker heterogeneous parents, and randomly within a family type (for TDT with all affected offspring included) owing to the random number of affected offspring. It would be slightly more accurate to average power over all possible study sample sizes taking into account this variability than to average only over the number of families, assuming that all families contribute the same sample size. This would require a calculation of power for each possible sample size instead of only a calculation for each possible number of families.

An exact determination of the distribution of the overall sample size is computationally barely feasible. The distribution for one family can be found by enumerating all cases and averaging over family types weighted by their frequency in the study. The distribution for N families requires evaluating the convolution of this distribution with itself N times. There are efficient methods to shortcut this computation but implementing them would be a major undertaking.

Asymptotics provides a compromise solution. The variance of the contribution of one family can be calculated, and the mean, $\mathcal{N}$, has been calculated. The mean and variance of the study sample size are N times the corresponding figures for one family. Assuming normality of the distribution of sample sizes (the large sample size approximation), one could average power over the normal distribution.

We did not implement either of these possibilities, owing largely to the close correspondence of our results with the simulations of Knapp. Also, accounting for a variable number of families in the study improved the agreement with the results of Knapp only slightly; accounting for the variable contribution of families would, in our opinion, make a lesser change.

(ii) *Poisson distribution of number of affected offspring*

The assumption that the number of offspring in a family is Poisson distributed is strong. The only use of this assumption is to obtain the distribution of the number of affected offspring by family type. An arbitrary distribution could be used. This distribution could be used to model, for example, cases in which

https://doi.org/10.1017/S0016672304006743 Published online by Cambridge University Press

having an affected offspring decreases the probability of future offspring.

6. Computer program

A standard Fortran95 program that performs the calculations described is available as source and as a PC or a Macintosh executable. Inputs to the program are the genetic parameters and inclusion criteria of the study; it calculates either power for a fixed number of families examined or the number of families needed to achieve a specified power. For links to the latest version of the program see the entry tdtasp at http://odin.mdacc.tmc.edu/anonftp/.

7. Glossary of symbols

$\mathcal{F}$. When applied to an allele pair, F_{AD}, the frequency of that pair in the haplotypes of the population. When applied to a family type, $\mathcal{F}_f$, the population frequency of the type – the product of the population frequencies of the four allele pairs of the parents.

$I_H(f)$. 1 if family type f meets the heterogeneity conditions for the study, otherwise is 0.

k. The minimum number of affected offspring required for a family to be included in the study. Also, the minimum number of affected offspring required for a family or its affected offspring to be on the list from which families are randomly chosen to be examined for study suitability. For the TDT with all affected offspring used, k is chosen by the study designer; for the ASP, k is fixed at 2; for TDT with only one affected offspring per family used, k is 1.

λ. The mean number of total offspring per family – assumed to be the same for all families. The distribution of the number of offspring is Poisson.

$\mathcal{L}(f)$. The proportion of the list constituted of family type f or offspring of family type f depending on whether the list contains families or affected offspring.

n_a, n_b, n, p_t, $\hat{p}_t$. (TDT) n_a and n_b are the observed number of transmissions of alleles A and B to an affected offspring by a heterozygous parent. $n=n_a+n_b$ and $\hat{p}_t=n_a/n$. p_t is the true (alternative hypothesis) probability of transmission of A; p_t (f) is the probability of transmission of A by a heterozygous parent for family type f.

n_s, n_u, n, p_s, $\hat{p}_s$. (ASP) n_s is the number of times that the same allele is transmitted to the two affected sibs by a heterozygous parent; n_u is the number of times that different alleles are transmitted. $n=n_s+n_u$ and $\hat{p}_s=n_s/n$. p_s is the true (alternative hypothesis) probability of transmission of the same allele to two affected offspring. p_s (f) is the same probability for family type f.

N. The number of families from the list that must be examined to obtain the desired power.

$\mathcal{N}(f)$, $\mathcal{N}$. The average contribution to the study sample size of a family of type f. $\mathcal{N}$, the average (over family types) contribution to the study sample size of one family.

$n_H(f)$. The number of A/B heterozygous parents in family type f.

$\mathcal{N}_k(f)$. The mean number of affected offspring of a family of type f given that there are at least k affected offspring.

$p_a(f)$. The expected value of p_t or p_s (for TDT and ASP, respectively) for family type f. p_a is the expected value of this quantity averaged over family types.

$\mathrm{P_A}(f)$. The probability that a random offspring of family to type f is affected.

p_I. The probability that a random member of the list leads to a family that will be included in the study.

PowBin(p_a, N). The power of a one-sided one-sample binomial test of the null hypothesis that $p=0{\cdot}5$ against the alternative that $p>0{\cdot}5$. The sample size is N and the true probability is p_a. The significance level of the test is α.

Pen(o). The probability that offspring of type o is affected.

PowStudy(N). The power of a study resulting from the examination of N members of the list and from including in the resulting study all eligible families.

p_a. Generic notation for the alternative hypothesis value; for TDT it is p_t, for ASP, it is p_s.

$S(i;\mu)$. The probability of i or more events in a Poisson distribution with mean μ.

S_f. The proportion of families in the study that are of family type f.

θ. The recombination fraction between the marker and disease locus ($0\leqslant\theta\leqslant 0{\cdot}5$).

This work was suported in part by grants HG-02275, ES-09912, CA-34936 and CA-16672 from the National Cancer Institute, and by the personal generosity of the family of Robert R. Herring. I appreciate the constructive comments of Ralph McGinnis on this manuscript and computer program.

References

Abramowitz, M. & Stegun, I. A. (1964). *Handbook of Mathematical Functions*. Washington, DC, USA. US Government Printing Office.

Chen, W. & Deng, H. (2001). A general and accurate approach for computing the statistical power of the transmission disequilibruim test for complex disease Genes. *Genetic Epidemiology* **21**, 53–67.

Knapp, M. (1999). A note on power approximations for the transmission/disequilibruim test. *American Journal of Human Genetics* **64**, 1177–1185.

McGinnis, R. E. (1998). Hidden linkage: a comparison of the affected sib pair (ASP) test and transmission/disequilibrium test (TDT). *Genetics* **62**, 159–179.

McGinnis, R. (2000). General equations for P_t, P_s, and the power of the TDT and the affected-sib-pair test. *American Journal of Human Genetics* **67**, 1340–1347.

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

Tu, I. P. & Whittemore, A. S. (1999). Power of association and linkage tests when the disease alleles are unobserved. *American Journal of Human Genetics* **64**, 641–649.

Zhu, X. & Elston, R. C. (2001). Transmission/disequilibrium tests for quantitative traits. *Genetic Epidemiology* **20**, 57–74.