# A Log-Linear Approach to Case-Parent–Triad Data: Assessing Effects of Disease Genes That Act Either Directly or through Maternal Effects and That May Be Subject to Parental Imprinting

C. R. Weinberg,[1] A. J. Wilcox,[2] and R. T. Lie[3]

[1]Biostatistics Branch and [2]Epidemiology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC; and [3]Section for Medical Statistics, University of Bergen, Bergen

## Summary

**We describe a log-linear method for analysis of case-parent–triad data, based on maximum likelihood with stratification on parental mating type. The method leads to estimates of association parameters, such as relative risks, for a single allele, and also to likelihood ratio $\chi^2$ tests (LRTs) of linkage disequilibrium. Hardy-Weinberg equilibrium need not be assumed. Our simulations suggest that the LRT has power similar to that of the $\chi^2$ "score" test proposed by Schaid and Sommer and that both can outperform the transmission/disequilibrium test (TDT), although the TDT can perform better under an additive model of inheritance. Because a restricted version of the LRT is asymptotically equivalent to the TDT, the proposed test can be regarded as a generalization of the TDT. The method that we describe generalizes easily to accommodate maternal effects on risk and, in fact, produces powerful and orthogonal tests of the contribution of fetal versus maternal genetic factors. We further generalize the model to allow for effects of parental imprinting. Imprinting effects can be fitted by a simple, iterative procedure that relies on the expectation-maximization algorithm and that uses standard statistical software for the maximization steps. Simulations reveal that LRT tests for detection of imprinting have very good operating characteristics. When a single allele is under study, the proposed method can yield powerful tests for detection of linkage disequilibrium and is applicable to a broader array of causal scenarios than is the TDT.**

Received November 24, 1997; accepted for publication February 13, 1998; electronically published March 27, 1998.

Address for correspondence and reprints: Dr. Clarice R. Weinberg, Biostatistics Branch, MD A3-03, National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, NC 27709. E-mail: weinberg@niehs.nih.gov


0002-9297/98/6204-0032$02.00

## Introduction

The genotypes of triads consisting of index cases and their parents provide a rich source of information for assessment of linkage disequilibrium, as has recently been reviewed elsewhere (Spielman and Ewens 1996). Previous analytic approaches, such as the transmission/disequilibrium test (TDT), focus on the transmission of alleles from parents to affected offspring. This approach is able to detect both effects attributable directly to alleles inherited by the case and effects of unidentified genes that can be presumed to be in linkage with such alleles. The TDT has provided a powerful method to test for linkage in the presence of association. There are, however, indirect pathways of genetic influence not detectable by this method. One example would be effects of a mother's genotype on the development of a fetus, through the intrauterine environment. The alleles that she transmits to her child could be irrelevant under such a mechanism, and purely maternal effects would therefore be undetectable by the TDT. Parental imprinting might also be important to risk, through a mechanism in which the effect of a given allele on the offspring is greater or lesser depending on the parental source of that allele.

These alternative varieties of genetic effect have not been addressed within the context of studies of affected individuals and their parents. Appreciating the possible relevance of the maternal genotype, Mitchell (1997) has suggested inclusion of the child's maternal grandparents in a family-based study. Unfortunately, such studies can be impracticable, because the grandparents are often not available. The TDT could, in principle, also be extended to search for evidence of parental imprinting in studies of case-parent triads, but this has not been described.

We propose a likelihood-based method of analysis for case-parent–triad data that can detect effects of the mother's as well as the offspring's genotype (Wilcox et al., in press) and that can readily be generalized to detect parental imprinting. Our log-linear approach produces estimates of relative risks that are inherently adjusted for population stratification. It is flexible in that it allows

for the possibility that individuals with a single copy of the variant allele have a different risk than do individuals with two copies.

The purpose of the present paper is twofold. In the simple situation in which only the genes inherited by the child affect risk, we show that our method yields a likelihood ratio test (LRT) that can serve as an alternative to the TDT for testing for linkage disequilibrium. We simulate case-parent–triad studies to compare the power of the two, with and without a background of maternal genetic effects. Second, we extend our approach to include possible effects of parental imprinting. We then propose a different test statistic for imprinting, developed in the spirit of the TDT. The power of this "transmission asymmetry test" (TAT) is contrasted with that of the LRT, again via simulations.

The paper is structured as follows. We first describe the log-linear method for the usual causal scenario in which the inherited genotype is responsible for the effect. We then discuss the close relationship of this method to the conditional-on–parental genotype (CPG) maximum-likelihood approach described by Schaid and Sommer (1993). We then describe how the model would be generalized to allow for possible maternal genetic effects. We present results of simulations to compare the power of the LRT, the TDT, and the score statistic, as tests for linkage disequilibrium. The simulations are done with and without a background of maternal genetic effects and under dominant, recessive, and gene-dose models for causation. We then generalize the log-linear model to allow for possible effects due to imprinting, and we pose a second possible imprinting test, developed as a natural extension of the TDT. Finally, we provide simulation results to demonstrate the operating characteristics of the two tests under various imprinting scenarios.

## Background

Building on insights of Falk and Rubinstein (1987), Spielman et al. (1993) introduced the $\chi^2$ TDT in 1993. This test is based on comparison of the proportion of heterozygous parents who have transmitted a particular allele to their affected child with the expected proportion (.5). Others (Schaid and Sommer 1993) have proposed score-statistic methods based on the likelihood, which can be tailored to particular modes of transmission, such as dominance. Conditional likelihood methods have also been developed, which condition on the number of copies of the variant allele carried by each of the parents and exploit Mendelian inheritance from parent to child (Self et al. 1991).

These family-based approaches can be preferable to comparisons of randomly sampled cases and controls, because case-control studies can find unimportant associations that are not etiologic but reflect population structure. The TDT and the mating-type–stratified likelihood-based methods overcome this problem by comparing the genotype of cases to that of their parents, whose nontransmitted chromosomes serve as ethnically matched genetic controls, even in a population without random mating and not in Hardy-Weinberg equilibrium (HWE). For these methods to be valid, the only required assumption is that, under the null hypothesis, there is Mendelian transmission. (The affected-family–based controls method [Thomson 1995] was similarly motivated but leads to valid inference only under restrictive assumptions [Spielman et al. 1993] and will not be considered further here.)

## The Log-Linear Likelihood Approach

Our likelihood-based approach allows for effects of the inherited genotype and is easily generalized to a wider range of causal scenarios. We assume that a particular allele, called the "variant," is of interest and that a series of cases and their biologic parents have been genotyped. (A gene may have just two alleles, or multiple alleles may have been grouped into two functional categories, one of which is the "variant.") We do not need to assume HWE. Let "M," "F," and "C" denote the number of copies of the variant carried by the mother, father, and child (case), respectively. We assume mating symmetry, in the sense that, in the population at large, the probability that F = 2 and M = 1 is the same as the probability that F = 1 and M = 2, and so on. There are then six distinct mating types defined by the paired parental genotypes (Schaid and Sommer 1993).

If we had randomly sampled a set of children from the population, the corresponding child-parent triads could provide genotypes (M,F,C) that would, under Mendelian inheritance, fall into a multinomial with just the 15 possible categories shown in table 1. Under HWE, the relative frequencies (table 1, column 3) are simple polynomials in $p$, the allele frequency. If HWE does not apply, then we need to take the relative frequencies for the mating types into account, and the frequencies can be expressed, by means of Mendelian inheritance, as in column 4 of table 1, where the sum of the relative-frequency parameters across the 15 categories is constrained to equal 1.

We are interested, however, not in the distribution of such triads in the population at large but in the distribution of such triads when the triad has been selected because the child has the condition under study. We can apply Bayes's theorem to write the conditional probability for (M,F,C), given that the child is a case. The counts still fall into a multinomial distribution with 15 categories, but the relative frequencies now are distorted

**Table 1**

**Population Frequencies of Case-Parent Triads, With and Without HWE**

| Mating Type: M,F,C Genotype | Theoretical Frequency: In HWE Population | Theoretical Frequency: When HWE Is Not Assumed |
|---|---|---|
| 1: | | |
| 2,2,2 | $p^4$ | $\mu_1$ |
| 2: | | |
| 2,1,2 | $p^3(1-p)$ | $\mu_2$ |
| 2,1,1 | $p^3(1-p)$ | $\mu_2$ |
| 1,2,2 | $p^3(1-p)$ | $\mu_2$ |
| 1,2,1 | $p^3(1-p)$ | $\mu_2$ |
| 3: | | |
| 2,0,1 | $p^2(1-p)^2$ | $\mu_3$ |
| 0,2,1 | $p^2(1-p)^2$ | $\mu_3$ |
| 4: | | |
| 1,1,2 | $p^2(1-p)^2$ | $\mu_4$ |
| 1,1,1 | $2p^2(1-p)^2$ | $2\mu_4$ |
| 1,1,0 | $p^2(1-p)^2$ | $\mu_4$ |
| 5: | | |
| 1,0,1 | $p(1-p)^3$ | $\mu_5$ |
| 1,0,0 | $p(1-p)^3$ | $\mu_5$ |
| 0,1,1 | $p(1-p)^3$ | $\mu_5$ |
| 0,1,0 | $p(1-p)^3$ | $\mu_5$ |
| 6: | | |
| 0,0,0 | $(1-p)^4$ | $\mu_6$ |

in a mathematically simple way by the corresponding relative risks. We reexpress the conditional probability of (M,F,C), conditional on the child being a case, as follows: $P[\text{M,F,C}|\text{D}] = P[\text{D}|\text{M,F,C}]P[\text{C}|\text{M,F}]P[\text{M,F}]/P[\text{D}]$.

If the allele under study is a marker for the disease gene, then the first factor will depend, in a complicated way, on all three counts—M, F, and C—through the recombination rate, $\theta$ (Schaid 1996). Consider "scenario A," in which a disease gene is under study ($\theta = 0$) and what matters is the number of copies of the variant allele inherited by the child (the parental genotypes are irrelevant once we know C). The multinomial would take the form represented in table 2, where $R_1$ ($R_2$) is the risk for a child with one copy (two copies) of the variant, divided by the risk for a child with no copies. The parameters $\mu_k$ now simply serve as stratification parameters for the six parental mating types. The same sort of structure arises under HWE, but a constant factor must be included to serve as a normalization parameter, ensuring that the relative frequencies sum to 1.

## Relation to CPG Maximum Likelihood

The specification of the scenario A likelihood, given by table 2, which includes stratum parameters for parental mating type, is functionally equivalent to the likelihood considered by Schaid and Sommer (1993) that was constructed to be CPG. The relative risks estimated via the multinomial of table 2 are, in fact, identical to the CPG maximum-likelihood estimates, with the advantage that the stratified likelihood can easily be fitted by use of standard software for Poisson regression. As Schaid and Sommer pointed out, only data from mating types 2, 4, and 5 are informative. In effect, when the stratified likelihood of table 2 is fitted, the counts for the noninformative mating types play a passive role and have no influence at all on estimation, standard errors, or significance tests.

## The Combined Model, Also Allowing for Maternal Effects

A major advantage to the log-linear approach is its ease of generalization to other causal scenarios. First, we will consider "scenario B," in which the mother's genotype is directly relevant to risk. Later, we will generalize to a scenario with parental imprinting effects.

A multinomial model exactly analogous to that developed for scenario A can be developed for scenario B, in which the mother's genotype is now what matters. Notice that, although scenario A involves (apparently) preferential transmission of the variant allele to cases whereas scenario B does not, both scenarios are susceptible to the same simple Bayes's theorem approach (as

**Table 2**

**Frequencies in Case-Parent Triads, under Scenario A**

| Mating Type: M,F,C Genotype | Theoretical Frequency[a] |
|---|---|
| 1: | |
| 2,2,2 | $R_2\mu_1$ |
| 2: | |
| 2,1,2 | $R_2\mu_2$ |
| 2,1,1 | $R_1\mu_2$ |
| 1,2,2 | $R_2\mu_2$ |
| 1,2,1 | $R_1\mu_2$ |
| 3: | |
| 2,0,1 | $R_1\mu_3$ |
| 0,2,1 | $R_1\mu_3$ |
| 4: | |
| 1,1,2 | $R_2\mu_4$ |
| 1,1,1 | $2R_1\mu_4$ |
| 1,1,0 | $\mu_4$ |
| 5: | |
| 1,0,1 | $R_1\mu_5$ |
| 1,0,0 | $\mu_5$ |
| 0,1,1 | $R_1\mu_5$ |
| 0,1,0 | $\mu_5$ |
| 6: | |
| 0,0,0 | $\mu_6$ |

[a] $R_1$ and $R_2$ are the relative risks associated with inheritance of one or two copies of the variant allele, respectively; and $\mu_j$ is the stratum parameter for the $j$th mating-type category.

given above) to derive the probabilities for the respective multinomial distribution.

Under such a generalization, the expected count in each cell (M,F,C) of the 15-nomial can be modeled in a log-linear form that captures both possible scenarios at once, as follows:

$$\ln[\mathrm{E}(n_{\mathrm{M,F,C}})] = \gamma_j + \beta_1 I_{\{\mathrm{C}=1\}} + \beta_2 I_{\{\mathrm{C}=2\}} + \alpha_1 I_{\{\mathrm{M}=1\}} + \alpha_2 I_{\{\mathrm{M}=2\}} + \ln(2) I_{\{\mathrm{M}=\mathrm{F}=\mathrm{C}=1\}} \; . \qquad (1)$$

Scenario A and Scenario B are each special cases of this model. The index, *j,* is a function of (M,F) and corresponds to the parental-mating-type stratum. In equation (1), $I_{\{C=1\}}$ denotes a "dummy" independent variable that becomes 1 when C = 1 and is 0 otherwise. If the child's genotype and the mother's genotype do not combine multiplicatively in their joint effect on risk, then product terms can also be included as necessary. This model is easily modified to allow for either a dominant model ($\alpha_1 = \alpha_2; \beta_1 = \beta_2$), by addition of the two indicator variables, for the mother and for the child, or a recessive model ($\alpha_1 = 0; \beta_1 = 0$), by omission of the single-allele-indicator variables. We later extend this model to also allow for possible effects due to parental imprinting.

Such a model can be fitted by use of widely available (Poisson regression) software—for example, in GLIM (generalized linear interactive modeling package [Baker and Nelder 1978]) or SAS (via the GENMOD procedure)—to maximize the corresponding multinomial likelihood and to estimate the parameters on the basis of maximum likelihood. The only complication is that the term $\ln(2)I_{\{M=F=C=1\}}$ (corresponding to the 2 multiplier for the [1,1,1] category in table 2) is not entered as an independent variable but, rather, must be declared as an "offset," so that it is incorporated as a linear term with its coefficient constrained to be 1.

The estimation of the relative-risk parameters is then straightforward. For example, estimation of $R_1$ is based on exponentiating the estimate for $\beta_1$. The corresponding relative risk for a maternal effect associated with a single copy, which we denote by "$S_1$," is estimated by exponentiating the estimated $\alpha_1$. Confidence intervals can be derived by exponentiating the limits of the corresponding standard-error–based confidence intervals for the corresponding $\alpha$ and $\beta$ parameters. Our simulations (Wilcox et al., in press) demonstrated that, with 100 case-parent triads, the power to detect a relative risk of 2.5 was close to .8, the coverage of the nominally 95% confidence interval (95% CI) was consistent with 95%, and there was very little bias in estimation of the relative risks.

If the investigator is willing to assume HWE, a further simplification of the aforementioned model is possible. The stratum parameters are replaced in the linear argument by a linear term in M + F (see the Appendix). This carries the additional advantage that one can estimate the prevalence of the variant allele, using only case-parent–triad data. However, power for detection of violation of assumed HWE is evidently small (simulation results are not shown), suggesting that the assumption of HWE cannot be reliably verified in practice.

## Power Considerations

The model gives rise to $\chi^2$ LRTs in the usual way. For example, one can test for whether the child's genotype (with respect to the variant allele) has any effect on risk, by removing the two indicator variables corresponding to the child's genotype (while leaving the maternal indicator variables in the model) and computing twice the change in the logarithm of the maximized likelihood. Under the null hypothesis that the child's genotype does not matter within any of the parental-mating-type categories—that is, Mendelian transmission of the allele to cases—this LRT statistic will be distributed as $\chi^2$ with 2 df. By contrast, in the presence of linkage disequilibrium, the LRT will tend to be elevated compared with the $\chi^2$.

The separate effects due to maternal and inherited genotypes can easily be distinguished with the aforementioned combined model. In fact, we have shown that, once one conditions on mating type, the inherited number of copies, C, and the maternal number of copies, M, are statistically independent under Mendelian inheritance; hence, testing and estimation based on the model given above yield results that are completely orthogonal for the maternal and the inherited genotype (Wilcox et al., in press). One consequence is that the $\chi^2$ LRT statistic for the child's genotype, adjusting for possible maternal effects, is identical to an LRT statistic that does not adjust for maternal effects. Thus, an analysis that stratifies on parental mating type and ignores possible maternal effects will not cause one spuriously to attribute to the inherited genotype effects that are in fact maternal.

For detection of an effect of the inherited genotype (scenario A), the LRT and the TDT can be regarded as competitors, regardless of whether maternal effects are also present. We therefore performed simulations to compare the power of the LRT with that of the TDT for the same data. We also included, for comparison, the score statistic proposed by Schaid and Sommer (1993), which was developed on the basis of the same mating-type–stratified model but without allowance for a maternal contribution.

The data were simulated on the basis of several combinations of parameter values. All simulations and analyses were performed by use of the GLIM package (Baker and Nelder 1978). Scenario A (inherited genotype ef-

**Table 3**

**Fraction Rejecting the Null Hypothesis (No Association, No Linkage) for the Child's Genotype, in 1,000 Simulated Studies of 100 Case-Parent Triads, with 95% CI for Empirical Estimates of Power**

| | Fraction (95% CI) Rejecting Null Hypothesis, in | | |
|---|---|---|---|
| Model[a] | TDT | LRT | Score Test |
| Null ($R_1 = R_2 = 1 = S_1 = S_2$) | .050 (.04–.06) | .058 (.04–.07) | .050 (.04–.06) |
| Child only: | | | |
| Dominant: | | | |
| $R_1 = R_2 = 2.5$; $S_1 = S_2 = 1.0$ | .746 (.72–.77) | .842[b] (.82–.87) | .829[c] (.81–.85) |
| $R_1 = R_2 = 3.0$; $S_1 = S_2 = 1.0$ | .855 (.83–.88) | .924[b] (.91–.94) | .918[c] (.90–.94) |
| Recessive: | | | |
| $R_2 = 2.5$; $R_1 = S_1 = S_2 = 1$ | .347 (.32–.38) | .537[b] (.51–.57) | .552[c] (.52–.58) |
| $R_2 = 3.0$; $R_1 = S_1 = S_2 = 1$ | .501 (.47–.53) | .727[b] (.70–.76) | .742[c] (.71–.77) |
| Gene dose ($R_1 = 2$; $R_2 = 3$; $S_1 = S_2 = 1$) | .758[d] (.73–.79) | .684 (.65–.71) | .676 (.65–.71) |

[a] $R_1$ and $R_2$ are as in table 2. $S_1$ and $S_2$ are the relative risks associated with the mother carrying one or two copies, respectively, of the variant allele.

[b] LRT was significantly more powerful than TDT (two-sided $P < .01$).

[c] Score test was significantly more powerful than TDT (two-sided $P < .01$).

[d] TDT was significantly more powerful than either LRT or score test (two-sided $P < .01$).

fects) could alternatively be present or absent; the genetic model could be dominant, recessive, or a gene dose (in which a single copy increases risk but in which two copies have an even greater effect); the adjusted relative risks for the disease gene could be 2.5 or 3. We did not assume HWE. Population structure was such that, for a 20% subpopulation, the gene prevalence was .3 and the background risk was .05, in those without the variant; for the remaining, 80% subpopulation, the gene prevalence was .1, and the background risk was .01. For each choice of parameters and scenarios, 1,000 studies were simulated, each of which included 100 case-parent triads. In a very few simulated data sets in which the LRT could not be considered valid, because the maximum-likelihood estimate for either $R1$ or $R2$ was 0 or $\infty$—that is, the maximum point was not in the interior of the parameter space—the score test was substituted for the LRT. Although confidence intervals are given for the power, the results for the various tests are properly compared by paired analysis based on the simulations in which the results differed (one test rejected and the other did not). Significant differences are indicated.

In a given setting, one could use prior knowledge to design a more powerful LRT (or score statistic). For example, if one had prior evidence that the proper genetic model was dominant, then one could add the two indicator variables, as described above, and estimate a single relative-risk parameter in model (1). The resulting LRT, with 1 df, would have enhanced power, if the dominant were the true model. Similarly, if one suspects that a gene-dose effect is the proper model, then one can devise an LRT that is optimal for this alternative, by fitting a model that specifies linearity in C in model (1). This 1-df test would have enhanced power if, in fact, there were a gene-dose effect. Our simulations were run under the conservative assumption that no such prior information has been brought to bear on the problem. Thus, the investigator is assumed to be testing the null hypothesis: $R_1 = R_2 = 1$, against all possible alternatives, with the 2-df test.

## Results of Simulations for Power

Table 3 shows results under models in which only the inherited genotype of the child influences risk. All three tests had estimated type 1 error rates that were fully consistent with the nominal .05, as shown in row 1. The LRT based on our log-linear model and the score test had similar power, and, under both dominant and recessive models, both markedly outperformed the TDT. By contrast, under a model in which the number of copies of the variant allele was important—in that a single copy conferred a relative risk of 2 to the child, whereas two copies conferred a relative risk of 3—the TDT outperformed both of the likelihood-based methods.

All three tests had good operating characteristics against a background of a maternal genotype effect, despite the inevitable correlation between mother and child (data not shown). In fact, all three generally performed better in the presence of maternal effects than they did under a pure scenario A. This difference most likely arises because, with an allele that is not common, the presence of maternal genetic effects enriches the distribution of parental genotypes, by, in effect, "oversampling" maternal carriers of the variant allele. Thus, when there is a maternal effect and when the allele is not common, the average number of informative families is increased, and the power for both the TDT and the LRT is correspondingly increased.

As mentioned above, one could improve on the power of the LRT, by tailoring it to a particular alternative, such as a gene-dose effect. If, under the gene-dose model, the effect of the number of copies of the allele carried by the child is modeled as linear, by revision of the argument of expression (1), this linearized LRT reveals power that was, in our simulations, identical to that of the TDT. The reasons for this equivalence are given in the Discussion section.

## Extension of the Model, to Allow for Imprinting

Potentially, the risk to the offspring could depend on whether the inherited copy is maternal or paternal in origin, because of imprinting mechanisms. To account for such mechanisms, the general model (1) can be extended as follows:

$$\ln[\mathrm{E}(n_{\mathrm{M,F,C}})] = \gamma_j + \beta_1 I_{\{\mathrm{C}=1\}} + \beta_2 I_{\{\mathrm{C}=2\}} + \alpha_1 I_{\{\mathrm{M}=1\}} + \alpha_2 I_{\{\mathrm{M}=2\}} + \varepsilon_{\mathrm{F}} I_{\{\text{F-derived copy}\}} + \varepsilon_{\mathrm{M}} I_{\{\text{M-derived copy}\}} + \ln(\eta) I_{\{\mathrm{M=F=C}=1\}}\ , \qquad (2)$$

where $\eta$ denotes the sum: $\exp(\epsilon_{\mathrm{F}}) + \exp(\epsilon_{\mathrm{M}})$. The parameter interpretation is that $I_{\mathrm{F}} = \exp(\epsilon_{\mathrm{F}})$ (or that $I_{\mathrm{M}} = \exp[\epsilon_{\mathrm{M}}]$) estimates the risk for a child who inherits a paternally derived (or maternally derived) copy of the allele, divided by the risk for a child who inherits no copy of the allele. All mating types except 1 and 6 should be informative for assessment of imprinting effects. For every category but one in table 2, the parental origin of the alleles is unambiguous. The one troublesome category is (1,1,1), in which, under Mendelian inheritance, half the children would have inherited the paternal copy and half would have inherited the maternal copy. If we had complete data on parental origin, then model (2) could be fit directly, suggesting that statistical methods for missing data can be applied. Here the missing data are just the parental origin for triads in the (1,1,1) cell—that is, heterozygous cases with two heterozygous parents.

Table 4 shows the 16-cell multinomial that we would need with hypothetical complete data on parental origin of the alleles. We can directly observe 14 of the 16 counts; for two of the categories, however, we can observe only the sum. This is because the parental source of the allele must be treated as unknown for the category (1,1,1).

The likelihood corresponding to the observable data has, as the expected count for the (1,1,1) cell, the sum $(I_{\mathrm{M}} + I_{\mathrm{F}})S_1R_1\mu_4$. Once all the relative-risk parameters have been estimated, the maximized log likelihood for the observed data is simple to compute.

**Table 4**

**Frequencies in Case-Parent Triads, under Model 2**

| M,F,C Genotype and Mating Type | Parental Origin of Allele(s) | Theoretical Frequency[a] |
|---|---|---|
| 1: | | |
| 2,2,2 | MF | $I_{\mathrm{M}}I_{\mathrm{F}}S_2R_2\mu_1$ |
| 2: | | |
| 2,1,2 | MF | $I_{\mathrm{M}}I_{\mathrm{F}}S_2R_2\mu_2$ |
| 2,1,1 | M | $I_{\mathrm{M}}S_2R_1\mu_2$ |
| 1,2,2 | MF | $I_{\mathrm{M}}I_{\mathrm{F}}S_1R_2\mu_2$ |
| 1,2,1 | F | $I_{\mathrm{F}}S_1R_1\mu_2$ |
| 3: | | |
| 2,0,1 | M | $I_{\mathrm{M}}S_2R_1\mu_3$ |
| 0,2,1 | F | $I_{\mathrm{F}}R_1\mu_3$ |
| 4: | | |
| 1,1,2 | MF | $I_{\mathrm{M}}I_{\mathrm{F}}S_1R_2\mu_4$ |
| 1,1,1 | M | $I_{\mathrm{M}}S_1R_1\mu_4$ |
| 1,1,1 | F | $I_{\mathrm{F}}S_1R_1\mu_4$ |
| 1,1,0 | | $S_1\mu_4$ |
| 5: | | |
| 1,0,1 | M | $I_{\mathrm{M}}S_1R_1\mu_5$ |
| 1,0,0 | | $S_1\mu_5$ |
| 0,1,1 | F | $I_{\mathrm{F}}R_1\mu_5$ |
| 0,1,0 | | $\mu_5$ |
| 6: | | |
| 0,0,0 | | $\mu_6$ |

[a] $I_{\mathrm{M}}$ and $I_{\mathrm{F}}$ are the relative risks associated with inheriting a copy from the mother or the father, respectively. $R_1$, $R_2$, $S_1$, $S_2$ are as in tables 2 and 3.

A problem of colinearity arises in the fully parameterized model (2), because $\mathrm{C} = I_{\{\mathrm{C}=1\}} + 2I_{\{\mathrm{C}=2\}} = I_{\{\text{F-derived copy}\}} + I_{\{\text{M-derived copy}\}}$. Consequently, model (2) in its full form is not statistically identifiable. This means that, when imprinting by both parents is added to a full-background model, only one additional parameter is fitted. Moreover, the change in minus twice the log likelihood is (under the null that there is no parental imprinting), against that full-background model, a $\chi^2$ with 1, not 2, df. There is a corresponding problem in interpretation, in that, against the fully elaborated background model (1), a model that now includes paternal imprinting will fit as well as one that includes maternal imprinting. Because of this colinearity problem, meaningful constraints need to be imposed to allow useful inference with regard to parental imprinting, on the basis of model (2). The simplest strategy is to leave out one of the parental imprinting parameters. Or one can either assume a dominant model or a recessive model or choose the one that fits best, thus reducing the number of parameters that must be estimated. Any such parameter restriction will yield statistical identifiability.

## Using the Expectation-Maximization (EM) Algorithm to Maximize the Imprinting Likelihood

To estimate the parameters, we apply the EM algorithm (Dempster et al. 1977). To fit this algorithm in

this simple situation, one estimates the expected value of the missing cell counts, conditional on the current estimates of all the parameters and on the observed data. If "$n_{1,1,1}$" denotes the observed count of triads in the (1,1,1) cell, then the estimated count when the child's copy was paternal in origin is given by

$$n_{1,1,1}\frac{\hat{I}_F}{\hat{I}_F + \hat{I}_M} ,$$

where $\hat{I}_F$ and $\hat{I}_M$ denote the current maximum-likelihood estimates for the imprinting relative risks for the father and the mother, respectively. The estimated counts are then treated as real data, and the likelihood in table 4 is maximized by use of standard software (the M step). Then the estimation (the E step) is repeated. In this way, the E and M steps are alternated until convergence is achieved. In the simulations that we have analyzed, this normally takes $\leqslant\sim 12$ iterations and always occurs in $\leqslant 30$ iterations. The log likelihood must, of course, not be based on the estimated data and the pseudocomplete data of table 4 but, rather, on the observed data likelihood, for which the fitted probability for the (1,1,1) cell is given by

$$\left(\hat{I}_M + \hat{I}_F\right)\hat{S}_1\hat{R}_1\hat{\mu}_4 .$$

## An Alternative TDT-Like Test for Imprinting

One can compute a 1-df test, TAT, as follows: Calculate, among heterozygous mothers not married to heterozygous fathers, the number who transmitted and the number who did not transmit the variant allele to the affected offspring. Denote these numbers as "*a*" and "*b,*" respectively. Then do the same for the heterozygous fathers not married to heterozygous mothers, enumerating both a number, *c,* for the transmitters and a number, *d,* for the nontransmitters. Families in which both parents are heterozygous (mating type 4) must be excluded as uninformative. The resulting two-by-two table made up of *a–d* is then tested for equality of the transmission rates, by means of the usual $\chi^2$ statistic for comparison of proportions. This yields a transmission-asymmetry statistic and a 1-df $\chi^2$ TAT.

## Imprinting Simulations

Imprinting was simulated under the same population-structure model described above for comparing the power of the LRT with that of the TDT. For each of several combinations of parameter values, 1,000 studies were simulated, with 100 case-parent triads in each. The models considered included, in addition to a model that was fully null (except for mating-type effects), one with purely paternal imprinting, with paternal transmission of the variant conferring a relative risk of 2.5 ($I_F = 2.5$), and one with purely maternal imprinting, with maternal transmission of the variant conferring a relative risk of 2.5 ($I_M = 2.5$). Each of these was also simulated against a context of background effects involving, in turn, the genotype(s) of the child or the mother, conferring an additional relative risk of 2.5, both under dominant models. Although data were simulated under a dominant model, dominance was not imposed in the modeling, since this would generally not be known by the investigator.

The likelihood tests were done in two ways—by testing imprinting (for one parent) against the correct background model and by testing imprinting against a fully parameterized model, allowing for $R_1$, $R_2$, $S_1$, and $S_2$. For the simplest case, where there is only parental imprinting, the imprinting effect was also tested against a model in which the child's genotype might also be independently related to risk, to see how much power would be lost in making that unnecessary adjustment. All tests were based on $\chi^2$ statistics (1 df), which are inherently two sided. The TAT was also computed for each simulated study, so that its power could be assessed.

Parameter estimation for the imprinting relative risk, $I_F$ or $I_M$, was also done under various choices of background models. These approaches are, unlike the maternal and child tests, not orthogonal and can give very different results, especially for power. Bias in estimation was assessed by comparing the exponentiated average of the estimated coefficients to the known relative risk, 2.5. Confidence intervals are given for these estimates, based on the empirical standard errors from the 1,000 simulated studies.

## Results of Analysis of Imprinting Simulations

Results are given in table 5. Models for which the correct background model was used in the adjustment have been specified. The EM algorithm converged consistently by the 30th iteration. The estimated imprinting relative risks were very close to the known 2.5, except that there was slight upward bias for the sample size considered. As would be expected, the power results for pure imprinting scenarios ($R_1 = R_2 = S_1 = S_2 = 1$) were symmetric for maternal and paternal imprinting, regardless of whether the child's genotype was (needlessly) adjusted for (see table 5, rows 1 and 4). Power was very good, >.90, against the properly specified null background model. By contrast, power was poor against a fully parameterized model. This would be expected, given the colinearities in the independent variables. The power for the TAT was approximately as poor as that for the fully parameterized background analysis. In sim-

**Table 5**

**Results for Tests for Detection of Imprinting, with Use of LRT, Based on 1,000 Simulated Studies Each**

| Scenario Assumed ($N$ = 100 Families) | Overall Estimated Relative Risk (95% CI); Fraction of Tests Rejecting Imprinting Null Hypothesis [Power for TAT][a] | | |
|---|---|---|---|
| | Mating Type Only | Mating Type and C (Rows 1, 2, 4, and 5) or Mating Type and M (Rows 3 and 6) | Mating Type, C and M |
| $I_F = 2.5$: | | | |
| $I_M = R_1 = R_2 = S_1 = S_2 = 1$ | $I_F$: 2.48 (2.44–2.52); .924[b] | $I_F$: 2.53 (2.47–2.59); .761 | $I_F$: 2.55 (2.44–2.66); .328 [.329] |
| $R_1 = R_2 = 2.5$; $S_1 = S_2 = I_M = 1$ | Invalid model | $I_F$: 2.56 (2.50–2.61); .873[b] | $I_F$: 2.54 (2.42–2.67); .275 [.391] |
| $S_1 = S_2 = 2.5$; $R_1 = R_2 = I_M = 1$ | Invalid model | $I_F$: 2.60 (2.53–2.67); .735[b] | $I_F$: 2.59 (2.49–2.70); .352 [.322] |
| $I_M = 2.5$: | | | |
| $I_F = R_1 = R_2 = S_1 = S_2 = 1$ | $I_M$: 2.50 (2.45–2.54); .928[b] | $I_M$: 2.55 (2.49–2.61); .777 | $I_M$: 2.55 (2.44–2.66); .315 [.285] |
| $R_1 = R_2 = 2.5$; $S_1 = S_2 = I_F = 1$ | Invalid model | $I_M$: 2.54 (2.49–2.59); .877[b] | $I_M$: 2.60 (2.48–2.72); .264 [.388] |
| $S_1 = S_2 = 2.5$; $R_1 = R_2 = I_F = 1$ | Invalid model | $I_M$: 2.56 (2.51–2.62); .868[b] | $I_M$: 2.64 (2.52–2.78); .281 [.323] |

[a] Type 1 error rate .05.

[b] Model is not overparameterized compared with the true scenario.

ulations of the fully null model (not shown) for the LRT and the TAT, the type 1 error rates were consistent with the nominal .05.

For the more complex settings, in which there is, in addition to imprinting, an independent effect of either the child's genotype (scenario A) or the mother's genotype (scenario B), the power remained generally high for the LRT, provided that no unnecessary adjustments were included in the modeling. The one exception to this was in the scenario in which paternal imprinting occurs simultaneously with a maternal genetic background, in which the power against the properly specified background dropped to .74. Power remained low for the TAT.

## Discussion

With case-parent–triad data, the relative risks associated with a particular variant allele can be estimated by maximum likelihood, by use of widely available statistical software. The log-linear model allows for causal scenarios in which the child's own genotype is directly relevant to risk, in which the mother's genotype is directly relevant, in which parental imprinting plays an important role, or in which the truth is some combination of these. In the special case where risk is determined by the child's own genotype, the LRT is closely related both to the CPG method described by Schaid and Sommer (1993) and to the conditional maximum-likelihood approach described by Self et al. (1991) and applied by others to multigene analyses (Langholz et al. 1995; Thomas et al. 1995).

A model including imprinting is somewhat more difficult to fit than are the models involving either scenario A or scenario B, in that an iterative missing-data procedure is required to develop the maximum-likelihood estimates. However, the simpler, TDT-inspired TAT for imprinting had low power compared with that of the LRT. This is presumably at least in part because the TAT does not use the information from mating type 4, in which both parents are heterozygous, whereas the LRT makes full use of these families.

We had previously reported that a study of 100 case-parent triads would yield a power of ~.80 for detection of an effect of the maternal genotype (a relative risk of 2.5, dominant model) (Wilcox et al., in press). Our current results suggest that power for detection of maternal or paternal imprinting may be even higher, ~.90, for the same population structure and sample size.

The relative-risk parameters that arise in these models, for imprinting, maternal effects, and effects of the child's own genotype, are interpretable as etiologically relevant "population-structure–adjusted" relative risks, in that the stratification on parental mating type has adjusted for effects of population structure. Thus, the model provides a way to estimate parameters and to test linkage disequilibrium for the usual causal scenario, but it also allows for generalization to a broader range of causal scenarios (e.g., scenario B) that do not necessarily include any appearance of distortion in transmission from parent to child.

We have described these results in the context of a "variant" allele, without clearly specifying whether the gene is considered to be a candidate disease gene (perhaps with low penetrance) or a marker that is in linkage with a disease gene. If the gene under study is in fact a marker, then the model laid out, in table 2, for scenario A is not quite correct, because the risk in a child who carries a certain number of copies of the marker depends also on the genotypes of the child's parents. Because of recombination, the risk in a homozygous child with one homozygous parent should be higher than the risk in a

homozygous child with two heterozygous parents. Nevertheless, the model in table 2 is correct under the null hypothesis of no linkage disequilibrium, and hence tests based on this model are statistically valid, if not optimal, even for studies of marker alleles.

Because it tests for linkage and association simultaneously, the TDT can be thought of as a test for linkage in the presence of known association, or as a test for both linkage and association (Spielman and Ewens 1996). The log-linear model that we have described yields an LRT for the effect of the child's genotype (scenario A) that is also simultaneously a test for association and linkage, and thus the LRT can be regarded as a competitor with the TDT. Our simulations suggest that, despite the inherent disadvantage of having 2 df rather than 1 df, both the LRT and the closely related score statistic proposed by Schaid and Sommer (1993) often outperform the TDT.

This enhanced power of the likelihood-based methods under a dominant or a recessive model is most likely due to how they use parental information. The TDT considers only transmission and treats all heterozygous parents as independent. The likelihood-based methods exploit the added information related to joint transmission from pairs of parents. This distinction is most clear when one considers the triads that fall into the (1,1,1) category. For the TDT, these triads, even if numerous, do nothing but decrease the $\chi^2$ statistic, since one heterozygous parent has transmitted a copy of the gene whereas the other heterozygous parent has not. Although each such family adds two to the denominator of the computed TDT statistic, it adds nothing to the numerator. On the other hand, the likelihood-based methods use the information that the child in such a triad did inherit a copy of the gene, and larger than expected counts in the (1,1,1) category will increase the value of the test statistic. Conceptually, looking only at which heterozygous parent did (or did not) transmit the gene to an affected child, as is done by the TDT, sacrifices important information related to what the child actually received as the *joint* transmission from the two parents. Thus, if parents who are both heterozygous jointly transmit a single copy to the affected child more than twice as often as they jointly transmit no copies, this provides important evidence of a genetic effect. Analysis of parents singly, as in the TDT, overlooks this evidence.

Our power results were identical for the TDT and the linearized LRT, the latter being a test in which model (1) has been modified to have an argument linear in C. The reason for this equivalence is quite simple. Under the linearized model, the maximum-likelihood estimate for the coefficient of C can be shown to be the logarithm of $b/c$, where $b$ ($c$) is the number of heterozygous parents who did (did not) transmit the variant to their offspring. In the linearized LRT, $b/c$ is the estimated relative risk for carrying a single copy of the variant and $(b/c)^2$ is the estimated relative risk for carrying two copies of the variant. The TDT is defined as $(b - c)^2/(b + c)$. Both tests are thus testing the null hypothesis that, conditional on $b + c$, $b$ will, on average, be half of the total. The two tests are asymptotically equivalent, the TDT being the score test for the linearized model ($R_2 = R_1^2$) (Schaid and Sommer 1994) and the LRT being the test based on the change in the maximized likelihood. Apparently, a sample of 100 families is close enough to infinite for this equivalence to hold empirically. Thus the TDT can be regarded as a linearized version of the LRT, and in this sense the (not necessarily linearized) LRT provides a generalization of the TDT.

It follows that the TDT will be statistically optimal only when the relative risk associated with carrying two copies of the variant is the square of the relative risk associated with carrying one copy. This gives additional insight into its relatively poor performance under the dominant and the recessive models, which violate that pattern.

Finally, the LRT, the TDT, and the score test all assume Mendelian transmission under the null hypothesis, and all are vulnerable to distortion if this assumption fails. For studies of birth defects, such failure is not implausible—for example, homozygosity for the variant allele could be incompatible with survival of the early embryo. Spielman and Ewens (1996) discussed this issue in relation to distortion of the meiotic process itself and suggested that one could rule out such a phenomenon by studying the unaffected siblings. Such a strategy should work well if the penetrance is low. Another issue has to do with survival of the affected fetus to term, since we must consider not just the occurrence of a birth defect such as spina bifida but also its occurrence among fetuses surviving to birth and therefore eligible for study. However, we have shown elsewhere (Wilcox et al., in press) that poor survival of affected fetuses has no effect on the multinomial distribution for case-parent triads, provided only that the conditional probability of survival among babies with the defect does not depend on the case's or the parents' genotype.

In summary, our proposed method for analysis of case-parent triads provides an easy-to-apply test of linkage disequilibrium that can distinguish between effects of the offspring's genotype and prenatal effects of the mother's genotype and that can be adapted to test for parental imprinting. The model can be constructed to make use of HWE for studies of well-mixed populations, but this assumption is not required in general. Simulations suggest that the LRT based on the log-linear model is, under both recessive and dominant models, more powerful than the TDT. Estimation of population-structure–adjusted relative risks is also straightforward, with use of standard software for Poisson regression. Models

that include an imprinting scenario require some specialized software, but the resulting likelihood-ratio statistic yields a powerful test that outperforms a TDT-inspired alternative.

## Acknowledgments

We thank Drs. Norman Kaplan, Marcy Speer, and David Umbach for helpful comments and discussions.

## Appendix

### Analysis to Assess and Exploit HWE

The theoretical distribution of case-parent triads under the assumption of HWE can be generated from model (1), by substitution of the logarithms of the polynomials in $p$ (shown in table 1) for the $\gamma_j$ and by inclusion of a normalizing constant to ensure that the probabilities sum to 1.0.

The resulting distribution can be fitted by use of standard log-linear software and, in fact, can be shown to be a reduced version of the mating-type–stratified model. To see this, note that the polynomials in $p$ shown in table 1 can be written in the following form: $p^{M+F}(1-p)^{4-M-F}$, and this can be rewritten as $\{[p/(1-p)]^{M+F}\}(1-p)^4$. Taking logarithms, we can write the logarithm of the expected count under the HWE model as follows:

$$\gamma + \phi(M + F) + \beta_1 I_{\{C=1\}} + \beta_2 I_{\{C=2\}} + \ln(2) I_{\{M=F=C=1\}} .$$

The parameter, $\phi$, is the logit of $p$, the prevalence of the variant allele. Thus, if HWE holds, one can estimate the allele prevalence in the population by using case-parent triads only, by taking the anti-logit of $\phi$, which is $1/[1+\exp(-\phi)]$. A confidence interval for the prevalence can be based on the confidence interval for $\phi$, by taking the anti-logit of the two limits. A second consequence is that one can test the hypothesis that HWE holds, by means of case-parent–triad data. To do this, one performs an LRT based on comparison of the likelihood for the HWE model and that for the mating-type model. Comparing any two such models leads to a 4-df $\chi^2$ statistic, because the mating-type model requires six parameters, whereas the HWE model requires only two.

## References

Baker RJ, Nelder JA (1978) The GLIM system, release 3. Numerical Algorithms Group, Oxford

Falk C, Rubinstein P (1987) Haplotype relative risks: an easy, reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 51:227–233

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Series B 39:1–38

Langholz B, Tuomilehto-Wolf E, Thomas D, Pitkaniemi J, Tuomilehto J, DiMi Study Group (1995) Variation in HLA-associated risks of childhood insulin-dependent diabetes in the Finnish population. I. Allele effects at A, B, and DR loci. Genet Epidemiol 12:441–453

Mitchell LE (1997) Differentiating between fetal and maternal genotypic effects, using the transmission test for linkage disequilibrium. Am J Hum Genet 60:1006–1007

Schaid D (1996) General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 13:423–449

Schaid DJ, Sommer SS (1993) Genotype relative risks: methods for design and analysis of candidate-gene association studies. Am J Hum Genet 53:1114–1126

——— (1994) Comparison of statistics for candidate-gene association studies using cases and parents. Am J Hum Genet 55:402–409

Self S, Longton G, Kopecky K, Liang K (1991) On estimating HLA-disease association with application to a study of aplastic anemia. Biometrics 47:53–61

Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 59:983–989

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: 506–516

Thomas D, Pitkaniemi J, Langholz B, Tuomilehto-Wolf E, Tuomilehto E, DiMe Study Group (1995) Variation in HLA-associated risks of childhood insulin-dependent diabetes in the Finnish population. I. Haplotype effects. Genet Epidemiol 12:455–466

Thomson G (1995) Mapping disease genes: family-based association studies. Am J Hum Genet 57:487–498

Wilcox A, Weinberg C, Lie R. Distinguishing the effects of maternal and offspring genes through studies of "case-parent triads." Am J Epidemiol (in press)