TDTcovariates083004final.tex

Covariate Adjustment in Family-based Association Studies

Alice S. Whittemore,[1] Jerry Halpern[2] and Habibul Ahsan[3]

[1]Division of Epidemiology, Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA

[2]Division of Biostatistics, Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA

[3]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY

Corresponding Author:

Alice S. Whittemore, Ph.D.

Stanford University School of Medicine

Department of Health Research and Policy

HRP Redwood Building, Room T204

Stanford, CA 94305-5405

Phone: (650) 723-5460

Fax: (650) 725-6951

E-mail: alicesw@stanford.edu

Keywords: confounding, covariate adjustment, efficient score statistic, genetic association, transmission/disequilibrium test, transmission distortion

ABSTRACT

Family-based tests of association between a candidate locus and a disease evaluate how often a variant allele at the locus is transmitted from parents to offspring. These tests assume that in the absence of association, an affected offspring is equally likely to have inherited either one of the two homologous alleles carried by a parent. However transmission distortion has been documented in families in which the offspring are unselected for phenotype. Moreover if offspring genotypes are associated with a risk factor for the disease, transmission distortion to affected offspring can occur in the absense of a causal relation between gene and disease risk. We discuss the appropriateness of adjusting for established risk factors when evaluating association in family-based studies. We present methods for adjusting the transmission/disequilibrium test (TDT) for risk factors when warranted, and we apply them to data on CYP19 (aromatase) genotypes in nuclear families with multiple cases of breast cancer. Simulations show that when genotypes are correlated with risk factors the unadjusted test statistics have inflated size, while the adjusted ones do not. The covariate-adjusted tests are less powerful than the unadjusted ones, suggesting the need to check the relation between genotypes and known risk factors to verify that adjustment is needed. The adjusted tests are most useful for data containing a large proportion of families that lack disease-discordant sibships, i.e., data for which multiple logistic regression of matched sibships would have little power. Software for performing the covariate-adjusted tests is available at http://www.stanford.edu/dept/HRP/epidemiology/COVTDT.

INTRODUCTION

Two decades ago, geneticists noted that genetic association with disease can be detected by evaluating excess transmission of particular alleles from parents to affected offspring. This strategy has been formalized in the haplotype relative risk test and the transmission/disequilibrium test (TDT) and their extensions [Field et al., 1986; Falk and Rubinstein, 1987; Terwilliger and Ott, 1992; Spielman, et al. 1993]. These tests are not biased by population stratification, a problem that can be serious for large case-control studies of small genetic effects [Marchini et al., 2004]. A fundamental assumption of the family-based tests is that, under the null hypothesis of no association between disease and variant allele, a parent is equally likely to have transmitted either of his two homologous alleles to an affected offspring (hereafter called *Mendelian transmission*). Large deviations from Mendelian transmission are interpreted as evidence that disease risk varies with alleles of the polymorphism or of a neighboring locus.

However deviations from Mendelian transmission can occur for other reasons. These include meiotic drive (biased segregation during meiosis), gametic selection (differential success of gametes in achieving fertilization) and postzygotic viability selection for or against certain genotypes. Zollner et al. (2004) studied 148 nuclear families ascertained without reference to phenotype, and found evidence for tranmission distortion spread broadly throughout the genome. At some loci, the distortion can be appreciable [Eaves et al., 1999]. We shall use the term *neutral distortion* for departures from Mendelian transmission that are unrelated to phenotypes for the disease of interest. As noted by Spielman et al. (1993), such distortion can cause spurious disease-genotype association in transmission/disequilibrium tests based only on affected offspring.

More problematic to transmission-based association tests is *nonneutral distortion*, wherein tranmission distortion differs by disease status of offspring. Such differential distortion might occur if genotypes are associated with a risk factor for the disease. For example, the CYP17 gene encodes an

enzyme that functions at key branch points in human steroidogenesis. Carriers of the variant A2 allele of a polymorphism in this gene appear less likely than noncarriers to use estrogen therapy (ET) for menopausal symptoms [Feigelson, 1999]. Since ET is a risk factor for breast cancer [Writing Group for Women's Health Initiative, 2002], women with breast cancer are likely to have used the therapy and thus may be less apt to carry the A2 allele than expected according to their parental genotypes. Failure to adjust for ET use could induce a spurious negative association between A2 carrier status and breast cancer, if carrier status has no effect on breast cancer risk. Such failure also could mask a true, causal association between A2 status and risk that is independent of ET use (see Figure 1A).

Nonneutral distortion also can occur if genotypes are associated with one or more co-morbid conditions [Smoller et al., 2000; Robins et al., 2001]. Such conditions are particularly likely to occur in clinic based data, because individuals with multiple disorders are more apt to seek medical care and receive diagnostic evaluation. Bias can occur when the candidate gene is unrelated to the disease of interest but is in linkage disequilibrium with a gene that affects the co-morbid condition.

Here we extend family-based tests to evaluate such departures from Mendelian transmission while accomodating the effects of covariates such as age and ET use. In applying these methods, we must avoid controlling for unmodifiable covariates that lie in a causal pathway between genotype and disease. For example, endogenous estrogen levels may vary with genotype of the CYP19 gene, which encodes the enzyme aromatase that converts androgens to estrogen. Since estrogens are involved in both onset of menarche and development of breast cancer, certain CYP19 genotypes may increase breast cancer risk by increasing estrogen levels (which also may cause early age at menarche, an established risk factor for breast cancer). Thus age at menarche is a marker for estrogen levels, which may lie on the causal pathway between CYP19 genotypes and breast cancer (Figure 1B). If so, then controlling for age at menarche when examining association between CYP19 genotypes and breast cancer risk would be counterproductive, unless interest focused on a possible association between genotype and risk that is independent of the estrogen pathway marked by age at menarche.

In the next section we establish notation and introduce the test statistics as efficient score statistics of likelihood functions that include covariates. We then apply a covariate-adjusted TDT to CYP19 genotype data from nuclear families with multiple cases of breast cancer. We use simulations to examine the tradeoffs between bias and power loss when considering covariate adjustment. We conclude with general recommendations for dealing with covariates and potential confounding in family-based association studies.

NOTATION AND ASSUMPTIONS

We wish to evaluate whether an offspring's genotype influences his disease risk, while allowing the possibility that his genotype also influences his risk factors for the disease (hereafter called covariates). These possibilities are illustrated in Figure 2. The likelihood-based framework for covariate-adjusted tests of genotype-disease association is similar to that developed for those without covariates (hereafter called "no-covariate" TDT's) [Schaid and Rowland, 1998; Clayton, 1999; Whittemore and Tu, 2000; Shih and Whittemore, 2002]. We illustrate the theory by applying it to a binary disease outcome and a diallelic polymorphism with a variant and normal allele, using nuclear family data with known parental genotypes. Extension to data involving multiple markers and missing parental genotypes is similar to that described elsewhere for the no-covariate TDT.

Suppose that for N unrelated nuclear families we have gathered data on the genotypes, covariates and binary disease statuses of the offspring, and the genotypes of their parents. Let $h = 0, 1$ or 2 denote the number of variant alleles in an offspring's genotype. We begin by considering the possibility of neutral distortion, without consideration of covariates. Specifically, we assume that the probability of offspring genotype h, given his parental genotypes g, is

$$\Pr(h|g) = \frac{P_M(h|g)e^{\lambda_h}}{\sum_{h'=0}^{2} P_M(h'|g)e^{\lambda_{h'}}},\ \lambda_0 = 0. \tag{1}$$

Here $P_M(h|g)$ denotes Mendelian transmission probability and λ_h is a scalar parameter, with λ_0

equated to 0 to insure identifiability. In this notation,

$$\frac{\Pr(h|g)}{\Pr(0|g)} = \frac{P_M(h|g)}{P_M(0|g)} e^{\lambda_h}, \quad h = 1, 2,$$

which states that the likelihood ratio of genotype h for an offspring with parental genotype g is distorted from the Mendelian ratio by the factor e^{λ_h}. Thus $\lambda_1 = \lambda_2 = 0$ corresponds to Mendelian transmission.

To address the possibility that genotypes may influence covariates (upper arrow in Figure 2), we assume an exponential family model [Hogg and Craig, 1971] for the distribution of an offspring's covariates, given his genotype:

$$\Pr(z|h) = \exp\left[\xi_h z + S(z) + \xi_{h0}\right]. \tag{2}$$

Here $z = (z_1, ..., z_p)^T$ denotes a p-dimensional column vector of covariates, $\xi_h = (\xi_{h1}, \xi_{h2}, ..., \xi_{hp})$ is a row vector of parameters, and

$$\xi_{h0} = \ln\left[\int \exp\left[\xi_h z + S(z)\right] dz\right]^{-1}, \quad h = 0, 1, 2,$$

is determined to insure that the probabilities (**??**) sum or integrate to one over the covariate space. The components of the vectors $\xi_h - \xi_0$ have interpretions as log-odds-ratios, since for two covariate vectors z and z',

$$\frac{\Pr(h|z)}{\Pr(0|z)} \div \frac{\Pr(h|z')}{\Pr(0|z')} = \frac{\Pr(z|h)}{\Pr(z|0)} \div \frac{\Pr(z'|h)}{\Pr(z'|0)} = e^{(\xi_h - \xi_0)(z - z')}.$$

From Bayes' Rule, the probability that an offspring has genotype h, given his parents' genotype g and his covariates z, is

$$\Pr(h|g, z) = \frac{P(h|g)\Pr(z|h)}{\sum_{h'=0}^{2} P(h'|g)\Pr(z|h')}, \tag{3}$$

where $P(h|g)$ is given by (**??**). Here we have assumed that, given his own genotype, an offspring's covariates are independent of the genotypes of his parents. Substituting (**??**) and (**??**) into (**??**) gives

$$\begin{aligned} \Pr(h|g,z) &= \frac{P_M(h|g)\exp(\lambda_h + \xi_{h0} + \xi_h z)}{\sum_{h'=0}^{2} P_M(h'|g)\exp(\lambda_{h'} + \xi_{h'0} + \xi_{h'} z)} \\ &\equiv \frac{P_M(h|g)\exp(\nu_h z_*)}{\sum_{h'=0}^{2} P_M(h'|g)\exp(\nu_{h'} z_*)},\ \nu_0 = 0. \end{aligned} \tag{4}$$

Here $z_*^T = (1, z^T)$, and $\nu_h = (\nu_{h0}, ..., \nu_{hp})$, with $\nu_{h0} = \lambda_h + \xi_{h0} - \xi_{00}$ and $\nu_{hj} = \xi_{hj} - \xi_{0j}$, $j = 1, ..., p$. The vectors ν_1 and ν_2 measure association between offspring genotype and covariates.

Some special cases of model (**??**) warrant comment. First, when there is no genotype-covariate association (i.e., when $\xi_h \equiv \xi$ in (**??**)), then, since the intercepts ν_{h0} equal λ_h and the regression coefficients ν_{hj} are 0, $j = 1, ..., p$, (**??**) reduces to the neutral distortion probabilities (**??**). Second, when there is neither genotype-covariate association nor neutral distortion (i.e., $\xi_h \equiv \xi$ in (**??**) and $\lambda_1 = \lambda_2 = 0$ in (**??**)), then $\nu_1 = \nu_2 = 0$ and the probabilities (**??**) reduce to the Mendelian probabilities $P_M(h|g)$. Third, when genotypes and covariates are associated, the intercepts ν_{10} and ν_{20} are generally nonzero, even without neutral distortion. Specification that the intercepts ν_{10} and ν_{20} are 0 (with arbitrary regression coefficients ν_{hj}, $j = 1, ..., p$) corresponds not only to the assumption of no neutral distortion, but also to the additional assumption that Mendelian transmission holds at the "baseline" covariate levels $z_j = 0$, $j = 1, ..., p$. Since our primary concern here is with confounding by covariates rather than neutral distortion, we shall assume hereafter that $\lambda_1 = \lambda_2 = 0$.

Model (**??**) requires estimating the $2(p+1)$ parameters in ν_1 and ν_2, and more parsimonious submodels may be desirable. For example, setting $\nu_1 = \nu_2$ (equivalently, setting $\xi_1 = \xi_2$ in (**??**)) indicates that the covariate distribution among carriers of one variant allele is similar to that among carriers of two variants (a dominant model for the effect of genotype on covariates). Alternatively, setting $\nu_1 = 0$ (equivalently, setting $\xi_1 = \xi_0$ in (**??**)) indicates that the covariate distribution among carriers of one variant allele equals that among those with a normal genotype (a recessive genotype-

covariate model).

We illustrate these models by considering a single binary exposure having prevalence 20% among normal homozygotes and 30% among all carriers of a variant allele A (a dominant genotype-covariate model). We let z denote an indicator for exposure, with $z = 1$ for exposed and $z = 0$ for unexposed, and write $\Pr(z = 1|h) = e^{\xi_{h1} z} / \left(1 + e^{\xi_{h1}}\right)$. This corresponds to model (**??**) with $S(z) = 0$ and $\xi_{h0} = -\ln\left(1 + e^{\xi_{h1}}\right)$. Equating $\Pr(z = 1|h = 0)$ to .2 and $\Pr(z = 1|h = 1) = \Pr(z = 1|h = 2)$ to .3 in equation (**??**) and solving for the ξ's yields

$$(\xi_{00}, \xi_{01}) = (-.223, -1.386) \ \textit{and} \ \ (\xi_{10}, \xi_{11}) = (\xi_{20}, \xi_{21}) = (-.357, -.847).$$

Table I gives the offspring genotype probabilities $\Pr(h|g, z)$, conditional on parental genotype and offspring exposure. Column 3 of the table gives the usual Mendelian transmission probabilities. Columns 4 and 5 show the probabilities $\Pr(h|g, z)$ for exposed ($z = 1$) and unexposed ($z = 0$) offspring, based on (**??**) with $\nu_1 = \nu_2 = (-.357, -.847) - (-.223, -1.386) = (-.134, .539)$. Comparison of columns 4 and 5 with column 3 illustrates two points. First, offspring genotypes that are inconsistent with parental genotypes under Mendel's laws remain so according to (**??**). Second, offspring of parental mating type $AA \times AB$ have a 50:50 chance of carrying two variants rather than one, regardless of their exposure level, in agreement with Mendelian inheritance. This agreement occurs because we have assumed the same exposure prevalence among carriers of one variant and two variants (a dominant model for the effects of genotype on exposure).

To address the possibility that both genotypes and covariates affect disease risk, we must specify a model $\varphi(h, z)$ for the probability of disease given genotype and covariates. We assume that $\varphi(h, z)$ depends on h only through a term $\beta c(h)$, where β is an unknown scalar parameter and $c(h)$ indicates how genotypes affect risk (lower arrow in Figure 2). For example, in the simulations we shall assume

a logistic regression model of the form

$$\varphi(h,z) = \Pr(y=1|h,z) = \frac{e^{\alpha+\beta c(h)+\delta z}}{1+e^{\alpha+\beta c(h)+\delta z}}, \tag{5}$$

where y is an indicator for disease, and $\delta = (\delta_1, ..., \delta_p)$ is a vector of parameters relating covariates to disease risk. For a dominant model, $c(h) = 1$ if $h = 1, 2$, with $c(0) = 0$. For a recessive model, $c(h) = 1$ if $h = 2$, with $c(h) = 0$ otherwise, while for an additive model, $c(h) = h$. The null hypothesis of interest is that $\beta = 0$, i.e., that genotypes are unrelated to disease risk.

We shall base inferences for β on the likelihood L of the offspring genotypes, given their phenotypes, their covariates and the genotypes of their parents. This likelihood is the product over families of family-specific contributions

$$\begin{aligned} L_i &= \prod_{j=1}^{n_i} \Pr\left(h_{ij}|y_{ij}, g_i, z_{ij}\right) \\ &= \prod_{j=1}^{n_i} \left\{ \frac{\Pr(h_{ij}|g_i, z_{ij})\varphi\left(h_{ij}, z_{ij}\right)^{y_{ij}} \left[1-\varphi\left(h_{ij}, z_{ij}\right)\right]^{1-y_{ij}}}{\sum_{h'=0}^{2} \Pr(h'|g_i, z_{ij})\varphi\left(h', z_{ij}\right)^{y_{ij}} \left[1-\varphi\left(h', z_{ij}\right)\right]^{1-y_{ij}}} \right\}. \end{aligned} \tag{6}$$

Here the subscripts i and j denote the j^{th} offspring of the i^{th} family, $j = 1, ..., n_i$, and $\Pr(h|g,z)$ is given by (**??**). In (**??**) we have assumed conditional independence of offspring phenotypes, given their genotypes and their covariates. This is a weaker assumption than that underlying the "no-covariate" TDT, which assumes that offspring phenotypes are conditionally independent, given just their genotypes.

The likelihood contributions (**??**) are conditioned on offspring phenotypes because families are ascertained on that basis. They are conditioned on parental genotypes because such conditioning avoids potential bias due to ethnic stratification of the parental population [Spielman et al., 1993]. The contributions also are conditioned on the offspring covariates, because such conditioning avoids specifying the joint distribution of the offspring covariates, which typically is complex and poorly

understood. Note that this conditioning on both the phenotypes and the covariates of the offspring precludes estimating the parameters that relate covariates to disease risk. Instead this relationship must be specified *a priori*. Thus, for example, we must specify the parameters α and δ in the logistic model (**??**).

The likelihood function L prompts several hypotheses about the offspring genotype probabilities, conditional on covariates and parental genotypes. These derive from comparison of the nested models shown in Figure 3. Model 1, the most general model of the four, allows arbitrary values for β and the two ν's. Model 2, with $\beta = 0$ and the ν's arbtrary, specifies that within each covariate level, allele transmission is independent of disease phenotype, although it need not follow Mendelian expectation. Model 3, with β arbitrary and $\nu_1 = \nu_2 = 0$, specifies that in families unselected for disease, alleles are transmitted according to Mendelian expectation. Model 4 ($\beta = \nu_1 = \nu_2 = 0$) specifies Mendelian expectation for allele tranmission, regardless of the offspring phenotypes or covariates. Models 2 and 4 both specify that disease risk $\varphi(h, z) = \varphi(z)$ depends only on covariates. The no-covariate TDT and its extensions are efficient score statistics evaluating the adequacy of Model 4 relative to Model 3. In contrast, the covariate-adjusted TDT's are efficient score statistics evaluating the adequacy of Model 2 relative to Model 1.

The covariate-adjusted TDT is based on the efficient score $\partial \log L/\partial\beta$ evaluated at $\beta = 0$ and with the ν's equated to their null maximum likelihood estimates. When standardized by an estimate of its null standard deviation, the test statistic has approximately a standard Gaussian distribution. As outlined in the Appendix, the efficient score has a form analagous to that for the no-covariate TDT. Both are sums over all offspring of terms

$$[y - \varphi(z)]\,[c(h) - \mu\,(z)]\,, \tag{7}$$

which measure the covariance between null trait residuals and null genotype residuals. Here $\varphi(z)$ is the

user-specified disease prevalence among individuals with covariates z, and $\mu(z)$ is the null expected value of the offspring's genotype value $c(h)$, given his parental genotypes and (for the covariate-adjusted TDT) his covariates. Both $\varphi(z)$ and $\mu(z)$ are assumed independent of z in computing the no-covariate TDT. Expression (**??**) shows that misspecification of $\mu(z)$ induces bias in the test statistic, for then the null expectation of $c(h) - \mu(z)$ is nonzero. In contrast, misspecification of the disease probabilities $\varphi(z)$ does not affect the null expectation of the standardized test statistic, since $w(z) = y - \varphi(z)$ serves only as a user-specified weight for the null genotype residual of each offspring. According to (**??**), positive residuals $c(h) - \mu(z)$ of affected offspring ($w(z) > 0$) contribute positively to the test statistic. Moreover affected offspring with low-risk covariates ($\varphi(z) << 1$ and $w(z) \sim 1$) contribute more than do those with high-risk covariates ($\varphi(z) \sim 1$ and $w(z) \sim 0$). In addition, positive residuals $c(h) - \mu$ of unaffected offspring ($w(z) < 0$) contribute negatively, and unaffected offspring with high-risk covariates ($w(z) \sim -1$) contribute larger negative values than do those with low-risk covariates ($w(z) \sim 0$). While misspecification of the weights could decrease power, in our limited simulations and data analyses we have found that weight specification has negligible impact on either the value of the test statistic or its power.

APPLICATION TO DATA

The need for covariate adjustement in family-based association tests was brought to our attention in the analysis of genotypes of a tetra-nucleotide (TTTA) repeat polymorphism in the CYP19 gene in 278 nuclear families with multiple cases of breast cancer. The polymorphism is characterized by a variable number of repeats, ranging from 7 to 13. Previous studies have suggested that carriers of 10 or more repeats have elevated breast cancer risk (see Dunning et al., 1999 for a review). We genotyped 299 affected and 213 unaffected daughters and 107 of their $2 \times 278 = 556$ parents, and found that carrier status of the allele containing 11 repeats (hereafter called allele A) was associated with increased breast cancer risk, with a (no-covariate) TDT statistic of 1.83 (one-tailed $p = .03$) [Ahsan

et al., 2004].

We were concerned about possible confounding of this association by nongenetic risk factors for breast cancer. To address this issue, we first performed conditional logistic regression (CLR) of the 183 phenotype-discordant sibships. The regression model included: carrier status of the A allele, age at risk (defined as age at breast cancer diagnosis for affected sibs and age at interview for unaffected sibs), age at menarche, parity, oral contraceptive use and ET use. The estimated odds-ratio relating CYP19 genotype to breast cancer risk was 1.9 (95% confidence interval: $0.9 - 3.5$, one-tailed $p = .09$). It is not clear whether this attenuated statistical significance reflects confounding due to failure to adjust for the risk factors, or power loss because only 183 of the 278 sibships were phenotype-discordant. The covariate most strongly correlated with disease was age at risk ($p < .001$ in the CLR). Indeed, as seen in the last column of Table II, daughters with breast cancer were, on average, younger than their unaffected sisters. Confounding by age at risk is possible because it also was related to genotype. Table II shows that homozygotes for the variant were younger than daughters with fewer than two variants, regardless of their disease status.

In an attempt to distinguish confounding from power loss, we examined possible departures from Mendelian transmission to affected and unaffected daughters, adjusted for age at risk. Motivated by the dominant disease-genotype association found for the no-covariate TDT [Ahsan et al., 2004], we assumed a dominant model for the effect of genotype on breast cancer. Motivated by the data in Table II, we assumed a recessive model for the effect of genotype on age at risk (coded as a continuous variable). We specified the null prevalence of breast cancer by age z years as $\varphi(z) = 1 - \exp\left[-\int_0^z I(s)\, ds\right]$, where $I(z)$ is an estimate of the US age-specific breast cancer incidence rate for the period 1973-77 (SEER 1981). This specification corresponds to a weight $w(40) = 1 - \varphi(40) = .99$ for an affected woman aged 40 years and weight $w(70) = -\varphi(70) = -.09$ for an unaffected woman aged 70 years. The resulting age-adjusted test statistic was 1.27 ($p = .10$). This result is consistent with the attenuated findings obtained by CLR. The consistency suggests either that the phenotype-genotype association seen in

the no-covariate TDT is due to confounding by age, or that both the covariate-adjusted TDT and CLR lack power to detect the association.

We attempted to distinguish the two explanations using simulated data. Specifically, we evaluated the power and size of the three tests (no covariate TDT, covariate-adjusted TDT, CLR) in samples with sizes comparable to the CYP19 genotype and breast cancer data. The results, described in the next section, suggest that the covariate-adjusted TDT forfeits about 25% of the power of the no-covariate TDT, while CLR based on only half the sibships forfeits about 50% of this power. Thus while the covariate-adjusted TDT is clearly more powerful than CLR for the CYP19 data, both are considerably less powerful than the no-covariate TDT. Therefore, although both covariate-adjusted analyses provided only weak evidence of association between CYP19 and breast cancer, we cannot exclude the possibility that both analyses lacked the power needed to detect a small increase in risk associated with carrier status of the A allele, independent of age at risk.

SIMULATIONS

We simulated genotype, phenotype and covariate data for a diallelic polymorphism in a candidate gene for 300 nuclear families, each with two offspring. We considered two sibship configurations: A) all 300 sibships were discordant for disease; and B) half the sibships were discordant and the remaining half consisted of two affected siblings. We assumed that genotypes were missing for both parents in half the families, and that one parental genotype was missing in the remaining half. We studied a single binary covariate with values $z = 1$ (exposed) and $z = 0$ (unexposed). For each data set of 300 families, we computed test statistics corresponding to the no-covariate TDT, the covariate-adjusted TDT and conditional logistic regression of both genotypes and covariates of the discordant sibships.

We generated the data for each family in the following four steps:

1) generate parental genotypes assuming random mating and Hardy-Weinberg genotype frequencies, with variant allele frequency equal to 10%;

2) given the parental genotypes, generate offspring genotypes assuming Mendelian transmission;

3) given the offspring genotypes, generate offspring exposure indicators according to two models: a) exposure prevalence of 20%, regardless of genotype. This value corresponds in (**??**) to parameter values $\nu_h = (0, 0)$, $h = 1, 2$, and specifies that exposure is unrelated to genotype; b) exposure prevalence of 40% among carriers of the variant and 20% among noncarriers. These values correspond to a dominant model for the effects of genotype on exposure, with parameters $\nu_1 = \nu_2 = (-0.29, 0.98)$ in equation (**??**);

4) given the offspring genotypes and covariates, generate offspring disease phenotypes according to the logistic regression model (**??**), with $c(h)$ taken to be an indicator for carrier status of the variant allele. We took $\alpha = -2.75$, and considered four models, depending on whether or not disease was associated with genotype ($\beta = 0$ or $\beta = 0.76$) and whether or not disease was associated with exposure ($\delta = 0 or \delta = 0.76$). These values correspond to risks of 6% in unexposed normal homozygotes, 12% in unexposed carriers of allele A and exposed normal homozygotes, and 23% in exposed carriers of allele A. We generated family data until we had obtained 300 families with the desired offspring phenotypes. In summary, we considered two family configurations (A and B); two models for association between genotype and covariate, and four models for disease risk in relation to genotype and covariate, a total of $2 \times 2 \times 4 = 16$ simulations.

The test statistics used to analyze the data assumed that offspring genotypes affect covariates according to the same dominant model used to generate the data. We evaluated various correct and incorrect specifications for the relation beween genotypes and disease (i.e., the weights), and found that the choice of weights had negligible effect on power. Here we report only the results based on the correct specification. For the CLR analyses, we used the t-test statistic for the coefficient β to test the relation between genotype and disease risk .

Table III gives results when all sib pairs are disease-discordant (Table IIIA) and when only half the sib pairs are discordant (Table IIIB). Each half of the table shows test size ($\beta = 0$) and power ($\beta > 0$)

in four blocks, depending on whether or not the covariate is correlated with genotype and with disease risk. Only the fourth block (covariate correlated with both genotype and disease risk) corresponds to confounding of the genotype-phenotype relationship. Thus, as expected, the nominal and empirical test sizes in Table III are similar for blocks 1-3, while in the fourth block the sizes of the no-covariate TDT are inflated.

The covariate-adjusted TDT has power comparable to that of CLR in situations when the latter can use all sibships (Table IIIA). However both suffer appreciable power loss compared to the no-covariate TDT. When half the sibships are disease concordant and thus are excluded from the CLR analysies (Table IIIB), the power loss of this method is substantial. In this case the covariate-adjusted TDT has power intermediate between that of the other two tests.

DISCUSSION

We have extended the TDT to accomodate potential confounding by established risk factors in tests of genetic assocation using nuclear families. Likelihood-based arguments and simulations show that when the covariates are associated with both genotype and disease risk, the empirical type-I error rates for the extended tests are similar to their nominal values, while those for the unadjusted test statistics are inflated.

The covariate adjusted tests are based on the probability distribution of the offspring genotypes, given parental genotypes, offspring disease status and offspring covariates. Because of this joint conditioning on both disease status and covariates, the method does not allow evaluation of association between covariates and disease. Instead, this relationship must be specified *a priori*. Misspecification of the relationship does not affect the type I error rate, although it could decrease power. In simulations and in practice, however, we have found little variation in test power with various specifications of this relation. Although the covariate-adjusted tests presented here handle missing parental data using likelihood-based methods, the underlying theory also could be applied to extensions based on

minimal sufficient statistics for the nuisance parameters [Horvath and Laird, 1998; Rabinowitz and Laird, 2000] or based on projections onto subsets of the parameter space [Rabinowitz, 2002; Allen et al., 2004].

The robustness of the covariate-adjusted statistics is purchased at the price of decreased power relative to that of the unadjusted TDT statistics. This tradeoff between robustness and power raises fundamental issues for the design of family-based association studies. Gathering and typing DNA from parents is costly, and indeed impossible when parents are deceased. An alternative strategy would genotype only siblings and compare genotypes of affected and unaffected sibs. Kraft and Thomas (2004) recently have reviewed several methods for analyzing age-at-onset data for sibships, covering a spectrum of levels of control for potential population stratification. Conditional logistic regression of genotypes and risk factors in matched sibship data provides perhaps the most attractive option for controlling potential confounding by established risk factors. Such analyses are simple to apply and interpret, and provide a measure of the strength of association in addition to a p-value. However our simulations have shown that this option can lose substantial power (relative to transmission-based tests) when a significant proportion of the families lack discordant sibships. Nevertheless, when designing a family-based genetic association study it may be advantageous (both economically and logistically) to restrict the analysis to sibships, without parents. An exception is the situation when genotypes of extended pedigrees have already been collected for other purposes (such as linkage analysis), and an appreciable fraction of the sibships in these pedigrees lack phenotype-discordant sibs. In this case the covariate-adjusted statistics provide an alternative strategy. It is important to examine evidence for correlation between genotypes and risk factors when considering the need for adjustment, and to adjust for a covariate only when it appears to be correlated with genotype yet does not lie on a causal pathway between genotype and disease.

Yet another option for dealing with confounding is unconditional logistic regression of all affected and unaffected family members, whose phenotypes and genotypes are available, and ignoring the

relationships of family members [Slager and Schaid, 2001, 2003; Whittemore and Halpern, 2003]. For this strategy, robust variance estimators are used to accomodate within-family correlations in covariates. A weakness of this approach is potential sensitivity to bias from population stratification.

In conclusion, we suggest the following strategy for dealing with potential confounding of genetic association tests by established risk factors when analyzing data from parents and offspring. First, one should check informally whether the risk factor is associated with the genotype of interest. If not, the no-covariate TDT is the preferred analytic method. If there is evidence for such association, and if most of the families contain discordant sibships, conditional logistic regression of matched sibships is simple and yields estimates of the strength of the association. If an appreciable fraction of the families lack discordant sibships, greater power can be expected with covariate adjustment using the methods presented here. Software for using these methods is freely available at http://www.stanford.edu/dept/HRP/epidemiology/COVTDT.

APPENDIX

A family's contribution to the "no-covariate" likelihood function [Clayton, 1999; Whittemore and Tu, 2000] is the joint probability $\Pr(G, H|Y)$ of the parent and offspring genotype data G and H, respectively, conditional on the offspring phenotypes Y. To include covariates, we take a family's likelihood contribution to be the joint probability $\Pr(G, H|Y, Z)$ of family genotype data, conditioned also on offspring covariates Z. The likelihood depends on a parameter $\theta = (\beta, \nu, \gamma)$, where β relates genotype to trait, $\nu = (\nu_1, \nu_2)$ relates offspring genotypes to covariates in model (**??**), and γ denotes the parameters in the null distribution $\Pr(G)$ of parental genotypes.

When parental genotypes are known, the likelihood factors as the probability $\Pr(G|Y, Z)$ of the parental genotypes, given offspring phenotypes and covariates, times the probability $\Pr(H|G, Y, Z)$ of the offspring genotypes, given parental genotypes, offspring phenotypes and offspring covariates.

The vector U of efficient scores at $\beta = 0$ obtained from the second factor $\Pr(H|G, Y, Z)$ is used to construct the test statistic. Specifically, the test is based on the β-component $U_\beta = U_\beta(\nu)$ of this score vector, evaluated at $\nu = \widehat{\nu}$, where ,$\widehat{\nu}$ maximizes the null probability (**??**) of the offspring genotypes, given parental genotypes and offspring covariates. Standard likelihood theory gives the asymptotic distribution of U_β as Gaussian with mean zero and variance

$$V(U_\beta) = J_{\beta\beta} - J_{\nu\beta} J_{\nu\nu}^{-1} J_{\beta\nu}. \tag{8}$$

Here for example, $J_{\nu\beta}$ is the null expectation of the product $U_\beta U_\nu^T$. The test statistic is

$$T = \frac{U_\beta}{\sqrt{V(U_\beta)}}, \tag{9}$$

evaluated at $\widehat{\nu}$.

When parental genotypes are incomplete, the usual likelihood-based arguments give the score and information for the incomplete family genotype data in terms of moments of the corresponding functions for the complete data, taken over the distribution of the complete data given the observed data. A disadvantage of tests based on this score is their dependence on the model for the distribution of parental genotypes and the resulting possibility of biased inferences for β due to misspecification of this distribution. To address this problem, we use a "partial score" [Clayton, 1999], whose γ-components are the logarithmic derivatives of the parental genotype probabilities, and whose β- and ν-components are the logarithmic derviates of the offspring genotype probabilities, given their phenotypes, their covariates, and the parental genotype information available.

The test is based on the β-component of this partial score, which now depends on the parental genotype parameter γ in addition to ν. The asymptotic variance of this β-component and the form of the test statistic are similar to (**??**) and (**??**), with γ estimated from the null likelihood for the family

genotype data, given the offspring covariates (see Shih and Whittemore, 2003 for details).

ACKNOWLEDGEMENTS

This work was supported by NIH grants CA94069 and CA69417, and DOD grant DAMD170010213. The authors thank Regina Santella for genotyping and Yu Chen for assistance with data analysis. The authors wish to thank an anonymous reviewer for comments that greatly improved an earlier version of the paper.

REFERENCES

Ahsan H, Whittemore AS, Chen Y, Santella R. Variants in Estrogen-biosynthesis Genes CYP17 and CYP19 and Breast Cancer Risk: A Family-based Genetic Association Study. Breast Ca Res (in press)

Chen Allen AS, Satten GA, Tsiatis AA. 2004. Semiparametric estimation of haplotype-disease association with unphased genotype data in family-based studies (submitted).

Clayton D. 1999. A generalization of the transmission/disequilibrium test to uncertain-haplotype transmission. Am J Hum Genet 65:1170-77.

Dunning AM, Healey CS, Pharoah PD, Teare MD, Ponder BA, Easton DF. 1999. A systematic review of genetic polymorphisms and breast cancer risk. Cancer Epidemiol Biomarkers Prev 8:843-54.

Eaves IA, Bennett ST, Forster P, Ferber KM, Ehrmann D, Wilson AJ, Bhattacharyya S, Ziegler AG, Brinkmann B, Todd JA. 1999. Transmission ratio distortion at the INS-IGF2 VNTR. Nat Genet 22:324-5.

Falk CT, Rubinstein P. 1987 Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 51:227-33.

Feigelson HS, McKean-Cowdin R, Pike MC, Coetzee GA, Kolonel LN, Nomura AM, Le Marchand L, Henderson BE. 1999. Cytochrome P450c17alpha gene (CYP17) polymorphism predicts use of hormone replacement therapy. Cancer Res 59:3908-10.

Field LL, Fothergill-Payne C, Bertrams J, Baur MP. 1986. HLA-DR effects in a large German IDDM dataset. Genet Epidemiol Suppl 1:323-8.

Hogg RV, Craig AT. 1971. Introduction to Mathematical Statistics (Third Edition). New York: The MacMillan Co..

Horvath S, Laird NM. 1998. A discordant-sibship test for disequilibrium and linkage: no need for parental data. Am J Hum Genet 63:1886-97.

Kraft P, Thomas D. 2004. Case-sibling gene-association studies for diseases with variable age at onset. Stat Med (in press).

Marchini J, Cardon LR, Phillips MS, Donnelly P. 2004. The effects of human population structure on large genetic association studies. Nat Genet 36:512-17.

Rabinowitz D. 2002. Adjusting for population heterogeneity and mis-specified haplotype frequencies when testing nonparametric null hypotheses in statistical genetics. J Am Stat Assoc 97:742-51.

Rabinowitz D, Laird N. 2000. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered 50:211-23.

Robins JM, Smoller JW, Lunetta KL. 2001. On the validity of the TDT in the presence of comorbidity and ascertainment bias. Genet Epidemiol 21:326-36.

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.

Shih MC, Whittemore AS. 2002. Tests for genetic association using family data. Genet Epidemiol 22:128-45.

Slager SL, Schaid DJ. 2001. Evaluation of candidate genes in case-control studies: a statistical

method to account for related subjects. Am J Hum Genet 68:1457-62.

Slager SL, Schaid DJ, Wang L, Thibodeau SN (2003) Candidate-gene association studies with pedigree data: controlling for environmental covariates. Genet Epidemiol 24:273-83.

Smoller JW, Lunetta KL, Robins J. 2000. Implications of comorbidity and ascertainment bias for identifying disease genes. AJ Med Genet 96:817-22.

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 52:506-16.

Surveillance, Epidemiology, and End Results: Incidence and Mortality Data, 1973-77. 1981. NCI Monograph 57, NIH Publ No. 81-2330, US DHHS.

Terwilliger JD, Ott J. 1992. A haplotype-based "haplotype relative risk" approach to detecting allelic associations. Hum Hered 42:337-46.

Whittemore AS, Halpern J. 2003. Logistic regression of family data from retrospective study designs. Genet Epidemiol 25:177-89.

Whittemore AS, Tu IP (2000) Detection of disease genes by use of family data. I. Likelihood-based theory. Am J Hum Genet 66:1328-40

Writing Group for Women's Health Initiative (WHI). 2002. Risks and benefits of estrogen plus proestin in healthy postmenopausal women: Principal results from the Women's Health Initiative randomized controlled trial. JAMA 288:321-33.

Zollner S, Wen X, Hanchard NA, Herbert MA, Ober C, Pritchard JK. 2004. Evidence for extensive transmission distortion in the human genome. Am J Hum Genet 74:62-72.

Figure Captions

1. Possible associations between genotypes of estrogen metabolizing genes and breast cancer risk. A) carriers of the A1 allele of CYP17 are less likely to use estrogen therapy than noncarriers. Since estrogen therapy is an established breast cancer risk factor, failure to adjust for its use could produce a spurious negative CYP17-breast cancer association, or mask a positive association. B) CYP19 genotypes may increase circulating estrogen levels, which could cause both early age at menarche and increased breast cancer risk. If so, controlling for age at menarche (a surrogate marker on the causal pathway between CYP19 genotype and breast cancer) would be counter-productive, unless the aim is to detect a causal relation between CYP19 and breast cancer that is independent of any effect of CYP19 on age at menarche.

2. Possible causal associations relating genotypes of a candidate gene to risk factors and to disease risk. The relation between genotype and risk factor may be take any one of several forms (dominant, recessive, co-dominant, etc), and need not be the same as the relation between genotype and disease risk.

3. Nested models for the joint effects of genotypes, phenotypes and covariates on parental transmission probabilities. Model 1 allows departures from Mendelian genotype transmission according to offspring risk factors (association parameter ν arbitrary) and disease phenotypes (association parameter β arbitrary). Model 2 specifies that, within each covariate level, allele transmission is independent of disease phenotype ($\beta = 0$ but ν arbitrary). Model 3 specifies that in families unselected for disease, alleles are transmitted according to Mendelian expectation ($\nu = 0$ but β arbritrary). Model 4 specifies Mendelian expectation for allele transmission, regardless of the offspring phenotypes or covariates ($\beta = \nu = 0$). The no-covariate TDT and its extensions evaluate the adequacy of Model 4 relative to Model 3. In contrast, the covariate-adjusted TDT's evaluate the adequacy of Model 2 relative to Model 1.