When a Case Is Not a Case: Effects of Phenotype Misclassification on Power and Sample Size Requirements for the Transmission Disequilibrium Test with Affected Child Trios

Steven Buyske, Guang Yang, Tara C. Matise, Derek Gordon
Departments of Genetics and Statistics and Biostatistics, Rutgers University, Piscataway, N.J., USA

Abstract
Phenotype misclassification in genetic studies can decrease the power to detect association between a disease locus and a marker locus. To date, studies of misclassification have focused primarily on case-control designs. The purpose of this work is to quantify the effects of phenotype misclassification on the transmission disequilibrium test (TDT) applied to affected child trios, where both parents are genotyped. We compute the non-centrality parameter of the distribution corresponding to the TDT statistic when there is linkage and association of a marker locus with a disease locus and there is phenotype misclassification. We verify our analytic calculations with simulations and provide an example sample size calculation. In our simulation studies, the maximum absolute difference between statistical power computed by simulation and analytic methods is 0.018. In an example sample size calculation, we observe that to maintain equivalent power, the required sample size increases when the disease prevalence decreases or when the misclassification rate increases. A 39-fold sample size increase is required when the misclassification rate is 5% and the disease prevalence is 1%. Given the potentially substantial power loss for the TDT in the presence of misclassification, we recommend that researchers incorporate phenotype misclassification into their study design for genetic association using trio data. We have developed freely available software that computes power loss for a fixed sample size or sample size for a fixed power in the presence of phenotype misclassification.

Key Words
Genome-wide association • Family-based association • Diagnostic error • Diagnosis error • Phenotyping error

Introduction

Genome-wide association studies are now successfully mapping susceptibility genes for complex trait diseases. These studies have featured large sample sizes or susceptibility genes with a large effect [1–4].

As additional association studies are done, efforts will be made to map genes of smaller effect. One of the critical issues that will come into play with such studies is phenotype and genotype accuracy. Inaccuracy in either of these measures can substantially decrease the power, or equivalently, increase the sample size requirements to map the genes [5–11]. The effects of misclassification are disproportionate in terms of power loss. Studies to map genes with smaller effect will lose more power than studies to map genes with larger effects [6, 12].

To date there has been a large focus on genotype misclassification error. For example, as of June 2008, there are over 100 publications listed in Pubmed using the phrase ‘genotyping error’. Pompanon et al. [13] and Gordon and Finch [6, 14] provide literature reviews.
For the same error rates, phenotype error has a much larger effect on power to detect association [10, 11] than does genotype error. Yet there have been relatively fewer publications in this area; as of June 2008, there are 30 publications listed in PubMed using the phrase “diagnostic error” OR “diagnosis error” OR “phenotype error” AND “genetic”. Furthermore, these publications have focused largely on case-control designs. Methodological studies to quantify power loss for case-control designs have been performed, using the linear trend test or the genotypic test of association [10, 11]. Power in the presence of phenotype misclassification has been implemented in the Power for Association With Error (PAWE-3D) webtool [15].

Table 1. Glossary of notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p, q = 1 - p$</td>
<td>Frequency of $A$ (disease) and $a$ (wild type) alleles, respectively, at disease locus</td>
</tr>
<tr>
<td>$f, f' = 1 - f$</td>
<td>Frequency of $M$ and $m$ alleles, respectively, at marker locus</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Linkage disequilibrium (LD) coefficient between the marker locus and the disease locus</td>
</tr>
<tr>
<td>$\delta' = \begin{cases} \frac{\delta}{\min[pf', qf']} &amp; \delta \geq 0 \ \frac{\delta}{\min[pf, qf']} &amp; \delta &lt; 0 \end{cases}$</td>
<td>Standardized linkage disequilibrium (LD) coefficient between the marker locus and the disease locus [25]</td>
</tr>
<tr>
<td>$\theta$</td>
<td>Recombination fraction between the marker locus and the disease locus. Unless otherwise indicated, this value is set to 0.</td>
</tr>
<tr>
<td>$P_{AM} = \delta + pf, P_{aM} = f - P_{AM}, P_{Am} = p - P_{AM}, P_{am} = 1 - f - p + P_{AM}$</td>
<td>Frequency of haplotypes containing alleles $A$ and $M$, $a$ and $M$, $A$ and $m$, and $a$ and $m$, respectively</td>
</tr>
<tr>
<td>$\phi_{aa} = \Pr(affected \mid having aa genotype), \phi_{Aa}, \phi_{AA}$</td>
<td>Penetrance of $aa$ genotype, $Aa$ genotype, and $AA$ genotype, respectively</td>
</tr>
<tr>
<td>$\gamma_i = \frac{\phi_{Aa}}{\phi_{aa}}$</td>
<td>Heterozygote genotype relative risk [26]</td>
</tr>
<tr>
<td>$\gamma_i = \frac{\phi_{AA}}{\phi_{aa}}$</td>
<td>Homozygote genotype relative risk [26]</td>
</tr>
<tr>
<td>$\phi$</td>
<td>True prevalence of disease in the population under study</td>
</tr>
<tr>
<td>$T_{obs}, T_{aff}, T_{unaff}$</td>
<td>The number of transmissions of the marker $M$ allele from heterozygous parents to observed affected (who may or may not be truly affected), truly affected, and truly unaffected offspring, respectively</td>
</tr>
<tr>
<td>$NT_{obs}, NT_{aff}, NT_{unaff}$</td>
<td>The number of transmissions of the marker $m$ allele from heterozygous parent to observed affected (who may or may not be truly affected), truly affected, and truly unaffected offspring, respectively</td>
</tr>
<tr>
<td>$D_{obs}, D_{aff}, D_{unaff}$</td>
<td>Event that child is observed to be affected, truly affected, and truly unaffected, respectively</td>
</tr>
<tr>
<td>$N, N_{aff}, N_{unaff} = N - N_{aff}$</td>
<td>Number of trios with an observed affected child, truly affected child, and truly unaffected child, respectively</td>
</tr>
<tr>
<td>$\pi = \Pr(D_{obs} \mid D_{unaff})$</td>
<td>Probability of misclassifying a truly unaffected child as an affected child. In this work, we only consider situations in which a truly unaffected child is misclassified as an observed affected child.</td>
</tr>
<tr>
<td>$\phi_{obs} = \phi + (1 - \phi) \pi$</td>
<td>Observed prevalence of disease in the presence of misclassification</td>
</tr>
<tr>
<td>$E(.)$</td>
<td>Expectation operator on random variables</td>
</tr>
<tr>
<td>$C = p \phi_{AA} + (q - p)\phi_{Aa} - q \phi_{aa}$</td>
<td>Expression defined by Deng and Chen [24]; used to simplify some of our later formulas</td>
</tr>
</tbody>
</table>
Table 2. Settings for simulation parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \phi )</td>
<td>0.05</td>
<td>0.15</td>
</tr>
<tr>
<td>( p )</td>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>( \delta' )</td>
<td>0.5</td>
<td>1.0</td>
</tr>
<tr>
<td>( \gamma_2 )</td>
<td>2.25</td>
<td>6.25</td>
</tr>
<tr>
<td>( \pi )</td>
<td>0.01</td>
<td>0.05</td>
</tr>
</tbody>
</table>

For our simulations, we specify that the recombination fraction \( \theta \) between the disease locus and SNP marker locus is 0. Also, we specify that \( f = p \); that is, frequencies of the disease allele and the SNP marker allele in coupling with the disease allele are equal. These simulation parameters are used under three different modes of inheritance: Dominant (\( \gamma_1 = \gamma_2 \)); Recessive (\( \gamma_1 = 1 \)); and Multiplicative (\( \gamma_1 = \sqrt{2} \)), where \( \gamma_1 = \phi_{Aa}/\phi_{aa} \). We have a total of \( 2^5 \times 3 = 96 \) parameter settings, each of which corresponds to a simulation of 10,000 replicates of trio genotype data.

We are aware of only one paper that has looked at the issue of phenotype misclassification in family-based studies of association, authored by Scott and Rogus [16]. These authors studied whether typing control trios increases power for the transmission disequilibrium test (TDT) [17] under certain genetic model conditions. As an additional analysis, they considered sample size requirements of the TDT statistic applied to unaffected child trios in the presence of misclassification. In the work of Scott and Rogus, misclassification means that a true affected child trio is misclassified as an observed unaffected child trio. To our knowledge, there has been no methodological work on the issue of power loss for a fixed sample size or sample size increase for a fixed power for the TDT statistic when true unaffected child trios are misclassified as observed affected child trios. Such misclassification can occur with childhood disease phenotypes such as atypical autism [18], glycogen storage diseases [19], asthma [20], and certain types of cancer [21].

The purpose of our work is to provide a general formula for the non-centrality parameter of the distribution corresponding to the TDT statistic when there is linkage and association of the marker locus with the disease locus and there is phenotype misclassification. Determination of this non-centrality parameter enables us to perform power and sample size calculations for the TDT in the presence of phenotype misclassification. We verify our analytic calculations with simulations and provide an example sample size calculation.

### Methods

#### Test Statistic

To be consistent with the work performed by Scott and Rogus [16], we consider the TDT statistic in this work. The TDT is one of the most commonly used statistics for susceptibility gene mapping in pedigrees [17, 22]. The statistic has the advantage that it is robust to population stratification and it can be used as a test of linkage in the presence of association for multiplex families [23]. In this work, our sampling frame is a trio, consisting of one affected child and their parents. We specify that all members of the trio are genotyped, and we consider biallelic disease and marker loci. We use the same notation as used by Deng and Chen [24] as shown in Table 1.

#### Non-Centrality Parameter of TDT in Presence of Misclassification

For the TDT statistic, Deng and Chen [24] documented that the non-centrality parameter (NCP) without misclassification is given by the formula

\[
\lambda_{\text{TDT}, \text{D}} = \frac{[E(T_{\text{aff}}) - E(NT_{\text{aff}})]^2}{[E(T_{\text{aff}}) + E(NT_{\text{aff}})]},
\]

where

\[
E(T_{\text{aff}}) = 2N_{\text{aff}} [f f' + \delta(f' - \theta) C/\phi],
\]

and

\[
E(NT_{\text{aff}}) = 2N_{\text{aff}} [f f' + \delta(\theta - f) C/\phi],
\]

To quantify power in the presence of misclassification, we compute the NCP

\[
\lambda_{\text{TDT}, \text{Misclass}} = \frac{[E(T_{\text{obs}}) - E(NT_{\text{obs}})]^2}{[E(T_{\text{obs}}) + E(NT_{\text{obs}})]}.
\]

The derivation reduces to computing the quantities \( E(T_{\text{obs}}) \) and \( E(NT_{\text{obs}}) \). Once we have the NCP, we can calculate power for a fixed sample size at a given significance level, or minimum sample size requirements for a fixed power and significance level. We refer to the power calculated by determination of the NCP as analytic power.

#### Simulation Study to Verify Analytic Results

As our calculation of analytic power is asymptotic and implicitly assumes independence of parental genotypes, we perform simulation studies to confirm the accuracy of our analytic calculations. We use a factorial design [27] to determine settings for each of the parameters in one set of our simulations. The parameters we consider are: prevalence (\( \phi \)), disease allele frequency (\( p \)), standardized linkage disequilibrium coefficient (\( \delta' \)), disease mode of inheritance (MOI), and homozygote genotype relative risk \( \gamma_2 \). Table 2 presents the low and high settings of each of these parameters. For MOI, there are three settings: Dominant (\( \gamma_1 = \gamma_2 \)); Recessive (\( \gamma_1 = 1 \)); and Multiplicative (\( \gamma_1 = \sqrt{2} \)), where \( \gamma_1 \) is the heterozygote genotype relative risk. We have a total of \( 2^3 \times 3 = 48 \) simulation parameter settings.

In our simulations, the sample size is fixed at 1000 observed trios per replicate, the recombination fraction \( \delta \) is fixed at 0, and the frequencies of the disease allele and the SNP marker allele in coupling with the disease allele are constrained to be equal (\( p = f \)). Genotypes for each trio are simulated using the method implemented in the FASTSLINK software [28]. LD between the disease
and marker locus is generated by specifying haplotype frequencies instead of disease and marker allele frequencies [29]. Power for various significance levels (0.1, 0.05, 0.01, and 0.001) is calculated as the proportion of replicates out of 10,000 whose TDT statistic exceeds the threshold corresponding to the various significance levels. Under the null hypothesis of no linkage ($\theta = 0.5$) or no association ($\theta = 1$), the TDT is asymptotically distributed as a central $\chi^2$ with 1 degree of freedom [23]. We refer to power determined by simulations methods as simulation power.

**Results**

**Non-Centrality Parameter of TDT in Presence of Misclassification**

We perform an example sample size calculation to document the effects of misclassification on the minimum number of trios (MNT) necessary to achieve 80% power with the TDT at the 0.00001 significance level. The parameters of our disease model are: $\phi = 0.01, 0.025, 0.05, 0.075,$ or 0.10; $p = f = 0.50$; $\delta = 1.0$; Multiplicative MOI with $\gamma_2 = 2.5$; and $\pi = 0.01, 0.02, 0.03, 0.04,$ and 0.05.

**Example Sample Size Calculation in the Presence of Phenotype Misclassification**

We perform an example sample size calculation to document the effects of misclassification on the minimum number of trios (MNT) necessary to achieve 80% power with the TDT at the 0.00001 significance level. The parameters of our disease model are: $\phi = 0.01, 0.025, 0.05, 0.075,$ or 0.10; $p = f = 0.50$; $\delta = 1.0$; Multiplicative MOI with $\gamma_2 = 2.5$; and $\pi = 0.01, 0.02, 0.03, 0.04,$ and 0.05.

**Comparison of Analytic and Simulation Power**

In Figures 1 and 2, we present scatter plots of analytic power versus simulation power for the 48 simulation parameter settings for misclassification rates of 1 and 5%, respectively, and 4 significance levels per set of simulation parameter settings, a total of 192 values in each figure. Each plot closely follows the line ‘$y = x$’. If we consider the absolute difference [simulation power – analytic power] for each set of simulations parameters and significance levels, for the 1% misclassification rate the minimum observed difference is 0, the median difference is 0.001, and the maximum difference is 0.018 (fig. 1), while for the 5% misclassification rate the minimum observed difference is 0, the median difference is 0.003, and the maximum difference is 0.018.
is 0.016 (fig. 2). These results suggest that our analytic calculations are adequate to determine power in the presence of misclassification error when designing a study.

Example Sample Size Calculation in the Presence of Phenotype Misclassification

In figure 2, we present a contour plot of the MNT as a function of disease prevalence and misclassification rates. Each solid contour line represents an integer multiple of 2000 MNT. For example, the area to the right of the dashed contour line contains MNT less than or equal to 1000. The area between the first and second solid contour lines (counting from the right side) contains MNT between 2000 and 4000. For misclassification rate 0 the MNT is 691.

The results for our example sample size calculation are consistent with results for the \( \chi^2 \) test of independence for case-control genotype data [10], and the linear trend test for case-control genotype data [11]. Specifically, power is reduced, or equivalently, minimum sample size is increased, in the presence of misclassification as compared with correctly classified data. Scott and Rogus documented sample size increase in the presence of misclassification for the TDT applied to unaffected child trios [16].

Another of our findings is that power loss increases as the disease prevalence decreases or the misclassification rates increase. These results are consistent with results for the \( \chi^2 \) test of independence for case-control genotype data [10], and the linear trend test for case-control genotype data [11]. Also, Scott and Rogus documented that required sample size increases as the misclassification rate increases for the TDT applied to unaffected child trios [16].

Overall, our findings suggest that phenotype misclassification rate is an important parameter to consider when designing studies of affected child trios that use the TDT statistic for analysis, especially when the prevalence of the disease in the population being studied is smaller. Greenberg [31] provides guidelines on how to improve phenotype quality for psychiatric disorders, with a focus on epi-
lepsy as an example. In this paper we discuss misclassification in terms of \( \pi \), the probability of misclassifying an unaffected child as affected. Investigators may also be interested in the proportion of unaffected among the observed affected, which is given by \((1 - \phi) \pi / (\phi + (1 - \phi) \pi)\).

We have software, written in C, that computes power loss for a fixed sample size or sample size increase for a fixed power for the TDT statistic in the presence of phenotype misclassification. This software may be obtained by contacting the corresponding author.

References


22. Purcell S, Cherny SS, Sham PC: Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics 2003;19:149–150.

23. Greenberg DA: There is more than one way to test for transmission distortion. Genet Epidemiol 2004;26:132–141.


32. Greenberg DA: There is more than one way to collect data for linkage analysis. What a study of epilepsy can tell us about linkage strategy for psychiatric disease. Arch Gen Psychiatry 1992;49:745–750.