MedCrave
Step into the World of Research

Journal of Investigative Genomics

Research Article

Open Access

CrossMark

Calculating exact p-values from the McNamara transmission/ disequilibrium test statistic

Abstract

The transmission/disequilibrium test (TDT) is a popular method for analyzing genetic data in studies of complex disease. It is often assumed that the P-values for the test are well-calculated using the asymptotic, chi-squared distribution. However, that is not always an accurate assumption. A formula is derived for the exact P-value of the TDT McNamara statistic and we show that the asymptotic P-values for the McNemar statistic can often depart considerably from the exact P-values, even when sample sizes are relatively large. Notably, the asymptotic P-values can be either too large or too small, leading to either false positive or false negative results. Since the exact P-value for this statistic is simple to calculate, it will be preferable to do so. We also anticipate that our derivation may find utility in other applications of the McNemar statistic where the underlying variables are binomially-distributed.

Keywords: transmission/disequilibrium test, TDT, mcnemar statistic, exact p-value, disease gene mapping

Volume 2 Issue 4 - 2015

Steven J Schrodi,[1,2] Hywel B Jones[3]
[1]Center for Human Genetics, USA
[2]Computation and Informatics in Biology and Medicine, USA
[3]Department of Genetics, Stanford University, USA

Correspondence: Steven J Schrodi, Center for Human Genetics, 1000 N Oak Ave_MLR, Marshfield, WI, 54449 USA, Tel 715 221 6443, Fax 715 389 4950, Email Schrodi.Steven@mcrf.mfldclin.edu

Received: June 29, 2015 | **Published:** September 02, 2015

Introduction

Genetic association studies have become increasingly common in recent years. These studies aim to detect an increase in the frequency of a disease predisposing variant in a population of affected as compared to a control population. In most cases, the predisposing allele cannot be interrogated directly, and instead a dense set of genetic markers is used as a surrogate. The association study then aims to detect a significant difference in the frequency of one or more alleles at the markers. Such an increase depends on the existence of linkage disequilibrium between a predisposing allele and one or more genetic markers. Because linkage disequilibrium only extends over a short distance, the most commonly used genetic markers are single nucleotide polymorphisms(SNPs) as they are numerous enough to provide a dense coverage of the genome Reich et al.,[1] Reich et al.[2] and easily and inexpensively assayed on high density arrays with well-validated analysis techniques Guo et al.[3]

The simplest experimental design for a genetic association study is to compare a population of cases (patients with the disease being studied) with a population of controls (unaffected individuals). This classic case-control design has been extensively studied in the field of epidemiology with many refinements being incorporated (unequal numbers of cases and controls, different methods for "matching", related cases and unrelated controls, etc) Breslow & Day,[4] Risch & Teng,[5] Teng & Risch,[6] Slager & Schaid.[7] A significant drawback of the case-control design is the potential for confounding which can lead to false positive and false negative results. This arises when an unknown factor causes the populations to differ, even though it may not contribute to the phenotype being examined. In terms of genetics, this may arise when one population is more homogenous than the other. For example, suppose that cases and controls are sampled from different geographic locations exhibiting different genetic histories. Allele frequency discrepancies at a particular marker between cases and controls may be due to the sampling bias rather than disease status. In this simple example, it would likely be relatively easy to tell that the two populations were not well matched (that is, that the background level of relatedness was not equal in the two populations). A simple analysis of markers from across the genome would show that the cases showed greater genetic homogeneity for all the markers Devlin & Roeder,[8] Pritchard et al.,[9] Ardlie et al.[10] However, this confounding (or stratification) can exist in much subtler forms and can lead to spurious results arising from case-control studies. Debate as to the extent of this bias between cases and controls is ongoing and several methods have been developed to either remove genetic background outliers or adjust by principal components derived from large numbers of SNPs Price et al.[11] Additionally, the primary hypothesis tested with case-control designs is independence between disease status and genotype counts. This may have limitations in that truly causative variants generate many genetic patterns in data sets that are not fully interrogated by basic analyses conducted on case-control data: 1 there is well-described decay of statistical association patterns with declining linkage disequilibrium from causal sites Schrodi et al.,[12] Garcia et al.[13] Hardy-Weinberg disequilibrium will exist in affected individuals at the causal site under many disease models Nielsen et al.[14] and Guo et al.[3] causative variants tend to segregate in families with disease status (i.e., linkage) Mohr,[15] Bernstein,[16] Haldane & Smith.[17]

An alternative design is the Transmission/Disequilibrium Test (TDT) Spielman et al.[18] this is a family based method that requires the parents of the affected individual to be available for genotyping. The idea is qualitatively similar to the case-control design except that the population of controls comprises of the non-transmitted alleles from the parents. That is, of the four parental alleles, two are transmitted to the affected child. The other two are not transmitted and hence should be a random sample from the population from which the cases were selected. These two alleles are used as the control genotype. In this way, the case and control population are well matched. Importantly, only heterozygous parents are informative and so the effective sample size may be much smaller than the total number of families in the study. Hence, highly polymorphic markers that tag chromosomes can be a significant advantage when conducting transmission-based tests. Subtly, the test evaluates the simple hypothesis of Mendel's law of segregation for parents to offspring, rather than independence

between disease status and genotypes. Similar to case-control studies and affected sibling pair linkage studies, the TDT aims to combine signal across a large number of small families and as such may lose substantial power under diseases models of high locus heterogeneity. It should be noted that numerous extensions to the TDT have been proposed including those that extend to larger families and multiplex situations Martin et al.[19]

Methods

Data from a TDT association study are analyzed by comparing the transmitted allele to the un transmitted allele. Under the disease model, a causative allele should be more often transmitted to affected offspring than the alternative allele(s) at the site interrogated. Under the disease model, the difference in the frequencies of transmission for each allele is greater than expected under Mendel's law of segregation – the null hypothesis – where each allele would have equal probabilities of being transmitted to the offspring. A McNemartest statistic was originally proposed as the TDT test statistic Spielman et al.[18] assuming a biallelic marker, segregating alleles A_1 and A_2 takes the form of

$$T = \frac{\left(X_1 - X_2\right)^2}{X_1 + X_2}; \tag{1}$$

Where X_1 and X_2 are the number of transmissions of the A_1 and A_2 alleles respectively, for the parents that are heterozygous at the locus evaluated. Researchers tend to use the asymptotic result for calculating p-values from this statistic using the Chi-Squared limiting distribution with one degree of freedom. Let N denote the total number of transmissions from heterozygous parents ($N=X_1+X_2$) then,

$$\lim_{N\to\infty} \frac{1}{dt} P\left[T \in \left(t, t+dt\right)\right] = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{t}{2}\right) \tag{2}$$

The density of X_1 under the null hypothesis of no linkage and no association with disease under Mendel's first law is simply

$$P\left[X_1 = x\right] = \binom{N}{x} 2^{-N} \tag{3}$$

For finite values of N, eqn (2) does not strictly hold and hence using this limiting distribution to determine a p-value is prone to error. For example, the variance of T is

$$Var\left[T\right] = \frac{1}{N^2}\left\{E\left[\left(2X_1 - N\right)^4\right] - N^2\right\} = \frac{2\left(N-1\right)}{N}, \tag{4}$$

As opposed to 2 under the limiting distribution. This departure is non-negligible for small values of N

The exact density of T can be derived, and we use this to calculate the appropriate p-value and examine the rate of convergence to the p-value calculated under the limiting distribution. As the McNemar statistic is commonly used in numerous scenarios within genetics and other fields, there may be additional applications for the exact density of T.

$$P\left[T = t\right] = P\left\{\left[X_1 = \frac{N+\sqrt{Nt}}{2}\right] U \left[X_1 = \frac{N-\sqrt{Nt}}{2}\right]\right\} \tag{5}$$

Since these are disjoint events,

$$= P\left[X_1 = \frac{N+\sqrt{Nt}}{2}\right] + P\left[X_1 = \frac{N-\sqrt{Nt}}{2}\right] \tag{6}$$

Employing eqn (3),

$$= \frac{N!}{2^N \left(\frac{N+\sqrt{Nt}}{2}\right)! \left(\frac{N-\sqrt{Nt}}{2}\right)!} \tag{7}$$

So, for an observed value for the statistic T=t_{obs}, a p-value can be directly calculated analytically with

$$P\left[T \geq t_{obs}\right] = \frac{N!}{2^N} \Sigma_{u \geq t_{obs}} \left[\left(\frac{N+\sqrt{Nu}}{2}\right)! \left(\frac{N-\sqrt{Nu}}{2}\right)!\right]^{-1} \tag{8}$$

Results

To exemplify the use of eqn (8), suppose that one observed 60 transmissions of the A_1 allele from a total of 100 informative transmissions. The T statistic will take a value of 4. Using the limiting Chi-Squared distribution with one degree of freedom as the null distribution, the p-value would be calculated as 0.0455, whereas, using eqn (8) yields an exact p-value for the McNemar statistic of 0.0569. Thus, in this example, the asymptotic approach exaggerates the significance of these data. Further, the departure of the p-value calculated using the Chi-Squared distribution may be positive or negative. That is, the asymptotic test may be either anti-conservative of conservative depending on the parameter space. For example, for a highly significant example where A_1 is 90 transmissions from a total of 100 informative transmissions, the asymptotic p-value is1.24×10^{-15}, while the exact result is 3.06×10^{-17}.

Figure 1 shows the ratio of the exact to asymptotic p-values varying the numbers of transmitted alleles assuming a total number of 100 informative transmissions. When the number of transmissions is close to the null expectation of 50, the two p-values are very similar and therefore the ratio is close to unity. As the number of transmissions increases the p-value given by the Chi-Squared approximation is less than the exact p-value, giving appositive ratio. In this case, the approximation over-estimates the significance, potentially leading to false positive results. Note the region of the parameter space where the proportion of transmission is only slightly greater than the null may be the most realistic scenario for a study (e.g., transmission of predisposing allele of ~60% compared to the null of 50%). Thus, for realistic values or transmission, the asymptotic result can lead to false positive results where association is deemed to exist when it does not. For higher rlevels of transmission (>75%), the situation is reversed with the asymptotic p-value being greater than the exact p-value, underestimating the true significance of the data, leading to false negative results.

Table 1 presents asymptotic and exact p-values for a variety of different sample sizes and transmission frequencies. Here again, the asymptotic p-value can be either greater than or less than the true value. Simulation studies were also carried out to verify these results. Table 2 shows the p-value from one million simulations and the

Citation: Schrodi SJ, Jones H. Calculating exact p-values from the McNamara transmission/ disequilibrium test statistic. *J Investig Genomics*. 2015;2(4):85–88. DOI: 10.15406/jig.2015.02.00032

corresponding Chi-Squared probability. Again, using the asymptotic p-value can lead to substantial errors that may be conservative or anti-conservative. To calculate statistical power or carry out Bayesian derivations, the probability that the *T* statistic takes a given value under the alternative hypothesis is needed. That is, a formula analogous to eqn (7) for probabilities of transmission that deviate from one half. For a general transmission probability q , this is given as,

$$P\left[T = t_{obs}\right]_{disease\,\mathrm{mod}el} = \left[q\left(1-q\right)\right]^{\frac{1}{2}\left(N-\sqrt{Nt}\right)}\left[\left(1-q\right)^{\sqrt{Nt}}\left(\frac{N-{}^{N}\sqrt{Nt}}{2}\right)+q^{\sqrt{Nt}}\left(\frac{N+{}^{N}\sqrt{Nt}}{2}\right)\right] \quad (9)$$

Table 1 presents asymptotic and exact p-values for a variety of different sample sizes and transmission frequencies

			Number of informative transmissions			
			20	**60**	**100**	**200**
Proportion of transmissions to affected off spring	55%	Exact	0.825	0.529	0.368	0.09
		Asymptotic	0.655	0.439	0.317	0.157
		Ratio	1.26	1.18	1.16	0.57
	65%	Exact	0.263	0.027	0.003	1.3x10-5
		Asymptotic	0.189	0.02	0.004	2.2x10-5
		Ratio	1.46	1.35	0.75	0.6
	75%	Exact	0.041	0.041	5.6x10-7	4.2x10-13
		Asymptotic	0.025	0.025	5.7x10-7	1.5x10-12
		Ratio	1.64	1.64	0.98	0.27

Table 2 Shows the p-value from one million simulations and the corresponding chi-squared probability

Transmission	Replicates	Quintile	T Quintile	Chi-squared probability	%Error
20	1000000	0.95	3.2	0.0736	47.3
20	1000000	0.99	7.2	0.00729	-27.1
20	1000000	0.999	9.8	0.00175	74.5
20	1000000	0.9999	12.8	0.000347	246.6
40	1000000	0.95	3.6	0.0578	15.6
40	1000000	0.99	6.4	0.0114	14.6
40	1000000	0.999	10	0.00157	56.5
40	1000000	0.9999	14.4	0.000148	47.8
100	1000000	0.095	4	0.0455	-9
100	1000000	0.999	6.8	0.00932	-6.8
100	1000000	0.9999	10.2	0.00137	37.4

Citation: Schrodi SJ, Jones H. Calculating exact p-values from the McNamara transmission/ disequilibrium test statistic. *J Investig Genomics*. 2015;2(4):85–88. DOI: 10.15406/jig.2015.02.00032

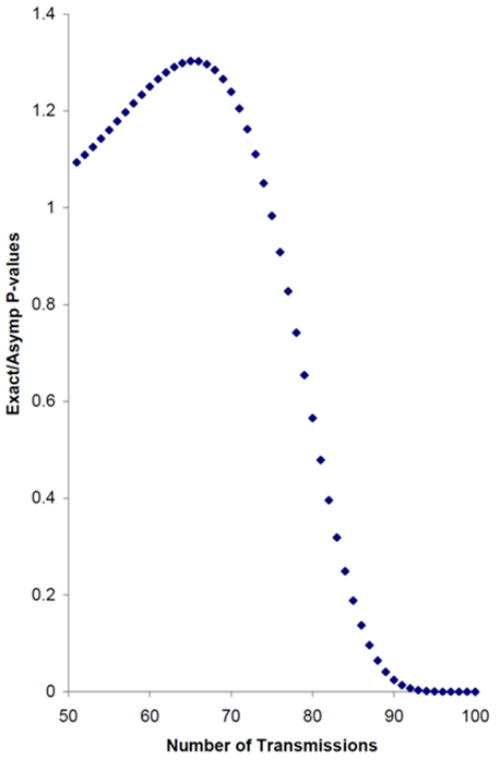

Figure 1 Shows the ratio of the exact to asymptotic p-values varying the numbers of transmitted alleles assuming a total number of 100 informative transmissions.

Conclusion

The TDT is a commonly-used method of carrying out disease mapping studies. Because it requires parental DNA to be available and that the parents are heterozygous for the marker being interrogated, samples sizes will often be modest. Thesis especially true if bi-allelic SNPs are being used. Results in Figure 1 & Table 1 show that the standard method of calculating a p-value by appealing to the asymptotic distribution can lead to both false positive and false negative results. Given the time and cost of genetic studies, such errors can be problematic. False positives results may lead are searcher to continue to pursue a region of the genome that does not harbor a predisposing allele. Conversely, false negatives may result in regions of the genome being excluded, even though they contain genetic factors that play a role in the disease of interest. Most notably, in the example given here, the asymptotic test is anti-conservative in the region of the parameter space most likely to be observed in a genetic association study. This leads to the dangerous situation where evidence for disease is believed to be proven at a given significance level when, in fact, it is not. Given the legion of problems that can arise from false positive and false negative results, it will be important to correctly calculate the probability of the observed data under the null hypothesis, especially when the sample size is limited.

Acknowledgements

None.

Conflict of interest

Author declares that there is no conflict of interest.

References

1. Reich DE, Cargill M, Bolk S, et al. Linkage disequilibrium in the human genome. *Nature*. 2001;411(6834):199–204.
2. Reich DE, Schaffner SF, Daly MJ, et al. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet*. 2002;32(1):135–142.
3. Guo Y, He J, Zhao S, et al. Illumina human exome genotyping array clustering and quality control. *Nat Protoc.* 2014;9(11):2643–2662.
4. Breslow NE, Day NE. Statistical methods in cancer research. Volume I - The analysis of case-control studies. *IARC Sci Publ*. 1980;(32):5–338.
5. Risch N, Teng J. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res*. 1998;8(12):1273–1288.
6. Teng J, Risch N. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases II. Individual genotyping. *Genome Res*. 1999;9(3):234–241.
7. Slager SL, Schaid DJ. Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. *Hum Hered.* 2001;52(3):149–153.
8. Devlin B, Roeder K, Bacanu SA. Unbiased methods for population-based association studies. *Genet Epidemiol.* 2001;21(4):273–284.
9. Pritchard JK, Stephens M, Rosenberg NA, et al. Association mapping in structured populations. *Am J Hum Genet*. 2000;67(1):170–181.
10. Ardlie KG, Lunetta KL, Seielstad M. "Testing for population subdivision and association in four case-control populations." *Am J Hum Genet.* 2000;71(2):304–311.
11. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38(8):904–909.
12. Schrodi SJ, Garcia VE, Rowland CM. A fine mapping theorem to refine results from association genetic studies. *ASHG Abstract.* 2009.
13. Garcia VE, Chang M, Brandon R, et al. Detailed genetic characterization of the interleukin-23 receptor in psoriasis. *Genes Immun*. 2008;9(6):546–555.
14. Nielsen DM, Ehm MG, Weir BS. Detecting marker disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. *Am J Hum Genet*. 1999;63(5):1531–1540.
15. Mohr J. A search for linkage between the Lutheran blood group and other hereditary characters. *Acta Pathol Microbiol Scand.* 1951;28(2):207–210.
16. Bernstein F. Zurgrundlegung der chromosomentheorie der vererbungbein menschen. *ZAbst Vererb.* 1931;57(1):113–138.
17. Haldane JBS, Smith CAB. A new estimate of the linkage between the genes for haemophilia and colour-blindness in man. *Ann Eugen*. 1947;14(1):10–31.
18. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet*. 1993;52(3):506–516.
19. Martin ER, Monks SA, Warren LL, et al. A test for linkage and association in general pedigrees: The pedigree disequilibrium test. *Am J Hum Genet.* 2000;67(1):146–154.

Citation: Schrodi SJ, Jones H. Calculating exact p-values from the McNamara transmission/ disequilibrium test statistic. *J Investig Genomics*. 2015;2(4):85–88. DOI: 10.15406/jig.2015.02.00032