A Generalization of the Transmission/Disequilibrium Test for Uncertain-Haplotype Transmission

David Clayton

MRC Biostatistics Unit, Institute of Public Health, Cambridge

Summary

A new transmission/disequilibrium-test statistic is proposed for situations in which transmission is uncertain. Such situations arise when transmission of a multilocus marker haplotype is considered, since haplotype phase is often unknown in a substantial number of instances. Even for single-locus markers, transmission is uncertain if one or both parents are missing. In both these situations, uncertainty may be reduced by the typing of further siblings, whose disease status may be unaffected or unknown. The proposed test is a score test based on a partial score function that omits the terms most influenced by hidden population stratification.

Introduction

Until recently, the literature on transmission/disequilibrium testing assumed, for the most part, that marker genotypes can be directly measured in affected cases and in both their parents and, therefore, that transmissions of haplotypes from parents to affected offspring can be counted. In two important situations this is not the case:

1. both parents may not always be available, particularly for diseases with late onset, and
2. since, if considered alone, binary markers such as single-nucleotide polymorphism (SNPs) may carry insufficient information, it may be necessary to consider haplotypes constructed from several such markers; however, haplotype phase is often uncertain.

In the former case, missing parental genotypes can be inferred from offspring genotypes, but Curtis and Sham (1995) have shown that restriction of the analysis to families in which such inference is certain leads to bias. Several alternative proposals have been published recently, mostly relying on the typing of further siblings, either unaffected or with unknown disease phenotype (Curtis 1997; Martin et al. 1997; Boehnke and Langefeld 1998; Horvath and Laird 1998; Schaid and Rowland 1998; Spielman and Ewens 1998). These approaches have been reviewed and compared by Monks et al. (1998).

Despite this work, no solution has yet been proposed for the problem of phase uncertainty for marker haplotypes, although the problems clearly have much in common. Again, our ability to infer parental haplotypes is enhanced by the typing of additional siblings, but it is to be expected that, here too, restriction of the analysis to families with known haplotype phase may lead to bias.

This report proposes a new and unified approach to these problems. In the next section, two different likelihood approaches to the analysis of transmission/disequilibrium studies are compared for the case in which all genotypes and transmissions are completely observed. The likelihood approach to uncertain transmission is then described, and its disadvantages are discussed. A new approach is described, and some extensions are outlined. These sections of the report emphasize the theoretical principles behind the method; detailed algebraic expressions are given in the Appendix.

Likelihoods, Score Tests, and the Transmission/Disequilibrium Test (TDT)

The idea that underlies TDTs is that, in the presence of association between a genetic marker and disease susceptibility (DS) (when such association is due to coincidence of linkage and gametic-phase disequilibrium between marker and DS gene), the probability of transmission of a marker gene from parents to an *affected* offspring is increased from the .5 value predicted by Mendelian inheritance. Several such tests have been proposed; the relationship between these is clarified by consideration of the likelihood for a disease-association model that is parametrized in terms of probabilities of disease, conditional on marker genotype, and of population-allele (or, more generally, haplotype) frequencies (Schaid 1996).

Received January 1, 1999; accepted for publication August 10, 1999; electronically published September 14, 1999.

Address for correspondence and reprints: Dr. David Clayton, MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Robinson Way, Cambridge CB2 2SR, United Kingdom. E-mail: david.clayton@mrc-bsu.cam.ac.uk

© 1999. The American Society of Human Genetics. All rights reserved.

0002-9297/1999/6504-0027$02.00

At the marker locus, the genotype g consists of a pair of haplotypes (i,j), and there is association between marker and disease when, at the population level, the probability of disease depends on the genotype. If this probability is represented by π_g, the genotype relative risk (GRR) is defined as $\phi_g = \pi_g/\pi_{(1,1)}$ (here, the genotype (1,1) has been taken as the reference genotype, so that $\phi_{(1,1)} = 1$). If there are H distinct haplotypes, then there are $G = H(H+1)/2$ distinct genotypes and, even for moderate H, many GRR parameters are required in order to model association. Thus, to obtain powerful tests, it may be necessary to consider a more restrictive model.

Falk and Rubinstein (1987) implicitly assumed a "genotype relative risk" model in which a single copy of the high-risk allele at a biallelic locus suffices to confer maximum risk on the genotype. This model has only a single association parameter and leads to a 1-df test statistic. However, the model does not easily generalize to multiallelic markers. Instead, most authors, following Terwilliger and Ott (1992), concentrate on a "haplotype-based" approach. This implicitly assumes the model $\phi_{(i,j)} = \theta_i\theta_j$, in which the two haplotypes act multiplicatively. The parameters θ_i may be regarded as haplotype relative risk (HRR) parameters (again, to ensure identifiability, it will be necessary to impose a constraint such as the "corner" constraint $\theta_1 = 0$). With this model, the (i,j) heterozygote genotype carries a relative risk equal to the geometric mean of the relative risks for the fully homozygote genotypes, $\phi_{(i,j)} = \sqrt{\phi_{(i,i)}\phi_{(j,j)}}$. The model represents marker-disease association with $H-1$ free parameters and therefore leads to association tests with $H-1$ df.

A full-probability model for the data must also describe the probability distribution of parental genotypes in the population. Again, unless simplifying assumptions are made, such a model can involve very many parameters. Accordingly, it is usual to assume no population stratification and Hardy-Weinberg equilibrium, so that the probability that a parent drawn from the population at random has genotype $g = (i,j)$ is

$$\Pr(g = (i,j)) = \begin{cases} \psi_i^2 & \text{if } i = j \\ 2\psi_i\psi_j & \text{otherwise} \end{cases}$$

and the two parents' genotypes are independent. The parameters ψ_i are the population haplotype relative frequencies and obey the constraint $\Sigma_i\psi_i = 1$. If the multiplicative model for HRRs also holds, it is easily shown that disease cases are also in Hardy-Weinberg equilibrium, with modified haplotype frequencies

$$\psi_i^* = \frac{\psi_i\theta_i}{\sum_j \psi_j\theta_j} \ . \tag{1}$$

It is generally more convenient to work with unbounded parameters, and, accordingly, henceforth the HRRs will be replaced by their logarithms, β_i, and the haplotype frequencies will be replaced by their multinomial logit transformations, γ_i; and $\theta_i = \exp\boldsymbol{\beta}_i$ and $\psi_i = \exp\boldsymbol{\gamma}_i/\Sigma_i\exp\boldsymbol{\gamma}_i$. With these transformations, the score functions are simply differences between observed and expected counts of haplotypes (see the Appendix).

The likelihood contribution of a parent-offspring trio ascertained via the affected offspring is the joint probability of parental and offspring genotypes, say PG and OG, conditional on disease in the offspring, say OD = 1; this, in turn, factorizes into two parts

$$\Pr(\mathrm{PG,OG}|\mathrm{OD}=1) = \Pr(\mathrm{PG}|\mathrm{OD}=1) \times\Pr(\mathrm{OG}|\mathrm{PG,OD}=1) \ . \tag{2}$$

The full-likelihood contribution and its two factors for the ith such trio will be denoted by $\mathrm{L}^{(\mathrm{F})}_{(i)}$, $\mathrm{L}^{(\mathrm{P})}_{(i)}$, and $\mathrm{L}^{(\mathrm{C})}_{(i)}$, respectively, so that, corresponding to equation(2),

$$\mathrm{L}^{(\mathrm{F})}_{(i)} = \mathrm{L}^{(\mathrm{P})}_{(i)}\mathrm{L}^{(\mathrm{C})}_{(i)} \ . \tag{3}$$

Detailed expressions for these likelihood contributions are given in the Appendix, but here it need only be noted that, whereas $\mathrm{L}^{(\mathrm{F})}_{(i)}$ and $\mathrm{L}^{(\mathrm{P})}_{(i)}$ depend on both sets of parameters, β and γ, the "conditional" likelihood contribution $\mathrm{L}^{(\mathrm{C})}_{(i)}$ depends only on the HRR parameters β. If the corresponding log-likelihood contributions are denoted by ℓ, the log likelihood decomposes additively: $\ell^{(\mathrm{F})}_{(i)} = \ell^{(\mathrm{P})}_{(i)} + \ell^{(\mathrm{C})}_{(i)}$. After summation over families, the total log likelihood decomposes in the same way (total log likelihoods will be indicated by omission of the subscript i).

This decomposition is central to what follows. Tests for no association ($H_0 : \boldsymbol{\beta} = 0$) can be constructed by use of either the full log likelihood, $\ell^{(\mathrm{F})}$, or the conditional log likelihood, $\ell^{(\mathrm{C})}$. The former extracts more information, but at a price; the extra term included, $\ell^{(\mathrm{P})}$, is the one that depends on the population model for parental genotypes, and the integrity of the added information depends strongly on correct specification of this model. Thus, $\ell^{(\mathrm{P})}$ depends on β, because the conditioning on presence of disease in offspring may induce deviation from Hardy-Weinberg equilibrium and from independence of the two parental genotypes. Since the model assumes that neither of these exists in the population, evidence of such deviations for the parents within the study constitutes evidence for disease-marker association. In contrast, the conditional-likelihood term $\ell^{(\mathrm{C})}$ does not depend on assumptions of the model for parental genotypes. This informal argument would suggest that methods based on the full log likelihood may be more powerful than those based on the conditional likelihood (Terwilliger and Ott 1992) but that they may give incorrect answers if the assumptions of the population model are not met (Spielman et al. 1993). For the latter reason, conditional-likelihood methods are usually preferred.

Tests of association can be constructed in two ways, from each of these likelihoods:

1. the log likelihood–ratio test, comparing twice the log likelihood at the global maximum-likelihood estimate, $\ell(\boldsymbol{\beta},\hat{\gamma})$, with the maximized likelihood under the null hypothesis, $\ell(0,\hat{\hat{\gamma}})$: $2[\ell(\hat{\boldsymbol{\beta}},\hat{\gamma}) - \ell(0,\hat{\hat{\gamma}})]$;
2. the score test based on the first-derivative vector of ℓ evaluated at $(0,\hat{\hat{\gamma}})$.

Both tests simplify if the conditional likelihood is used, since γ is not involved. The two testing strategies are asymptotically equivalent, leading to χ^2 tests on $H-1$ df. Table 1 sets out some tests previously proposed, classified by the testing strategy and likelihood on which they are based.

The approach proposed here for situations in which the transmission pattern may be uncertain is based on the score test, and this section concludes with some further notation for this approach. The score vector, denoted by **u**, is the vector of first derivatives of the log likelihood with respect to the parameters. This can be partitioned into two parts corresponding to the two sets of parameters:

$$\mathbf{u} = \begin{pmatrix} \mathbf{u}_\beta \\ \mathbf{u}_\gamma \end{pmatrix} = \begin{pmatrix} \dfrac{\partial \ell}{\partial \boldsymbol{\beta}} \\ \dfrac{\partial \ell}{\partial \boldsymbol{\gamma}} \end{pmatrix} .$$

The score vector is of length $2H$, with the first H elements, $\mathbf{u}_\beta$, concerning the HRR parameters β and with the next H elements, $\mathbf{u}_\beta$, concerning the logit-transformed gene frequencies γ. The information matrix, **J**, is of size $2H \times 2H$ and can be partitioned similarly:

$$\mathbf{J} = \begin{pmatrix} \mathbf{J}_{\beta\beta} & \mathbf{J}_{\beta\gamma} \\ \mathbf{J}_{\gamma\beta} & \mathbf{J}_{\gamma\gamma} \end{pmatrix} = -\begin{pmatrix} \dfrac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} & \dfrac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}^{\mathrm{T}}} \\ \dfrac{\partial^2 \ell}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\beta}^{\mathrm{T}}} & \dfrac{\partial^2 \ell}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^{\mathrm{T}}} \end{pmatrix}$$

$$= -\begin{pmatrix} \dfrac{\partial \mathbf{u}_\beta}{\partial \boldsymbol{\beta}^{\mathrm{T}}} & \dfrac{\partial \mathbf{u}_\beta}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} \\ \dfrac{\partial u_\gamma}{\partial \boldsymbol{\beta}^{\mathrm{T}}} & \dfrac{\partial u_\gamma}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} \end{pmatrix} .$$

Standard likelihood theory shows that, when evaluated at the correct parameter values, the variance-covariance matrix of **u** is given by the expected value of **J** (since, in the cases considered here, **J** does not depend on random variables, a distinction between the observed and expected values of **J** will not be made hereafter). Each set of parameters will require a linear constraint for identifiability (e.g., one β and one γ might be set to zero), leaving $2(H-1)$ free parameters, so that the rank of **J** will also be $2(H-1)$.

In the same way as the log likelihood, ℓ, the values of **u** and **J** can be decomposed into contributions of parents and of transmission from parents to offspring:

$$\begin{aligned} \mathbf{u}^{(\mathrm{F})} &= \mathbf{u}^{(\mathrm{P})} + \mathbf{u}^{(\mathrm{C})} , \\ \mathbf{J}^{(\mathrm{F})} &= \mathbf{J}^{(\mathrm{P})} + \mathbf{J}^{(\mathrm{C})} . \end{aligned}$$

Similarly, these arrays are sums of contributions for each parent-offspring trio, which will be denoted by "$\mathbf{u}_{(i)}$" and "$\mathbf{J}_{(i)}$." The full expressions for these contributions are given in the Appendix.

After the score vector has been calculated at the null hypothesis, $\boldsymbol{\beta} = 0$, the score test is given by

$$\mathbf{u}_\beta^{\mathrm{T}} \mathbf{V}_{\beta\beta}^{\ominus} \mathbf{u}_\beta , \qquad (4)$$

where $\mathbf{V}_{\beta\beta}^{\ominus}$ is a generalized inverse of the variance-covariance matrix of $\mathbf{u}_\beta$. Asymptotically, this is distributed as χ^2 with df equal to the rank of $\mathbf{V}_{\beta\beta}$ (usually $H-1$). Single-df tests for excess transmission of specific hap-

Table 1

Tests for Linkage/Gametic-Phase Equilibrium, Classified by the Likelihood on Which They Are Based and on Whether They Are Score or Likelihood-Ratio Test

Likelihood	Type	Test (Source)
$\ell^{(\mathrm{F})}$	Score	HRR (Falk and Rubinstein 1987)[a]
		Marginal homogeneity test (Bickeböller and Clerget-Darpoux 1995)
		Haplotype-based HRR, independence test (Terwilliger and Ott 1992)[a]
	Likelihood ratio	Likelihood-ratio test of Morris et al. (1997)[a]
$\ell^{(\mathrm{C})}$	Score	TDT test (Spielman et al. 1993)[a]
		Haplotype-based HRR, McNemar test (Terwilliger and Ott 1992)[a]
		Stuart-Maxwell test (Stuart 1955; Maxwell 1970; Bickeböller and Clerget-Darpoux 1995)
	Likelihood ratio	Extended TDT test (Sham and Curtis 1995)

[a] Test is for biallelic markers.

lotypes may be constructed by dividing the square of the appropriate element of $\mathbf{u}_\beta$ by its variance. Clayton and Jones (1999) have pointed out that both of these testing strategies lack power for large H and have suggested an alternative test based on a quadratic form in the score vector, using a haplotype "similarity" matrix. This approach could be extended to the case in which haplotype assignments are uncertain, along the lines suggested below; but there are a number of technical problems, and this will not be attempted here.

If the test is based on the conditional argument—that is, if $\mathbf{u}^{(C)}$ is used—then $\mathbf{V}_{\beta\beta}$ is given by the corresponding information matrix, $\mathbf{J}^{(C)}_{\beta\beta}$. However, if the full likelihood is used, a correction must be made for the fact that γ must be estimated. The estimated value of γ for fixed β, say $\hat{\hat{\gamma}}$, is obtained by solving the estimating equation $\mathbf{u}_\gamma = 0$. A linear approximation leads to

$$(\mathbf{u}_\beta)_{\gamma=\hat{\hat{\gamma}}} \approx \mathbf{u}_\beta - \mathbf{J}_{\beta\gamma}\mathbf{J}^{\ominus}_{\gamma\gamma}\mathbf{u}_\gamma \ , \qquad (5)$$

where the score and information component on the right-hand side are evaluated at the true values of β and γ. This, in turn, leads to the approximation

$$\tilde{\mathbf{V}}_{\beta\beta} \approx \mathbf{J}_{\beta\beta} - \mathbf{J}_{\beta\gamma}\mathbf{J}^{\ominus}_{\gamma\gamma}\mathbf{J}^{\mathrm{T}}_{\beta\gamma} \qquad (6)$$

for the variance of $(\mathbf{u}_\beta)_{\gamma=\hat{\hat{\gamma}}}$. To perform the score test in practice, $\mathbf{J}$ is evaluated at ($\boldsymbol{\beta} = 0$,$\boldsymbol{\gamma} = \hat{\hat{\boldsymbol{\gamma}}}$).

Incomplete Data

Likelihood methods generalize naturally to situations in which data are incomplete. Well-known results (Dempster et al. 1977; Little and Rubin 1987) give the score and information for incomplete data in terms of moments of the corresponding functions for the complete data, taken over the "posterior" distribution of the complete data, given all the available data. Thus, the data for the ith trio may be consistent with a set of possible genotypes and transmissions, each giving a different likelihood contribution. If these are denoted by $\mathrm{L}_{(j)}, j \in \mathcal{P}$, the log-likelihood contribution of such a trio is $\ell_{\mathcal{P}} = \log \Sigma_{j\in\mathcal{P}} \mathrm{L}_{(j)}$. If the score and information contribution corresponding to the complete data likelihoods $L_{(j)}$ are denoted by $\mathbf{u}_{(j)}$ and $\mathbf{J}_{(j)}$, the score contribution for a trio with incomplete data is a weighted mean of the possible complete-data contributions consistent with the observed data:

$$\mathbf{u}_{\mathcal{P}} = \frac{\sum_{j\in\mathcal{P}} \mathrm{L}_{(j)}\mathbf{u}_{(j)}}{\sum_{j\in\mathcal{P}} \mathrm{L}_{(j)}} \ . \qquad (7)$$

The variance-covariance matrix of $\mathbf{u}_{\mathcal{P}}$ is obtained from the corresponding information contribution:

$$\mathbf{V}_{\mathcal{P}} = \mathbf{J}_{\mathcal{P}} = \frac{\sum_{j\in\mathcal{P}} \mathrm{L}_{(j)}\mathbf{J}_{(j)}}{\sum_{j\in\mathcal{P}} \mathrm{L}_{(j)}} - \left\{ \frac{\sum_{j\in\mathcal{P}} \mathrm{L}_{(j)}\mathbf{u}_{(j)}\mathbf{u}^{\mathrm{T}}_{(j)}}{\sum_{j\in\mathcal{P}} \mathrm{L}_{(j)}} - \mathbf{u}_{\mathcal{P}}\mathbf{u}^{\mathrm{T}}_{\mathcal{P}} \right\} \ . \qquad (8)$$

These expressions could be used to compute the score vector and information matrix for incomplete data, so that, with use of expression (4) and equation (6), a score test for association could be calculated.

This approach provides a solution to the problem of uncertain-haplotype transmission, but it has a major disadvantage. Equations (7) and (8) require that $L_{(j)}$, $\mathbf{u}_{(j)}$, and $\mathbf{J}_{(j)}$ refer to the *full* likelihood, so that they should more correctly be written as $\mathrm{L}^{(F)}_{(j)}$, $\mathbf{u}^{(F)}_{(j)}$, and $\mathbf{J}^{(F)}_{(j)}$, respectively. Because of the summation over possible genotypes consistent with the available data, the score contributions from each trio no longer decompose into a parental contribution $\mathbf{u}^{(P)}_{\mathcal{P}}$ and a conditional-likelihood contribution $\mathbf{u}^{(C)}_{\mathcal{P}}$. Thus, this approach would not seem to lead to a test that is robust against deviation from the assumptions of the model for the distribution of parental genotypes.

A Partial-Likelihood Argument

In this section, a compromise between the full-likelihood approach and the conditional-likelihood approach is proposed. For complete data in which both parental genotypes and haplotype transmission to offspring are observed with certainty, this uses the parental term $L^{(P)}$, defined by equations (2) and (3), for inference concerning γ but reverts to the conditional part of the likelihood, $L^{(C)}$, for inference concerning β. Thus, the contribution of the ith such trio to the partial score function is

$$\mathbf{u}^{(*)}_{(i)} = \begin{pmatrix} \mathbf{u}^{(C)}_{\beta(i)} \\ \\ \mathbf{u}^{(P)}_{\gamma(i)} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial \ell^{(C)}_{(i)}}{\partial \boldsymbol{\beta}} \\ \dfrac{\partial \ell^{(P)}_{(i)}}{\partial \boldsymbol{\gamma}} \end{pmatrix} \ .$$

Because they are based on factorization (3), the parental and conditional score contributions $\mathbf{u}^{(P)}_{(i)}$ and $\mathbf{u}^{(C)}_{(i)}$ are independent, and the variance-covariance matrix of $\mathbf{u}^{(*)}_{(i)}$ is

$$\mathbf{V}^{(*)}_{(i)} = \begin{pmatrix} \mathbf{J}^{(C)}_{\beta\beta(i)} & 0 \\ 0 & \mathbf{J}^{(P)}_{\gamma\gamma(i)} \end{pmatrix} \ . \qquad (9)$$

Note that, because $\mathbf{u}^{(*)}_{(i)}$ are no longer true score functions, $\mathbf{V}^{(*)}_{(i)}$ is no longer equal to the matrix of derivatives of $\mathbf{u}^{(*)}_{(i)}$, which is

$$\mathbf{J}^{(*)}_{(i)} = \begin{pmatrix} \mathbf{J}^{(C)}_{\boldsymbol{\beta\beta}(i)} & 0 \\ \mathbf{J}^{(P)}_{\boldsymbol{\gamma\beta}(i)} & \mathbf{J}^{(P)}_{\boldsymbol{\gamma\gamma}(i)} \end{pmatrix} \ . \tag{10}$$

When the data are incomplete, consistent with a set of complete data scenarios $j \in \mathcal{P}$, the proposed partial-score function is a weighted mean of the complete-data partial-score functions over $\mathcal{P}$, using the full-likelihood contributions $\mathrm{L}^{(F)}_{(j)}$ as weights; that is,

$$\mathbf{u}^{(*)}_{\mathcal{P}} = \frac{\sum\limits_{j\in\mathcal{P}} \mathrm{L}^{(F)}_{(j)}\mathbf{u}^{(*)}_{(j)}}{\sum\limits_{j\in\mathcal{P}} \mathrm{L}^{(F)}_{(j)}} \ . \tag{11}$$

This approach does not entirely do away with the need for a model for the distribution of parental genotypes, but this model contributes only to the weights given to the complete-data-score contributions for β in the mean score and not to these contributions themselves. Intuitively, this strategy would be expected to be much more robust against departures from the population model than would a full-likelihood approach. The variance of $\mathbf{u}^{(*)}_{\mathcal{P}}$ may be estimated by an expression similar to that given in equation (8):

$$\mathbf{V}^{(*)}_{\mathcal{P}} = \frac{\sum\limits_{j\in\mathcal{P}} \mathrm{L}^{(F)}_{(j)}\mathbf{V}^{(*)}_{(j)}}{\sum\limits_{j\in\mathcal{P}} \mathrm{L}^{(F)}_{(j)}} - \left\{ \frac{\sum\limits_{j\in\mathcal{P}} \mathrm{L}^{(F)}_{(j)}\mathbf{u}^{(*)}_{(j)}(\mathbf{u}^{(*)}_{(j)})^{\mathrm{T}}}{\sum\limits_{j\in\mathcal{P}} \mathrm{L}^{(F)}_{(j)}} - \mathbf{u}^{(*)}_{\mathcal{P}}(\mathbf{u}^{(*)}_{\mathcal{P}})^{\mathrm{T}} \right\} \ , \tag{12}$$

and the matrix of derivatives of $\mathbf{u}^{(*)}_{\mathcal{P}}$ is given by

$$\mathbf{J}^{(*)}_{\mathcal{P}} = \frac{\partial \mathbf{u}^{(*)}_{\mathcal{P}}}{\partial(\boldsymbol{\beta}^{\mathrm{T}},\boldsymbol{\gamma}^{\mathrm{T}})} = \frac{\sum\limits_{j\in\mathcal{P}} \mathrm{L}^{(F)}_{(j)}\mathbf{J}^{(*)}_{(j)}}{\sum\limits_{j\in\mathcal{P}} \mathrm{L}^{(F)}_{(j)}} - \left\{ \frac{\sum\limits_{j\in\mathcal{P}} \mathrm{L}^{(F)}_{(j)}\mathbf{u}^{(*)}_{(j)}(\mathbf{u}^{(F)}_{(j)})^{\mathrm{T}}}{\sum\limits_{j\in\mathcal{P}} \mathrm{L}^{(F)}_{(j)}} - \mathbf{u}^{(*)}_{\mathcal{P}}(\mathbf{u}^{(F)}_{\mathcal{P}})^{\mathrm{T}} \right\} \ . \tag{13}$$

The total partial-score vector, $\mathbf{u}^{(*)}$, is obtained by summation of such contributions over families, as is its variance $\mathbf{V}^{(*)}$ and the matrix of derivatives $\mathbf{J}^{(*)}$.

Although the complete-data-score contributions to $\mathbf{u}^{(*)}_{\beta}$ and $\mathbf{u}^{(*)}_{\gamma}$ are independent, as demonstrated by the block-diagonal structure of the contributions $\mathrm{V}^{(*)}$ seen in equation (9), the process of averaging over $j \in \mathcal{P}$ to obtain the incomplete-data-score contributions leads to the variance contributions given by equation (12), which are *not* block diagonal. It is then necessary to take account of the need to estimate γ when tests for β are constructed. As outlined in the earlier discussion of the standard-likelihood theory, this may be done by adopting the linear approximation to the estimating equations, leading to equation (5). Because the information and variance matrices no longer coincide, the variance estimate for $(\mathbf{u}^{(*)}_{\beta})_{\gamma=\hat{\hat{\gamma}}}$ is now rather more complicated than the expression given in equation (6):

$$\tilde{\mathbf{V}}^{(*)}_{\beta\beta} \approx \mathbf{V}^{(*)}_{\beta\beta} + \mathbf{J}^{(*)}_{\beta\gamma}(\mathbf{J}^{(*)}_{\gamma\gamma})^{\ominus}\mathbf{V}^{(*)}_{\gamma\gamma}(\mathbf{J}^{(*)}_{\gamma\gamma})^{\ominus\ \mathrm{T}}(\mathbf{J}^{(*)}_{\beta\gamma})^{\mathrm{T}} \\ -\mathbf{J}^{(*)}_{\beta\gamma}(\mathbf{J}^{(*)}_{\gamma\gamma})^{\ominus}\mathbf{V}^{(*)}_{\gamma\beta} - \mathbf{V}^{(*)}_{\beta\gamma}(\mathbf{J}^{(*)}_{\gamma\gamma})^{\ominus\ \mathrm{T}}(\mathbf{J}^{(*)}_{\beta\gamma})^{\mathrm{T}} \ . \tag{14}$$

The theory of the proposed partial-score function has been set out above in some generality and would allow consistent estimation of β as well as the testing of hypothetical values. However, some simplification is possible when the null hypothesis $\boldsymbol{\beta} = 0$ is tested. In this special case, we can embed the null hypothesis within the model for parental genotypes, so that these are assumed to be sampled at random from the population. The dependence of $\mathrm{L}^{(P)}_{(j)}$ and $\mathbf{u}^{(P)}_{\gamma(j)}$ on β may then be ignored, leading to a number of simplifications:

1. $\mathbf{J}^{(*)}_{(i)}$ is now block diagonal, given by the right-hand side of equation (9);
2. similarly, the derivative matrix in the uncertain-transmission case, $\mathbf{J}^{(*)}_{\mathcal{P}}$, is given by equation (12); and
3. equation (14) reverts to the simpler form of equation (6).

This simplification becomes essential when, as below, more-extended nuclear-family structures are considered. In such cases, the dependence of $L^{(P)}$ on β becomes more complicated and, particularly when multiple affected offspring are considered, may involve ascertainment corrections. In such circumstances, consistent estimation of β under uncertain-haplotype transmission may prove difficult, whereas testing the null hypothesis remains straightforward.

To perform a score test of $H_0 : \boldsymbol{\beta} = 0$, the score equations $\mathbf{u}^{(*)}_{\gamma} = 0$ are solved for the maximum-likelihood estimate $\hat{\hat{\gamma}}$, with $\boldsymbol{\beta} = 0$. The vector $\mathbf{u}^{(*)}_{\beta}$ is then used in expression (4), to obtain an asymptotic χ^2 statistic.

Additional Offspring

Uncertainty of parental genotype and transmission to the affected offspring, whether due to missing parental genotype or phase uncertainty, may be reduced by typing further offspring in the family. For the moment, assume that these are either unaffected or of unknown disease status. Although, strictly speaking, transmission to unaffected offspring provides some information about the association parameters β, since high-risk haplotypes will be *less* likely to be transmitted to unaffected offspring, the amount of information provided is negligible when the risks of disease conditional on genotype, π_g, are small. In these circumstances, unaffected offspring and offspring with unknown disease status can be treated in exactly the same way; such offspring are ignored in the conditional-likelihood contributions $\mathrm{L}^{(\mathrm{C})}_{(i)}$ and, consequently, in the $\mathbf{u}^{(*)}_{\beta}$. However, they are used when the set, $\mathcal{P}$, of genotypes consistent with the available data is calculated and when the likelihood weights, when score contributions over $\mathcal{P}$ are averaged, are calculated.

Additional *affected* offspring are more difficult to deal with, since these potentially contribute to $L^{(\mathrm{C})}$ and, therefore, to $\mathbf{u}^{(*)}_{\beta}$. However, if the null hypothesis allows for the presence of *linkage* between marker locus and disease-susceptibility locus, only specifying the gametic-phase equilibrium, then the contributions of multiple affected offspring to the family contribution to $L^{(\mathrm{C})}$ are not independent. In this situation, some authors have advocated use of only the first affected offspring, but this may lead to considerable loss of information. More-satisfactory approaches have been suggested by Martin et al. (1997) and by Horvath and Laird (1998). The approach proposed here can be extended to deal with this difficulty, by application of the general results, reported by Huber (1967), concerning "robust" variance estimation when a likelihood is misspecified. When the null hypothesis allows for linkage and more than one affected offspring may be included for each family, the expected value of $\mathbf{u}^{(*)}$ is still zero, but its variance is incorrectly estimated by the results given above. When the method of Huber is followed, an alternative estimate is provided by the empirical variance-covariance matrix of the contributions of each family to $\mathbf{u}^{(*)}$. Thus, if families 1,…,N are consistent with genotype configurations $\mathcal{P}_1, \dots, \mathcal{P}_N$, then the overall score is $\mathbf{u}^{(*)} = \Sigma_{i=1}^{N}\mathbf{u}^{(*)}_{\mathcal{P}_i}$, and its variance can be estimated by

$$\mathbf{V}^{(*)} = \sum_{i=1}^{N} \mathbf{u}^{(*)}_{\mathcal{P}_i}(\mathbf{u}^{(*)}_{\mathcal{P}_i})^{\mathrm{T}} - \frac{1}{N}\mathbf{u}^{(*)}(\mathbf{u}^{(*)})^{\mathrm{T}} \ . \qquad (15)$$

After estimation of γ, a robust variance estimate for $\mathbf{u}^{(*)}_{\beta}$ can be obtained from equations (14) and (15).

Discussion

This report has set out the statistical theory behind the proposed new approach to transmission/disequilibrium testing. Numerical results concerning its performance in the situation in which one parent is unavailable have been provided by A. Cervino, A. Hill, and P. Donnelly (personal communication), who demonstrate that the method is indeed highly robust against violation of the assumption of no population stratification—at least in the situations that they considered. The method has a number of advantages over other approaches that have been proposed. First, it is efficient, making full use of whatever parental data are available. Second, by use of the robust variance estimator (15), more than one affected offspring per family may be used, even in the presence of linkage. Third, this would seem to be the only approach so far proposed that will deal with the problem of phase uncertainty for multilocus haplotypes. This will be an important problem as attention turns to SNP markers within candidate genes. (A computer program, "TRANSMIT," which implements the methods described in this report, is available from the author [Medical Research Council Biostatistics Unit].)

Appendix A

Likelihood, Score, and Information Contributions

Consider a marker with m alleles or possible haplotypes and a family (parent-offspring trio), f, in which the parental genotypes are (p,q) and (r,s) and the affected offspring is (p,r). In terms of the original gene frequency and HRR parameters, the full-likelihood contribution of the family is

$$\begin{aligned}\mathrm{L}^{(\mathrm{F})}_{(f)} &= \Pr(\mathrm{PG,OG}|\mathrm{OD} = 1) \\ &= \frac{\psi_p\psi_q\psi_r\psi_s\theta_p\theta_r}{\sum_i\sum_j\sum_k\sum_\ell \psi_i\psi_j\psi_k\psi_\ell(\theta_i\theta_k + \theta_i\theta_\ell + \theta_j\theta_k + \theta_j\theta_\ell)} \\ &= \frac{\psi_q}{\sum_i \psi_i}\frac{\psi_s}{\sum_i \psi_i}\frac{\psi_p\theta_p}{\sum_i \psi_i\theta_i}\frac{\psi_r\theta_r}{\sum_i \psi_i\theta_i} \\ &= \psi_q\psi_s\psi^*_p\psi^*_r \ ,\end{aligned}$$

where all the summations are over the range 1,…,m and ψ^*_i, the relative frequency of haplotype i in cases, is given by equation (1). The final simplified form makes it clear that this likelihood is equivalent to the suggestion, by Terwilliger and Ott (1992), that haplotypes p and r may be regarded as "case" haplotypes and that the untransmitted haplotypes q and s may be regarded as "controls." The two factors in the factorization (3) of this contribution are

$$\begin{aligned}\mathrm{L}^{(\mathrm{P})}_{(f)} &= \Pr(\mathrm{PG}|\mathrm{OD} = 1) \\ &= \psi_p\psi_q\psi_r\psi_s(\theta_p + \theta_q)(\theta_r + \theta_s)/\Big(\sum_i \psi_i\theta_i\Big)^2 \ ; \\ \mathrm{L}^{(\mathrm{C})}_{(f)} &= \Pr(\mathrm{OG}|\mathrm{PG,OD} = 1) \\ &= \frac{\theta_p}{(\theta_p + \theta_q)}\frac{\theta_r}{(\theta_r + \theta_s)} \ .\end{aligned}$$

Thus the conditional-likelihood contribution involves only the HRR parameters, whereas the parental contribution involves both sets of parameters.

All the score functions may be regarded as observed counts minus expected counts. When N_i is used to denote the number of times that haplotype i occurs in the two parents, the score contributions with respect to γ_i are

$$\frac{\partial \ell^{(F)}_{(f)}}{\partial \boldsymbol{\gamma}_i} = \frac{\partial \ell^{(P)}_{(f)}}{\partial \boldsymbol{\gamma}_i} = N_i - 2\psi_i - 2\psi_i^* \ . \qquad (A1)$$

When T_i is used to denote the number of times that haplotype i is transmitted to the affected offspring, the score contribution for β_i, based on the full likelihood, is $\partial \ell^{(F)}_{(f)}/\partial \boldsymbol{\beta}_i = T_i - 2\psi_i^*$. The corresponding score contribution based on the conditional likelihood has a similar observed-minus-expected form, but the expected frequencies are now based on transmission probabilities:

$$\frac{\partial \ell^{(C)}_{(f)}}{\partial \boldsymbol{\beta}_i} = T_i - E_i \ ;$$

$$E_i = (\Delta_{ip}\theta_p + \Delta_{iq}\theta_q)/(\theta_p + \theta_q), \\ +(\Delta_{ir}\theta_r + \Delta_{is}\theta_s)/(\theta_r + \theta_s) \ , \qquad (A2)$$

where Δ_{ij} is the Kronecker delta, taking the value 1 if $i = j$ and 0 otherwise. It follows that the score contribution for β_i based on the parental likelihood is $\partial \ell^{(P)}_{(f)}/\partial \boldsymbol{\beta}_i = E_i - 2\psi_i^*$. It is clear from this expression why this term is so dependent on the assumptions of no population stratification and Hardy-Weinberg equilibrium.

The important score functions are equations (A1) and (A2), and, under the null hypothesis, these take the values

$$\frac{\partial \ell^{(F)}_{(f)}}{\partial \boldsymbol{\gamma}_i} = \frac{\partial \ell^{(P)}_{(f)}}{\partial \boldsymbol{\gamma}_i} = N_i - 4\psi_i \ ;$$

$$\frac{\partial \ell^{(C)}_{(f)}}{\partial \boldsymbol{\beta}_i} = T_i - \frac{N_i}{2} \ .$$

The corresponding information contributions are obtained by differentiation of equations (A1) from equations (A2). At the null hypothesis, after it is noted that $\Sigma_k \psi_k = 1$,

$$\frac{\partial^2 \ell^{(F)}_{(f)}}{\partial \boldsymbol{\gamma}_i \partial \boldsymbol{\gamma}_j} = \frac{\partial^2 \ell^{(P)}_{(f)}}{\partial \boldsymbol{\gamma}_i \partial \boldsymbol{\gamma}_j} = 4(\Delta_{ij}\psi_i - \psi_i\psi_j) \ ;$$

$$\frac{\partial^2 \ell^{(F)}_{(f)}}{\partial \boldsymbol{\gamma}_i \partial \boldsymbol{\beta}_j} = \frac{\partial^2 \ell^{(P)}_{(f)}}{\partial \boldsymbol{\gamma}_i \partial \boldsymbol{\beta}_j} = 2(\Delta_{ij}\psi_i - \psi_i\psi_j) \ .$$

Only a small number of the elements of the matrix $[\partial^2 \ell^{(C)}/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T]$ (again evaluated at the null hypothesis) need to be updated. If $p \neq q$, the diagonal elements (p,p) and (q,q) must be incremented by $+\frac{1}{4}$, and the off-diagonal elements (p,q) and (q,p) must be incremented by $-\frac{1}{4}$. A similar procedure is used for r,s.

Electronic-Database Information

The URL for data in this article is as follows:

Medical Research Council Biostatistics Unit, http://www.mrc-bsu.cam.ac.uk (for TRANSMIT program)

References

Bickeböller H, Clerget-Darpoux F (1995) Statistical properties of the allelic and genotypic transmission disequilibrium test for multiallelic markers. Genet Epidemiol 12:865–870

Boehnke M, Langefeld CD (1998) Genetic association mapping based on discordant sib pairs: the discordant-alleles test. Am J Hum Genet 62:950–961

Clayton D, Jones H (1999) Transmission/disequilibrium tests for extended marker haplotypes. Am J Hum Genet 65: 1161–1169 (in this issue)

Curtis D (1997) Use of siblings as controls in case-control association studies. Ann Hum Genet 61:319–333

Curtis D, Sham PC (1995) A note on the application of the transmission disequilibrium test when a parent is missing. Am J Hum Genet 56:811–812

Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Society B 39:1–22

Falk CT, 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

Horvath S, Laird NM (1998) A discordant-sibship test for disequilibrium and linkage: no need for parental data. Am J Hum Genet 63:1886–1897

Huber P (1967) The behaviour of maximum likelihood estimates under non-standard conditions. In: Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability. Vol 1. University of California Press, Berkeley, pp 221–233

Little R, Rubin D (1987) Statistical analysis with missing data. John Wiley & Sons, New York

Martin ER, Kaplan NL, Weir BS (1997) Tests for linage and association in nuclear families. Am J Hum Genet 61:439–448

Maxwell A (1970) Comparing the classification of subjects by two independent judges. Br J Psychiatry 116:651–655

Monks SA, Kaplan NL, Weir BS (1998) A comparative study of sibship tests of linkage and/or association. Am J Hum Genet 63:1507–1516

Morris A, Curnow R, Whittaker J (1997) A likelihood ratio test for detecting patterns of disease-marker association. Ann Hum Genet 61:335–350

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

Schaid DJ, Rowland C (1998) Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. Am J Hum Genet 63:1492–1506

Sham P, Curtis D (1995) An extended transmission/disequilibrium test for multi-allelic marker loci. Ann Hum Genet 59: 323–336

Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450–458

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus. Am J Hum Genet 52: 506–516

Stuart A (1955) A test for the homogeneity of the marginal distributions in a two-way classification. Biometrika 32: 412–415

Terwilliger J, Ott J (1992) A haplotype based 'haplotype relative risk' approach to detecting allelic associations. Hum Hered 42:337–346