European Journal of Human Genetics (2004) 12, 964–970

www.nature.com/ejhg

ARTICLE

Combining the transmission disequilibrium test and case–control methodology using generalized logistic regression

Nico JD Nagelkerke*[,1,2], Barbara Hoebee[3], Peter Teunis[1] and Tjeerd G Kimman[4]

[1]*Computerization and Methodological Consultancy Unit, National Institute of Public Health and the Environment (RIVM), Bilthoven, The Netherlands;* [2]*Department of Medical Statistics, Leiden University Medical Centre, Leiden, The Netherlands;* [3]*Laboratory of Toxicology, Pathology, and Genetics, National Institute of Public Health and the Environment (RIVM), Bilthoven, The Netherlands;* [4]*Laboratory for Vaccine-Preventable Diseases, National Institute of Public Health and the Environment (RIVM), Bilthoven, The Netherlands*

To study the role of genetic factors in the etiology, susceptibility, or severity of disease, several methods are available. In a transmission disequilibrium test, genotypes of cases are compared to those of their parents to explore whether a specific allele, or marker, at a locus of interest appears to be transmitted in excess of what is expected on the basis of Mendelian inheritance. Such apparent excess transmission indicates that cases are being selected for that allele, thereby providing evidence that this allele is a risk factor for disease. In case–control studies, genotypes of cases are compared to those of controls from the same population to identify whether a specific allele is associated with disease. If so, either the allele at this locus or one in linkage disequilibrium with it may be causally related to the etiology of the disease. Here, we discuss the problem of combining a transmission disequilibrium test and a case–control comparison, in order to integrate all available information, and thereby increase statistical power. As the same cases are used in both approaches, the two results are not independent. However, parents of cases can be independently compared to controls. Both the issue of testing for a genetic effect and the estimation of relative risks under the multiplicative model using generalized logistic regression are discussed.
European Journal of Human Genetics (2004) **12**, 964–970. doi:10.1038/sj.ejhg.5201255
Published online 1 September 2004

Keywords: genetic epidemiology; transmission disequilibrium test; case–control studies

Introduction

Suppose that a specific disease or medical condition is under the genetic control of an autosomal locus. This locus may be known or suspected (ie a candidate gene),[1,2] for example, because its gene products are involved in the immune response to the disease. Alternatively, the causative locus may be in close linkage disequilibrium with another, known, locus (a 'marker'). A specific allele at the marker locus may thus appear to be a risk factor for disease, because it forms part of a haplotype that contains alleles at other loci that are causally related to the disease; essentially a form of confounding. The association between marker and disease may help to identify the causally related haplotype or locus (to 'map' it). In order to explore the magnitude of this genetic control, two genetic methodological approaches are commonly used: the transmission disequilibrium test (TDT) and case–control (CC) studies.

For the TDT we observe genotypes of *incident* cases of the condition, and in addition those of their parents or other close relatives.[3–9] In its original form, triplets composed of

*Correspondence: NJD Nagelkerke, National Institute of Public Health and the Environment (RIVM), PO Box 1, 3720 BA Bilthoven, The Netherlands. Tel: +31 30 2743690; Fax +31 30 2744456;
E-mail: nico.nagelkerke@rivm.nl
Received 11 November 2003; revised 7 June 2004; accepted 20 June 2004

a case and his/her two parents are observed, and it is tested whether a specific allele at a locus of interest is transmitted in excess of what is expected on the basis of Mendelian inheritance. If so, this would constitute evidence that the sample of cases is enriched for this allele due to an enhanced propensity to becoming a case. For the case–control methodology, one similarly identifies incident cases, and in addition one samples controls from the same background population as cases.[10,11] Again, enhanced susceptibility would increase the prevalence of the suspected allele in the sample of cases. A comparison of these two methodological approaches (CC and TDT) has recently been published.[12] This comparison also provides an excellent overview of the assumptions underlying each approach.

Statistical power of both the TDT and CC approach depends on sample sizes and the magnitude of the genetic effect. Often, finding a sufficient number of triplets is a cause for concern. Combining the two approaches (TDT and CC), and thereby increasing statistical power, is possible, either when:

a. In addition to the triplets required for the TDT, genetic information from suitable controls is available, or
b. in addition to a case–control sample, genotypic information on some, but not all, of the cases' parents are collected. This may be because of inability to obtain this information, or because the study was first designed as a CC study and a TDT component was later added to confirm associations detected in this study. Thus, in addition to triplets and controls, genotypes of affected individuals ('founder cases') without information on parental genotypes are available.

The former situation is expected to occur when a TDT study is carried out first and controls are subsequently collected in order to increase statistical power. The latter situation would normally arise when a CC study is carried out first and a TDT study is carried out subsequently to corroborate CC findings. As the TDT cases can be used for both the TDT and for comparison with the controls from the CC study, results from these methods are not statistically independent.

We aim to develop a simple overall estimator of the impact of the gene on the condition; that is, an estimator of the relative risk experienced by subjects carrying the allele of interest. This estimate and its standard error also provide a test for association, that is a test for whether the condition may indeed be under the control of the locus of interest, taking this dependence into account.

Statistical methods

We assume a biallelic polymorphism. One allele is the presumed risk allele, denoted with a '2', that is suspected to be associated with a higher disease incidence than the normal, reference allele, denoted with a '1'. Individuals who are homozygous for the susceptibility allele are denoted by 2/2, etc.

We denote the relative risk of disease (relative to homozygotes of the normal allele, ie 1/1) of individuals with one copy of the susceptibility allele (ie 1/2) by γ_1 ($\geqslant 1$), and the relative risk of individuals having two copies of the susceptibility allele (ie 2/2) by γ_2. We want to estimate γ_1 and γ_2 and test the null hypothesis that $\gamma_1 = \gamma_2 = 1$. Useful 'penetrance models' for the allele, with only one γ parameter, are:

(i) $\gamma_2 = \gamma_1 = \gamma > 1$;
(ii) $\gamma_1 = 1$ and $\gamma_2 = \gamma > 1$;
(iii) $\gamma_2 = \gamma_1 * \gamma_1 = \gamma * \gamma$

These correspond to a dominant, recessive, and multiplicative, effect of the susceptibility allele, respectively.[13]

Maximum likelihood

A possible method of analysis seems to be maximum likelihood. Assume that controls are drawn from the same population P from which affected children and possibly founder cases are recruited. We will denote the population frequency of the '2' allele in P by p. The three genotypes 1/1, 1/2 and 2/2 occur with probability $(1-p)^2$, $2p(1-p)$, p^2 respectively, assuming Hardy–Weinberg equilibrium (HWE). We assume the absence of parent-of-origin effects. That is, we assume alleles inherited from the mother to have the same effect as alleles inherited from the father.[14–16] When such effects are suspected, TDT and CC studies should neither be compared nor combined, as parents-of-origin effects cannot be inferred from CC studies.

Using Bayes' theorem and standard probability calculus, we can derive expressions for the probabilities of specific genotypes for the cases and their parents, conditional on the child being a case (Table 1).

Table 1 Probabilities of mating types and genotypes of cases

Mating type	*Offspring genotype*	*Probability*	*Number observed*
1/1, 1/1	1/1	$(1-p)^4/T$	n_1
1/1, 1/2	1/1	$2(1-p)^3p/T$	n_2
1/1, 1/2	1/2	$2(1-p)^3p\gamma_1/T$	n_3
1/1, 2/2	1/2	$2(1-p)^2p^2\gamma_1/T$	n_4
1/2, 1/2	1/1	$(1-p)^2p^2/T$	n_5
1/2, 1/2	1/2	$2(1-p)^2p^2\gamma_1/T$	n_6
1/2,1/2	2/2	$(1-p)^2p^2\gamma_2/T$	n_7
1/2, 2/2	1/2	$2(1-p)p^3\gamma_1/T$	n_8
1/2, 2/2	2/2	$2(1-p)p^3\gamma_2/T$	n_9
2/2, 2/2	2/2	$p^4\gamma_2/T$	n_{10}

T is chosen such that the probabilities add to 1, ie $T=(1-p)^2+2p(1-p)\gamma_1+p^2\gamma_2$.

As controls and founder cases are sampled independently of triplets, and the genotypes of controls are not influenced by γ_1 or γ_2, the likelihood of p, γ_1, γ_2 is

$$\begin{aligned}&\Pi \Pr(\text{genotypes of triplets}|\text{offspring is affected cases}; p, \gamma_1, \gamma_2)\\&\quad\times\Pi\Pr(\text{genotypes of controls}|\text{p})\\&\quad\times\Pi \Pr(\text{founder cases}|\text{p}, \gamma 1, \gamma 2)\}\end{aligned} \tag{1}$$

that is, the product of three multinomial likelihoods, and p, γ_1, γ_2 can all be estimated by maximizing this likelihood, under the assumption of HWE.

Let n_i $(i=1,\ldots,10)$ denote the frequencies of the 10 possible mating type/genotype of child combinations given in Table 1. Analogously, let m_1, m_2, m_3, k_1, k_2, and k_3 denote the observed numbers of unrelated controls and founder cases with marker genotype 1/1, 1/2, 2/2, respectively.

Then, it can be shown that the maximum likelihood estimators (MLEs) are

$$\hat{p}=\frac{m_2+2m_3+n_2+n_4+2n_5+n_6+2n_8+n_9+2n_{10}}{2(m_1+m_2+m_3+n_1+n_1+n_2+n_3+n_4+n_5+n_6+n_7+n_8+n_9+n_{10})}$$

$$\hat{\gamma}_1=\frac{(k_2+n_3+n_4+n_6+n_8)(2m_1+m_2+2n_1+n_2+2n_3+n_4+n_6+2n_7+n_9)}{2(k_1+n_1+n_2+n_5)(m_2+2m_3+n_2+n_4+2n_5+n_6+2n_8+n_9+2n_{10})}$$

$$\hat{\gamma}_2=\frac{(k_3+n_7+n_9+n_{10})(2m_1+m_2+2n_1+n_2+2n_3+n_4+n_6+2n_7+n_9)^2}{(k_1+n_1+n_2+n_5)(m_2+2m_3+n_2+n_4+2n_5+n_6+2n_8+n_9+2n_{10})^2}$$

In case for some of the children information on one of their parents is missing, methods for missing data, such as the EM algorithm,[17] or multiple imputation, can be used.[18] For example, under the assumption that the missing parents are missing at random, missing parents can be multiply imputed by sampling from the group where both parents are present, and the child and other parent have the same genotype as the case under consideration. The use of missing data techniques that make use of all available information should typically be more efficient than treating the case as a founder case by discarding information from the present parent. The assumption of random missingness may not be true; however, when missingness of the parent depends on whether he/she is also affected.

Alternative factorization of the likelihood

Several problems may arise if the conditions for calculating the above MLE are not met. First, the population may not be in HWE. If so, the MLEs of p, γ_1, γ_2 may be biased. However, information on γ_1 and γ_2 can be obtained independent of the Hardy–Weinberg assumption by the TDT. The TDT does not require this assumption as it considers transmission of alleles *conditional* on the parents' genotypes, and thus does not use the same information as the MLE does. The difference is perhaps best illustrated by the fact that the MLE may be calculated from parents and affected children even if all parents are homozygous, whereas then the TDT could not be computed. This robustness with respect to the (HWE) assumptions argues in favour of the TDT, although obviously the TDT may be less efficient than the MLE.

Second, while multiple imputation (as described above) is possible in the context of MLE, it will unfortunately fail if only one of the parents is available (and thus one of the parents is missing) for all children. Versatile methods, for example the 1-TDT, to analyze such data have been developed, but these appear not to be likelihood-based.[19] As yet there appears to be no obvious way of combining the 1-TDT with information from population controls.

Third, a complication that arose in the context of the example presented in this paper is that when using the above MLE all individuals are required to be from the same population. If not, the concept of population allele frequency (P) may be futile. Also, stratification, that is, a mismatch between cases and controls, for example, due to the two groups containing different mixes of ethnic groups, should be avoided. In our example,[20] the available control group was entirely ethnically 'Dutch', whereas among the affected children some were of foreign or mixed descent with the HWE assumption almost certainly not true. In the likelihood approach, inclusion of 'foreign' (ie not from population P) triplets is not allowed, or should be taken into account by including additional parameters, for example, p_1 and p_2 representing allele frequencies in (at least) two different populations.

To overcome these problems, we observe that the likelihood can alternatively be factorized as

$$\begin{aligned}&\{\Pi\Pr\ (\text{cases}|\text{parents of cases};\ p, \gamma_1, \gamma_2)\}\\&\quad\times\{\Pi\Pr\ (\text{parents of cases}|p, \gamma_1, \gamma_2)\\&\quad\times\Pi\Pr\ (\text{control}|p)\\&\quad\times\Pi\Pr(\text{founder cases}|p, \gamma_1, \gamma_2)\}\end{aligned} \tag{2}$$

Thus, the likelihood can be written as the product of two 'independent' factors (the two expressions in {}). The first factor specifies the distribution of the genotype of cases, *conditional* on the genotype of their parents, that is, the TDT in its likelihood formulation.[21] The second factor specifies the distribution of the genotype of parents, controls, and founder cases from the *same* population, but with controls and founder cases randomly sampled, but with parents selected for having an affected child. Now, the factor Pr(parents of cases|p, γ_1, γ_2) does contain information on γ_1 and γ_2 in addition to information on p. Essentially, the information that is 'lost' by using the TDT instead of the MLE is 'regained' by taking into account the parents in the second factor of the above partition of the likelihood.

The implications of this factorization, however, are more far-reaching. Specifically, and importantly, it implies that statistical inference, for example estimation or testing, on γ_1, γ_2 can be carried out on these two factors separately, and

subsequently combined using standard methods for combining independent studies. Thus, we could use the TDT to analyze the first factor and use case–control methods to analyze the second factor. If some affected children (cases) are from a different population than the controls (as in our example) and founder cases, they can be included in the first factor, but their parents need to be omitted from the second. This type of factorization of the likelihood has recently been used to develop optimal score tests to *test* for a genetic effect.[22,23]

Here, we propose a simple estimator of the genetic relative risk under the multiplicative model, using a logistic model approximation to the likelihood (cf below) and to test whether this estimate is statistically significantly different from 1. We will assume that the conditions for the valid use of both the TDT (few) and case–control methods (notably the absence of stratification, cases and controls arising from the same population) are fulfilled.

Estimation

We first consider separate (for TDT and parent–control separately) estimates of the parameters γ_1, γ_2.

TDT

From the expressions in Table 1 we can show that the probability r that a heterozygote parent transmits the risk allele is:

$\gamma_1/(1+\gamma_1)$, for heterozygous (1/2) parents, when the other parent is 1/1.
$(\gamma_1+\gamma_2)/(1+2\gamma_1+\gamma_2)$, for heterozygous parents when the other parent is also heterozygous (1/2).
$(\gamma_2)/(\gamma_1+\gamma_2)$, for heterozygous parents when the other parent is 2/2.

For the marginal probability (ie marginalized over the other parent) r that a heterozygous parent transmits the high susceptibility allele, under HWE, we have

$$r/(1-r) = \{p\gamma_2 + (1-p)\gamma_1\}/\{p\gamma_1 + (1-p)\} \qquad (3)$$

Taking the multiplicative model ($\gamma_1=\gamma$; $\gamma_2=\gamma^2$) we have that for the TDT the relative risk $\gamma = r/(1-r)$, where r is the probability of transmission of a risk allele. Note that in this case $r/(1-r)$ is independent of p, and does not require HWE, as it is independent of the co-parent's genotype. For other models one should either know p, or consider *both* parents simultaneously when evaluating the TDT in order to estimate γ_1 and γ_2. In this (multiplicative) case, however, parents may be considered separately.

Under the multiplicative model, γ can also be estimated using logistic regression, with the 1/0 outcome denoting – for informative heterozygous parents – whether the allele of interest has been transmitted or not. When no covariables are included, the exponent of the estimated intercept estimates γ. Under the multiplicative model the logistic model yields the exact relative risk.

Covariables, such as the age of the child, its diet, or the presence of certain alleles at different loci, could be included to explore whether the effect of the putative susceptibility allele on disease risk depends on other factors; a form of interaction or effect modification.[24]

If only one of the parents is known for some of the children, then multiple imputation, as described above, can be used. When only one parent is available for all children, the 1-TDT can be used to test for an association between allele and disease. Unfortunately, while this poses no problems for testing, the 1-TDT statistics T_1 and T_2 depend on both γ_1, and γ_2, making estimation more difficult.

Parents–controls–founder cases

We will assume that the (subset of) parents included in this (sub) analysis are from the same population as the controls and founder cases. For the multiplicative model, we do not need to assume that this population is in HWE, as relative risks do not depend on this assumption. However, we take a population in HWE as an example (Table 2). Genotypes of controls follow directly from the HWE assumption. For

Table 2 (a) Genotype probabilities for cases, parents and controls. (b) Relative risks for cases, parents and control under the multiplicative model

	1/1	*1/2*	*2/2*
(a)Genotype probabilities			
Case	$(1-p)^2/\{(1-p)^2+\gamma_1 2(1-p)p+\gamma_2 p^2\}$	$\gamma_1 2(1-p)p/\{(1-p)^2+\gamma_1 2(1-p)p+\gamma_2 p^2\}$	$\gamma_2 p^2/\{(1-p)^2+\gamma_1 2(1-p)p+\gamma_2 p^2\}$
Parent	$\{(1-p)^2\}\{(1-p)+p\gamma_1\}/T$	$\{(1-p)p\}\{(1-p)+\gamma_1+p\gamma_2\}//T$	$\{p^2\}\{(1-p)\gamma_1+p\gamma_2\}/T$
Control	$(1-p)^2$	$2(1-p)p$	p^2
(b) Relative risks			
Case	1	γ	γ^2
Parent	1	$(1+\gamma)/2$	γ
Control	1	1	1

T has the same value as in Table 1.

Table 3 Data set-up for generalized logistic regression using Poisson regression

Comments	*N*	*y*	*x*	*z*
TDT, 1/2 parent, 2 allele transmitted to child		1	1	0
TDT, 1/2 parent, 2 allele not transmitted to child		0	1	0
Parent–control–founder cases. 1/1 genotype		2 = founder case 1 = parent 0 = control	0	1
Parent–control–founder cases. 1/2 genotype		2 = founder case 1 = parent 0 = control	0.5	1
Parent–control–founder cases. 2/2 genotype		2 = founder case 1 = parent 0 = control	1	1

parents selected for having an affected child, we have to take appropriate sums over the 10 possible situations in order to calculate their genotype distribution.

Thus, for the parent–control study, for the multiplicative model, we have that the odds ratio of the association between being a parent and being 1/2 heterozygous and 2/2 homozygous (relative to being a control, and being1/1 homozygous) equals $(1+\gamma)/2$ and γ, respectively. Writing $\gamma = 1+\delta$, we have that these are $1+\delta/2$ and $1+\delta$ respectively, or – when γ is not very large (<3, say) – approximately $\sqrt{\gamma}$ and γ. We can thus estimate γ by using logistic regression with a covariable x having values 0, 0.5, and 1, for 1/1, 1/2 and 2/2 individuals respectively. Parents have 'response' value $y=1$ and controls $y=0$. We can estimate γ by $\exp(\rho)$, where ρ is the estimated coefficient of x. Standard errors and confidence intervals are automatically provided by most standard software (eg SAS). As the approximation depends on $\gamma = 1+\delta$ being not very large, logistic regression is only an approximation. For founder cases, the odds ratio of being a case and being 1/2 heterozygous and 2/2 homozygous (relative to being a control, and being1/1 homozygous) equals γ and γ^2 respectively. In order to incorporate these founder cases we assign them a 'response' value $y=2$. As, in this case, y can assume three different values (0, 1, and 2), standard logistic regression is inappropriate, and one should use the adjacent-category logit model, a generalized logistic regression model.[25–27]

Note that we implicitly assumed that the genotype distribution of the two parents of a case are independent (within the population of parents of cases), that is, that there is random (not assortative) mating. One should also be aware of other sources of bias. For example, if 2/2 parents would be less fertile, these parents would be underrepresented in the parent–control comparison for reasons unrelated to disease in their children.[28]

Note that the assumption of a multiplicative model can be tested using either the parents and controls, or – even better – cases and controls. Under this model, the relative risk of 2/2 cases should be the square of 1/2 cases. However, the power of such goodness-of-fit tests will usually be low.

Combining TDT and parents–controls–founder cases

For the multiplicative model, an overall estimator for γ is obtained by combining the two logistic regressions into a single one. A simple method for this is Poisson regression.

Records are created as follows. Each record consists of the same four variables, a number n of cases, an outcome variable y, a record type z, and a covariable x. The first group of records pertains to heterozygous informative parents, and y denotes whether the susceptibility allele has been transmitted ($y=1$) or not ($y=0$). The record type z is set = 0, and the covariable x is set = 1.

The second group of records pertains to parents, founder cases, and controls, as described above. The variable $x=0$, 0.5, or 1, depending on the frequency of occurrence of the susceptibility allele. The record type variable z is set = 1, and the outcome variable y assumes the value 0, 1, or 2, for controls, parents and founder cases, respectively. No intercept is required.

The required layout of the data set is shown in Table 3.

In SAS, for example, Poisson regression is carried out as follows. Copies y1 and x1 are made of x and y, and a variable $z1 = 1-z$ created. The following commands will do the analysis.

```
PROC GENMOD;
CLASS x1 y1;
MODEL n=x1*z y1*z x*y z1/ LINK=LOG ERROR=POISSON NOINT;
RUN;
```

The exponent of the coefficient of $x*y$ will estimate γ, that is, the relative disease risk.

Confidence intervals and *P*-values can be based on Wald type tests. As logistic regression yields only an approximation of the true disease relative risk for the parents–control comparison, so does the combined (generalized) logistic regression.

A complication may arise when some of the affected children are siblings. As they may share other risk factors (eg environmental) in addition to shared genes at the susceptibility locus; their observations cannot be treated as independent. Fortunately, the Generalized Estimating Equation (GEE) approach to logistic regression can be used to take such dependencies into account.[29] However, if the number of such siblings is small, ignoring the dependency among them is unlikely to seriously affect the parameter estimates.

Example

In a study carried out in two pediatric hospitals in The Netherlands, cases were 207 children hospitalized for a serious respiratory syncytial virus (RSV) infection. This infection is common in infants, but usually takes an uncomplicated course. It has been hypothesized that interleukin genes may play a role in the development of serious disease, requiring hospitalization.[30] In the study reported here, several polymorphic loci (each with two known different alleles), suspected to play a role in the immune response to this virus, were studied, but here we will only consider the gene coding for interleukin-4. Details of this study have been published elsewhere.[20] Briefly, parents of all children were approached for permission to enroll their children and were requested to send some scrapings from their oral mucous membrane for DNA testing. In addition, 447 random population controls were also genotyped in order to increase the power of the study by adding a CC component to the study, and to obtain background information on the population frequency of risk alleles.

Of the 193 mothers and 186 fathers who agreed to participate (several couples had more than one child enrolled), there were 114 informative parents of whom 65 transmitted the mutant allele suspected of being associated with serious disease. With β_{xy} estimated at 0.28 (standard error 0.19), γ_{TDT} is estimated at $\exp(0.283) = 1.33$ (one-sided *P*-value 0.067). This is somewhat suggestive of a positive association, and lack of statistical significance may have been due to a lack of statistical power. In the parents–control comparison the DNA of 447 (adult) controls, selected for being 'native Dutch'[31] was compared to that of 379 parents of whom 321 were classified as 'native Dutch', and included in the parents–control comparison. Of these parents, 223 were homozygous for the nonrisk allele, 93 were heterozygous and five were homozygous for the 'risk' allele. For the controls, these numbers were 342, 94, 11 respectively. The slope β of logistic regression of parent (parent = 1, control = 0) on half the number of risk alleles = 0.495 (standard error 0.29) one-sided *P*-value 0.045), yielding a γ of $\exp(0.495) = 1.64$.

A combined logistic regression yields $\beta = 0.345$ (standard error 0.16), giving $\gamma = 1.41$ (95% CI 1.03–1.93), which is significantly different (two-sided) from 1.

Note that the addition of controls has added some power to the TDT, but that the relative sizes of the standard errors indicates that most (approximately 2/3) of the information of the joint analysis still came from the TDT part of the data.

In would seem more efficient, where possible, instead of including only triplets and controls, also to include founder cases, that is, a true case–control study.

In order to illustrate this point, we took the above estimate of γ (1.41), and an allele frequency (of allele '2') of 0.16. With these parameter values we simulated studies in which, in addition to 114 informative parents and 379 parents, either (a) 447 controls or (b) 224 controls and 223 founder cases, were included.

For the 1000 simulations of study type a, we found a mean estimate of $\ln(\gamma)$ of 0.352 (true value $\ln(1.41) = 0.344$) , with a standard deviation of 0.16. The mean estimated standard error is 0.154. Clearly, the procedure appears to perform satisfactorily. Analysis of only the (simulated) TDT part yields a mean estimate of 0.353, with a standard deviation of 0.175. The mean standard error then equals 0.191. Thus, adding controls did indeed increase the power (reduce the standard error), but not by very much. The combined study has a standard error comparable to a TDT that is approximately 30–40% larger than the one actually done.

For the type *b* simulations we found a mean of 0.356 with a standard deviation of 0.125, and a mean estimated standard error of 0.128. Again, the procedure appears to function properly. As expected, the standard error is smaller when both controls and founder cases are available, as the combined TDT and CDC (with both controls and founder cases) has a standard error approximately equivalent to a TDT study 2.3 times as large.

Discussion

We presented a simple new method to combine the TDT and case–control methodologies. It complements separate TDT and case–control analyses. It integrates and presents the total evidence for the association between an allele and a disease. The integrated study is more powerful than the constituent parts. The design that makes the most efficient use of resources (genetic tests) appears to be one in which, in addition to the TDT information, both controls and founder cases are available.

As our approach combines the two methodologies of TDT and CC, it is also sensitive to the assumptions that underlie either of them, such as absence of population admixture, absence of parent-of-origin effects, or the assumption that the controls are from the same population

as (the subset) of parents with whom they are compared. Our method does not guard against bias that may arise when these assumptions are wrong. Violation of underlying assumptions may be suspected when the TDT and case–control analyses yield quite disparate estimates of the risk of disease associated with an allele. In such cases, researchers should emphasize the discrepancy of results rather than obscure it by lumping all the data together in a single likelihood.

So far, we have totally ignored parental affection status. For some diseases, parents of children may also be affected by the disease of interest, however, and one may wonder whether this should affect our calculations. We believe that parental affection status can safely be ignored, as the differences between parents and controls are based on selection on the affection status of the children. Similarly, controls are supposed to be a random population sample, and their affection status should similarly be ignored. If a disease is highly prevalent and controls have been selected for not being affected, the formulae presented should be adapted to reflect selection of controls. If multiple loci (eg markers) are of interest then we recommend separate analyses for all markers. However, an exception should be made for closely linked markers whose phase can be identified. In this situation it is probably simplest to list and identify the haplotypes of all individuals involved, treat the haplotypes as multiple alleles occurring at a single 'locus', and compare each haplotype in turn to all other ones combined.

Acknowledgements

We thank the late Dr LA Sandkuijl for his many useful suggestions.

References

1 Weiss KM: *Genetic variation and human disease*. Cambridge: Cambridge University Press, 1993.
2 Kimman TG: *Genetics of infectious disease susceptibility*. Boston, Dordrecht, London: Kluwer Academic Publishers, 2001, ISBN 0-7923-7155-0.
3 Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993; **52**: 506–516.
4 McGinnis RE, Ewens WJ, Spielman RS: The TDT reveals linkage and linkage disequilibrium in a rare disease. *Genet Epidemiol* 1995; **12**: 637–640.
5 Ewens WJ, Spielman RS: The transmission/disequilibrium test: history, subdivision, and admixture. *Am J Hum Genet* 1995; **57**: 455–464.
6 Spielman RS, Ewens WJ: The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet* 1996; **59**: 983–989.
7 Spielman RS, Ewens WJ: A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet* 1998; **62**: 450–458.
8 Spielman RS, Ewens WJ: TDT clarification. *Am J Hum Genet* 1999; **64**: 668.
9 Sham PC: *Statistics in human genetics*. London: Arnold, 1998.
10 Khoury MJ, Beaty TH, Cohen BH: *Fundamentals of genetic epidemiology*. New York: Oxford University Press, 1993.
11 Rothman KJ, Greenland S: *Modern epidemiology*, 2nd ed. Philadelphia: Lippincott-Raven, 1998.
12 Mitchell LE: Relationship between case–control studies and the transmission/disequilibrium test. *Genet Epidemiol* 2000; **19**: 193–201.
13 Schaid DJ: Likelihoods and TDT for the case–parents design. *Genet Epidemiol* 1999; **16**: 250–260.
14 Weinberg CR: Methods for detecton of parent-of-origin effects in genetic studies of case-parents triads. *Am J Hum Genet* 1999; **65**: 229–235.
15 Weinberg CR, Wilcox AJ, Lie RT: 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. *Am J Hum Genet* 1998; **62**: 969–978.
16 Cordell HJ, Barratt BJ, Clayton DG: Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene–gene and gene–environment interactions, and parent-of-origin effects. *Genet Epidemiol* 2004; **26**: 167–185.
17 Weinberg CR: Allowing for missing parents in genetic studies of case–parent triads. *Am J Hum Genet* 1999; **64**: 1186–1193.
18 Schafer JL: *Analysis of incomplete multivariate data*. London: Chapmann & Hall, 1997.
19 Sun F, Flanders WD, Yang Q, Khoury MJ: Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT. *Am J Epidemiol* 1999; **150**: 97–104.
20 Hoebee B, Rietveld E, Bont L *et al*: Association of severe respiratory syncytial virus bronchiolitis with interleukin-4 and interleukin-4 receptor alpha polymorphisms. *J Infect Dis*. 2003; **187**: 2–11.
21 Abel L, Müller-Myhsok B: Maximum-likelihood expression of the transmission/disequilibrium test and power considerations. *Am J Hum Genet* 1998; **63**: 664–667.
22 Whittemore AS, Tu I-P: Detection of disease genes by use of family data. I. Likelihood-based theory. *Am J Hum Genet* 2000; **66**: 1328–1340.
23 Tu I-P, Balise RR, Whittemore AS: Detection of disease genes by use of family data. II. Application to nuclear families. *Am J Hum Genet* 2000; **66**: 1341–1350.
24 Maestri NE, Beaty TH, Hetmanski J *et al*: Application of transmission disequilibrium tests to nonsyndromic oral clefts: including candidate genes and environmental exposures in the models. *Am J Med Genet* 1997; **73**: 337–344.
25 Agresti A: A survey of models for repeated ordered categorical response data. *Stat Med* 1989; **8**: 1209–1224.
26 McCullagh P, Nelder JA: *Generalized linear models*, 2nd ed London: Chapman & Hall, 1989.
27 Ananth CV, Kleinbaum DG: Regression models for ordinal responses: a review of methods and applications. *Int J Epidemiol* 1997; **26**: 1323–1333.
28 Pritchard JK, Rosenberg NA: Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 1999; **65**: 220–228.
29 Diggle P, Heagarty P, Liang K-Y, Zeger S: *Analysis of longitudinal data*, 2nd ed Oxford: Oxford University Press, 2002.
30 Hull J, Thomson A, Kwiatkowski D: Association of respiratory syncytial virus bronchiolitis with the interleukin 8 gene region in UK families. *Thorax* 2000; **55**: 1/23–1/27.
31 Viet AL, van Gils HWV, van den Hof S *et al*: *Risicofactoren en gezondheidsevaluatie nederlandse bevolking een onderzoek op GGD'en (Regenboog-project)*. Bilthoven (The Netherlands): National Institute of Public Health and the Environment (RIVM), 2001.