Original Paper

Human Heredity

Hum Hered 2009;67:88–103
DOI: 10.1159/000179557

Received: January 15, 2008
Accepted after revision: April 16, 2008
Published online: December 12, 2008

On Combining Triads and Unrelated Subjects Data in Candidate Gene Studies: An Application to Data on Testicular Cancer

Li Hsu[a] Jacqueline R. Starr[b] Yingye Zheng[a] Stephen M. Schwartz[c]

[a]Biostatistics and Biomathematics Program, Fred Hutchinson Cancer Research Center, [b]Departments of Pediatrics and Epidemiology, University of Washington; Children's Craniofacial Center, Children's Hospital and Regional Medical Center, and [c]Epidemiology Program, Fred Hutchinson Cancer Research Center; Department of Epidemiology, University of Washington, Seattle, Wash., USA

Key Words
Genetic association studies · Family studies · Population-based case control · Testicular cancer

Abstract
Combining data collected from different sources is a cost-effective and time-efficient approach for enhancing the statistical efficiency in estimating weak-to-modest genetic effects or gene-gene or gene-environment interactions. However, combining data across studies becomes complicated when data are collected under different study designs, such as family-based and unrelated individual-based (e.g., population-based case-control design). In this paper, we describe a general method that permits the joint estimation of effects on disease risk of genes, environmental factors, and gene-gene/gene-environment interactions under a hybrid design that includes cases, parents of cases, and unrelated individuals. We provide both asymptotic theory and statistical inference. Extensive simulation experiments demonstrate that the proposed estimation and inferential methods perform well in realistic settings. We illustrate the method by an application to a study of testicular cancer.

Introduction

The etiology of many common diseases such as cancer and coronary heart diseases is complex, involving an interplay of both genes and environment. Researchers have collected such comprehensive risk factor information on study participants under various design, yet often the statistical efficiency in estimating weak-to-moderate main and interaction effects is limited because of a limited sample size in each study. It is important to be able to combine and analyze these data across study designs to enhance the efficiency in a cost and time effective manner.

Two broad types of study designs, case-unrelated control and family-based, are often used in studies of associations of genetic polymorphisms with disease risk. The case-unrelated control study design is commonly used to study relatively rare, complex, usually late-onset phenotypes such as coronary heart disease and cancer. When such studies are population-based, unrelated controls are recruited from the same geographic or catchment area as the cases and matched to cases on such characteristics as age, gender, and self-reported ethnicity. In contrast, family-based studies use genotypes of blood relatives as a reference with which to compare cases' genotypes. Family-

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0001–5652/09/0672–0088$26.00/0

Accessible online at:
www.karger.com/hhe

Dr. Li Hsu
Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center
1100 Fairview Ave. N.
M2-B500, Seattle, WA 98109 (USA)
Tel. +1 206 667 2854, Fax +1 206 667 7004, E-Mail lih@fhcrc.org

based studies are inherently more robust to population substructure than case-unrelated control studies. However, there are also weaknesses in the family-based design. For example, parents may not be available for late-onset phenotypes. Moreover, in detecting the association of genes with disease risk, statistical power may be impaired, sometimes substantially, due to the correlation of genotypes among family members. This correlation, on the other hand, can be advantageous when assessing rare alleles or gene-gene interactions, since the chance of carrying the rare allele or the high risk alleles at both loci is higher in relatives of cases than in unrelated individuals [Witte et al., 1999; Hopper et al., 2005]. Many authors have compared extensively these two types of designs, see e.g. Thomas [2004]; Weinberg and Umbach [2005]; Laird and Lange [2006].

Joint estimation of the effects of gene and environment on disease risk has been of considerable interest to many investigators. It is worth noting that estimation, though related, is different from testing whether any of these variables is associated with disease risk. Testing an association requires one to devise a powerful test statistic under the null hypothesis of no association against a specific alternative. Estimation, on the other hand, involves estimating unbiasedly the magnitude of the effects of these variables on disease risk and is particularly useful in characterizing the role of gene and environment in disease development. Clearly testing and estimation are related – estimates of the effects along with the uncertainty of these estimates can be used for testing whether a particular variable is associated with disease risk while adjusting for other variables. To some extent estimation examines the association in a more general fashion than testing only. Many of the existing work are focused on testing rather than on estimation and they do not always offer an approach to obtaining unbiased risk estimates. This paper is therefore concerned with methods for joint estimation of the effects of these variables on disease risk.

There are different scenarios for data combining. One scenario is that some cases have family member controls whereas others have unrelated controls and the two sets of cases do not overlap with each other. Such data structures are common in a consortium setting, where some clinical sites have case-control samples while others have family-based samples. Another scenario is that both family-based and unrelated controls are collected for the same set of cases. This scenario is common when the number of cases is limited and there is a wish to take advantage of both designs.

Methods have been developed for combining data collected under different designs particularly for data that arise in the first scenario. These methods include for example a weighted estimate of odds ratios from various data sources [Kazeem and Farrall, 2005; Allen-Brady et al., 2006; Curtin et al. 2007], and a retrospective likelihood approach by Dudbridge [2006]. Nagelkerke et al. [2004] and Epstein et al. [2005] proposed a likelihood approach that accommodate both types of data combining. In their approach, controls' and/or parents of cases' genotypes were used to estimate the mating type, genotype, or allele frequencies, depending on the validity of assumptions regarding random mating and Hardy-Weinberg equilibrium. These population estimates in turn improved the parameter estimation efficiency. An appealing feature of these approaches is that they are likelihood-based and can therefore be understood and applied through established theory and inference procedures. However both approaches are limited to analysis of one marker at a time and no other covariates.

In this paper we propose a pseudo-likelihood approach that can flexibly accommodate multiple genetic markers, environmental covariates, and gene-gene and gene-environment interactions for hybrid data that consists of cases, parents of cases, and unrelated individuals. Such data conceivably could be available from studies of birth outcomes and conditions affecting children and young to middle-aged adults as in, for example, our study of testicular cancer. The approach is readily extended to allow for non-overlapping subsets of case-parents and case-unrelated controls. In fact, the proposed pseudo-likelihood function becomes a likelihood function in this situation. In subsequent sections, we describe this method for hybrid data and the extension to include case-parent sets in which genotypes are missing for one parent. An approach for simplifying the family-based likelihood is also proposed in the presence of multiple unlinked markers. We assess the finite sample performance of the proposed estimators under gene-gene and gene-environment interaction models. We also compare performance with that of existing approaches by Nagelkerke et al. [2004] and Epstein et al. [2005] under a major gene model and observed little difference in efficiency for the proposed pseudolikelihood approach. To demonstrate our method, we present results from an analysis of several candidate polymorphisms in a study of testicular cancer.

Method

A Pseudo-Likelihood Function for Combining Cases, Case-Parents, and Unrelated Controls

Consider a situation where the data consist of I cases, J unrelated controls, and parents of I cases. Let D denote the disease status of an individual, G the genotype of a candidate gene, and Z a vector of other risk factors for disease occurrence. For notational simplicity, we let X be a column vector representing G, Z, and possible interactions among G and Zs that might be incorporated in the model. Assuming that there are no parent-of-origin or maternal genotype effects, the logit form for the relationship between covariates X and disease risk $P(D = 1)$ is

$$\text{logit}\{P(D = 1 \mid X)\} = \alpha + \beta X, \tag{1}$$

where α is the intercept measuring the baseline log-odds of disease risk, $\log\{P(D = 1)/P(D = 0)\}$, for individuals with $X = 0$, and β is a row vector of regression coefficients for covariates X that quantify the log-odds ratios (OR) of disease risk for a unit increment in X. Candidate gene G is coded as 0, 1, or 2 for the number of variant alleles under the log-additive model. This model has been shown performing well even when the underlying true genetic model is not log-additive [Schaid, 1996]. The model also accommodates other codings for G such as a binary variable under a dominant or recessive model or two indicator variables with one for carrying one variant allele and the other for carrying two variant alleles under an unrestricted model. The genotypes for the parents of cases are denoted by $G_p = (G_f, G_m)^T$.

The data can be partitioned into two components, those from case-unrelated controls and those from case-parent triads, each of which contributes to a likelihood function. The product of the two likelihood functions gives

$$\mathcal{L}^{comb} = \mathcal{L}^{cc} \times \mathcal{L}^{fam}, \tag{2}$$

where $\mathcal{L}^{cc}$ is the likelihood function for cases and unrelated controls and $\mathcal{L}^{fam}$ is the likelihood function for family-based case-parent triads. The product of the two likelihood functions, $\mathcal{L}^{comb}$ where comb stands for 'combined', is not itself a likelihood function, however, because the same cases are used in both $\mathcal{L}^{cc}$ and $\mathcal{L}^{fam}$. We term $\mathcal{L}^{comb}$ in (2) a pseudo-likelihood function. Constructing $\mathcal{L}^{comb}$ is straightforward because each individual likelihood has been extensively studied, and the asymptotic properties of each are well known. However, deriving the asymptotic distribution for the maximum pseudo-likelihood estimators requires additional work because standard likelihood theory does not apply. First we give a brief review of $\mathcal{L}^{cc}$ and $\mathcal{L}^{fam}$ separately.

For case-unrelated control samples that are collected retrospectively, the proportions of cases and controls are predetermined by investigators. The random elements are the covariates collected on these cases and controls. The retrospective likelihood function therefore is a product of $P(X_i|D_i)$ over $i = 1, \ldots, I + J$ cases and controls. From Bayes' rule, $P(X_i|D_i) = P(D_i|X_i)P(X_i)/P(D_i)$, implying that a direct modeling of $P(X_i|D_i)$ would require estimating the distribution or parameters in the distribution of X_i, which often are not of main interest. To avoid estimating such a distribution or parameters, many researchers (e.g. Anderson [1972], Prentice and Pyke [1979] and Whittemore [1995]) have observed that the OR, $\exp(\beta)$, can be consistently estimated from a case-unrelated control sample as if it had been prospectively collected in a hypothetical population. In this hypothetical population, the baseline log-odds of disease risk is $\alpha^* = \alpha + \log\{\pi/(1 - \pi)\}$ with π being the proportion of cases among all cases and controls. The likelihood function $\mathcal{L}^{cc}$ can then be written as

$$\mathcal{L}^{cc} = \prod_{i=1}^{I+J} \left\{ \frac{\exp(\alpha^* + \beta X_i)}{1 + \exp(\alpha^* + \beta X_i)} \right\}^{D_i} \left\{ \frac{1}{1 + \exp(\alpha^* + \beta X_i)} \right\}^{1-D_i}.$$

Note that the logistic regression model with intercept α^* puts constraints on the distribution of X conditional on disease status. Nevertheless, Anderson [1972] and Prentice and Pyke [1979] have shown that despite this constraint, one may make statistical inference about $\hat{\beta}$ without any particular modifications as compared to the same likelihood function for prospectively collected data.

For the case-parent data, we follow Clayton and Jones [1999] and Cordell et al. [2004] and represent $\mathcal{L}^{fam}$ as

$$\mathcal{L}^{fam} = \prod_{i=1}^{I} P\left(G_i | D_i = 1, G_{ip}, Z_i\right).$$

From Bayes' rule,

$$P\left(G_i | D_i = 1, G_{ip}, Z_i\right) = \frac{P\left(D_i = 1 | G_i, Z_i\right) P\left(G_i | G_{ip}, Z_i\right)}{\sum_{G^* \in \mathcal{G}_{ip}} P\left(D_i = 1 | G^*, Z_i\right) P\left(G^* | G_{ip}, Z_i\right)}, \tag{3}$$

where $\mathcal{G}_{ip}$ is the set of offspring genotypes that are consistent with the parental genotypes G_{ip}. Assuming an equal probability of allele transmission conditional on parental genotypes and other risk factors Z_i, i.e., $P(G_i|G_{ip}, Z_i) = P(G^*|G_{ip}, Z_i)$ for any $G^* \in \mathcal{G}_{ip}$ and no parent-of-origin or maternal genotype effects, the likelihood function can be simplified to

$$\mathcal{L}^{fam} = \prod_{i=1}^{I} \frac{P\left(D_i = 1 | G_i, Z_i\right)}{\sum_{G^* \in \mathcal{G}_{ip}} P\left(D_i = 1 | G^*, Z_i\right)}.$$

This representation shows that the estimable parameters are the relative risks (RR) associated with G and Z, namely the ratio of $P(D = 1|G = g_1, Z = z_1)$ and $P(D = 1|G = g_0, Z = z_0)$ with g_0 and z_0 being the reference values and g_1 and z_1 being the high-risk allele and exposed for G and Z, respectively. Though algebraically related, the RR is not the same as the OR, $\exp(\beta)$. Considering a single covariate, the relationship between the two quantities under model (1) is

$$RR = \frac{\exp(\alpha + \beta)}{1 + \exp(\alpha + \beta)} \frac{1 + \exp(\alpha)}{\exp(\alpha)} = \frac{\exp(\beta)\{1 + \exp(\alpha)\}}{1 + \exp(\alpha)\exp(\beta)}.$$

This suggests that using $\mathcal{L}^{fam}$ alone cannot identify α and β uniquely. To allow for estimating the log-odds ratio β from $\mathcal{L}^{fam}$, we propose to fix α by assuming that the disease is rare, i.e. $\exp(\alpha) \approx 0$. In this situation, the relative risk RR $\approx \exp(\beta)$. For example, when the baseline disease probability is 0.7% ($\alpha = -5$), 1.8% ($\alpha = -4$), or 4.7% ($\alpha = -3$), an OR of 3 would correspond to RRs of 2.96, 2.90, and 2.74, respectively. The discrepancy between RR and OR is within 10% even when the baseline disease probability is as high as 4.7%. Additionally when there is no association between a covariate and disease risk, both the RR and OR are 1 regardless of the underlying disease prevalence. Note that Nagelkerke et al. [2004] and Epstein et al. [2005] also employed a rare disease assumption when using unrelated controls to estimate the allelic or mating type frequencies.

Assessing the effects of multiple unlinked markers on disease risk is trivial, as these markers can simply be incorporated as covariates in model (1). The same pseudo-likelihood function (2) can be used in obtaining $\hat{\beta}$, except that in the case-parent likelihood $\mathcal{L}^{fam}$, the denominator in equation (3) becomes a summation over all combinations of offspring genotypes that are consistent with parental genotypes at these marker loci.

The estimates ($\hat{\alpha}^*$, $\hat{\beta}$) can be obtained by maximizing the pseudo-likelihood function (2), and many statistical software programs e.g. R [R Development Core Team, 2006] have built-in routines to return parameter values that maximize (or minimize) an objective function such as $\mathcal{L}^{comb}$. Alternatively the estimates can be obtained by solving

$$\frac{\partial}{\partial(\alpha^*, \beta)} \log \mathcal{L}^{comb} = 0, \tag{4}$$

where

$$\frac{\partial}{\partial(\alpha^*, \beta)} \log \mathcal{L}^{comb} = S(\alpha^*, \beta) = \left\{ S^{(1)}(\alpha^*, \beta), S^{(2)}(\alpha^*, \beta) \right\}^T,$$

and

$$\begin{cases} S^{(1)}(\alpha^*, \beta) = \dfrac{\partial}{\partial \alpha^*} \log\left(\mathcal{L}^{cc}\right) \\ S^{(2)}(\alpha^*, \beta) = \dfrac{\partial}{\partial \beta} \log\left(\mathcal{L}^{fam}\right) + \dfrac{\partial}{\partial \beta} \log\left(\mathcal{L}^{cc}\right) \end{cases}$$

The Newton-Raphson algorithm can be used to iteratively solve the estimating equation (4). Following the law of large numbers and the central limit theorem [Billingsley, 1999], we showed that the solution to the estimating equation (4) was consistent for the true parameter values (see Appendix). The solution is also asymptotically normal with a 'sandwich' type variance that can be consistently estimated by

$$\hat{\Sigma} = \hat{\Sigma}_1^{-1} \hat{\Sigma}_0 \hat{\Sigma}_1^{-1},$$

where $\hat{\Sigma}_1^{-1}$ and $\hat{\Sigma}_0$ are given in the Appendix. The $\hat{\Sigma}_0$ accounts for the correlation of the two components of the likelihood function due to having data from the same cases in both components.

The proposed pseudo-likelihood function (2) is readily extended to allow for non-overlapping case-parent and case-control components without any modifications to the likelihood and the estimation procedure. The covariances i.e. the off-diagonal elements in Σ_0 is zero, and the variance is simply Σ_1^{-1}. In fact, the proposed pseudo-likelihood function becomes a likelihood function in this situation and the maximum likelihood estimator from this likelihood is asymptotically equivalent to the weighted average of estimates from individual components with weight being the inverse of their variances [e.g. Kazeem and Farrall, 2005]. In this case, it is also very similar to the Dudbridge's likelihood approach [Dudbridge, 2007] for which the software can be downloaded at http://www.mrc-bsu.cam.ac.uk/personal/frank/software/unphased/. The proposed pseudo-likelihood function is also readily extended to allow for singleton cases, that is, cases whose parental genotype data are not collected. Singleton cases contribute to the case-control likelihood $\mathcal{L}^{cc}$ but not $\mathcal{L}^{fam}$. For such cases the covariance in Σ_0 between case-control and case-parent components would be due only to the subset of shared cases in both components. The asymptotic theory in the Appendix includes both situations.

Incorporating Missing Genotypes in Parents

For some cases, one or both parents may be unavailable for genotyping. When both parents lack genotypes, we treat these cases as singletons and use them only in the case-unrelated control likelihood $\mathcal{L}^{cc}$. It is unlikely that such singleton cases contribute much information to $\mathcal{L}^{fam}$ when the association estimate in $\mathcal{L}^{fam}$ derives from the comparison between the transmitted and non-transmitted genotypes. In this section, therefore, we consider only cases that lack genotypes for one parent but not both. Many approaches have been proposed to allow for incomplete triads in family-based association studies (e.g. Clayton [1999], Weinberg [1999], Rabinowitz and Laird [2000], Rabinowitz [2002], Allen et al. [2003], Chen [2004], and Kistner and Weinberg [2005]). However, all of these methods deal with one marker or haplotype, and not all are readily generalizable to studies with many unlinked markers. One promising approach is to use multiple imputation [Little and Rubin, 2002] to impute the missing genotypes. This has been proposed for quantitative traits [Kistner and Weinberg, 2005], where complete triads are used to estimate the probability of each of a triad's possible genotypic configurations given the phenotype of the offspring; missing genotypes are then imputed based on these posterior probabilities given the observed data. More recently, Croiseau et al. [2007] proposed a Bayesian-based multiple imputation approach in family-based association studies. In this approach, a Dirichlet prior distribution was assumed for haplotype (genotype) frequencies and the posterior probabilities and imputed data sets were obtained via a data augmentation algorithm.

We modify the Kistner and Weinberg approach by using a combination of the EM algorithm [Dempster et al., 1977] and multiple imputation [Little and Rubin, 2002] to: (1) estimate the triad's genotypic configurations for each marker by using case-parents, and (2) fill in the missing genotypes based on these probabilities one marker at a time. We assume that the parental genotypes are 'missing at random' [Little and Rubin, 2002] such that the genotype distribution among genotyped parents is the same as that among the ungenotyped, or missing, parents. Note that this approach is different from the Kistner and Weinberg approach, in which only complete triads were used in calculating the joint probabilities of a triad. In our approach, we use both triads and dyads to estimate the frequencies of genotypic configurations by using the EM algorithm. The EM algorithm, though most efficient, does not generalize easily to multiple markers because it involves multiple levels of summation over unknown genotypes in calculating the likelihood function. To overcome this cumbersome manipulation, in the second step we propose to use a multiple imputation method to impute the missing genotypes based on estimates of the probability of triads' genotypic configurations obtained from the EM algorithm. Once missing genotypes are imputed, estimation procedures for $\hat{\beta}$ proposed earlier in this section are readily applied. Even though there are 10 parameters associated with the genetic configurations of triads for each marker that need to be estimated, the estimation occurs only at the missing data imputation stage. It therefore has less computational bearing on the estimation of $\hat{\beta}$ compared with methods that would require a simultaneous estimation of such parameters and β. It is worth noting that the frequency estimates of genotypic configuration are calcu-

Table 1. Parameters for the possible genotypic configurations of triads

Configuration	Parents		Offspring	Probability
M_1	0	0	0	μ_1
M_2	0	1	0	μ_2
M_3	0	1	1	μ_3
M_4	0	2	1	μ_4
M_5	1	1	0	μ_5
M_6	1	1	1	μ_6
M_7	1	1	2	μ_7
M_8	1	2	1	μ_8
M_9	1	2	2	μ_9
M_{10}	2	2	2	μ_{10}

In the 'Parents' and 'Offspring' columns, 0, 1, and 2 are the numbers of the copies of the variant allele carried by the given individual.

Table 2. Posterior probabilities of the genotypes for a missing parent conditional on the available parent and offspring

Offspring	Parent 1	Parent 2		
		0	1	2
0	0	$\frac{2\mu_1}{2\mu_1+\mu_2}$	$\frac{\mu_2}{2\mu_1+\mu_2}$	0
0	1	$\frac{\mu_2}{\mu_2+2\mu_5}$	$\frac{2\mu_5}{\mu_2+2\mu_5}$	0
1	0	0	$\frac{\mu_3}{\mu_3+\mu_4}$	$\frac{\mu_4}{\mu_3+\mu_4}$
1	1	$\frac{\mu_3}{\mu_3+2\mu_6+\mu_8}$	$\frac{2\mu_6}{\mu_3+2\mu_6+\mu_8}$	$\frac{\mu_8}{\mu_3+2\mu_6+\mu_8}$
1	2	$\frac{\mu_4}{\mu_4+\mu_8}$	$\frac{\mu_8}{\mu_4+\mu_8}$	0
2	1	0	$\frac{2\mu_7}{2\mu_7+\mu_9}$	$\frac{\mu_9}{2\mu_7+\mu_9}$
2	2	0	$\frac{\mu_9}{\mu_9+2\mu_{10}}$	$\frac{2\mu_{10}}{\mu_9+2\mu_{10}}$

We assume, with no loss of generality, that parent 2 is the parent missing genotype data.

lated using case-parents trios as by the hybrid study design only the parents of cases are collected and thus only the parents of cases, when missing, need to be imputed. This approach is also different from that of Croiseau et al.'s [2007] approach where we do not assume a prior distribution for genotype frequencies (a more detailed comparison is given in the Discussion Section).

Specifically, let $G_{it} = (G_{ip1}, G_{ip2}, G_i)$ denote the number of copies of the variant allele at a given locus carried by the first parent, the second parent, and the case in the ith triad. For each locus, there are 10 possible genotypic configurations $M_1, M_2, \ldots, M_{10}$ with frequencies denoted by the row vector $\mu = (\mu_1, \mu_2, \ldots, \mu_{10})$ (table 1). The first step is to estimate μ by using an EM algorithm that incorporates information from both triads and dyads. Without loss of generality, we assume that the genotypes for the second parent are unavailable and need to be imputed. In the E-step, the posterior probabilities of possible genotypes for the missing parent in each dyad are calculated conditional on the observed genotypes of the non-missing parent and the case at the current values of $\hat{\mu}$ (table 2). The $\hat{\mu}$ are then updated by a weighted average in the M-step. Briefly, let the first $i = 1, \ldots, n_1$ families be the triads, and the next $i = (n_1 + 1), \ldots, (n_1 + n_2)$ families be the dyads. Furthermore, let $\mathcal{G}(G_{ip1}, G_i)$ be the set of possible genotypes that the missing parent could have conditional on the genotypes for the available parent G_{ip1} and offspring G_i. Then the probability of M_j configuration can be estimated by

$$\hat{\mu}_j = \frac{n_1}{n_1+n_2}\frac{\sum_{i=1}^{n_1} I\left(G_{it}=M_j\right)}{n_1} + \frac{n_2}{n_1+n_2}\frac{\sum_{i=n_1+1}^{n_1+n_2}\sum_{G^*_{ip2}\in\mathcal{G}\left(G_{ip1},G_i\right)} I\left(G^*_{it}=M_j\right)P\left(G^*_{ip2}|G_{ip1},G_i\right)}{n_2},$$

where $G^*_{it} = (G_{ip1}, G^*_{ip2}, G_i)$. The estimate consists of two terms: the first term is the proportion of triads that have configuration M_j, and the second term is the proportion of dyads that could have configuration M_j weighted by its posterior probability. The overall estimate is a weighted average of triads and dyads with weights proportional to their sample sizes n_1 and n_2. We alternate between the E- and M- steps until $\hat{\mu}$ converges.

We then use the multiple imputation method to randomly sample a genotype from the multinomial distribution with posterior probabilities calculated according to table 2. The imputed genotypes along with all observed data are then treated as a complete dataset, which can be analyzed by using the estimation and inference procedure described in previous sections. To account for the variability of imputation, we create multiple complete datasets $b = 1, \ldots, B$ and obtain $\hat{\beta}_b$ and associated variance $\hat{\Sigma}_b$ for each of these datasets. The final estimate for β is

$$\bar{\beta}_B = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_b .$$

The variance of $\hat{\beta}$ is the sum of the average of within-imputation variance and the between imputation variance [Little and Rubin, 2002], given by

$$\sum_t = \frac{1}{B}\sum_{b=1}^{B}\hat{\Sigma}_b + \left(1+\frac{1}{B}\right)\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}_b-\bar{\beta}_B\right)'\left(\hat{\beta}_b-\bar{\beta}_B\right),$$

where $(1 + 1/B)$ is an adjustment for finite B. Depending on the proportions of triads and dyads, usually 3–10 imputations are adequate to achieve maximal efficiency [Rubin, 1987].

An Approach for Reducing Computation in the Calculation of $\mathcal{L}^{fam}$ in the Presence of Multiple Unlinked Markers

For case-parents likelihood, the denominator in equation (3) involves a large number of summation because the number of the combinations of untransmitted genotypes at different markers increases exponentially as the number of loci. Exhaustive searching for all these combinations is not difficult, but storing and analyzing all these combinations can take much computer space and may become prohibitive when the number of markers is large. In this section, we describe a simple approach that uses only a subset of these combinations for estimation.

The idea is as follows. It is well known that in population-based case-control studies, increasing the number of controls per case improves the efficiency of parameter estimates, however the improvement is not unlimited because of the number of cases is fixed [Breslow and Day, 1980]. If we intuitively consider all these combinations of untransmitted genotypes as 'pseudo-sibling controls', it is possible that there is a similar effect to case-control studies, that is, a much reduced subset of these combinations may be sufficient to achieve nearly full efficiency from using all combinations. If that is the case, we can create a subset of pseudo-sibling controls prior to the analysis and use the subset to identify disease associated polymorphisms, for example, in a large candidate gene association study in which the joint effects of markers from different gene are considered.

The most straightforward approach for choosing such subset is by random selection. We can randomly select a fixed number of combinations from the set of all possible genotypes conditional on the parental genotypes, excluding the ones that are actually transmitted to the diseased offspring. This strategy, although valid, may not be efficient in selecting 'controls', because the genotypes of these controls can be quite similar to each other and the affected offspring genotypes.

To overcome this over-matching, we pair up the alleles that are not transmitted to the diseased offspring and use this combination as the control genotypes. Since this combination is least similar to the observed offspring genotype, it is the most efficient choice if only one pseudo-control is chosen. Following this idea, an alternative strategy would be to choose paired controls such that for each randomly chosen pseudo-sibling control, the combination of alleles that are not transmitted to this pseudo-sibling control is also chosen. The reason is that in the conditional likelihood function for case-parents trios, the denominator is a summation of permutations in the matched set, and therefore the controls should be sampled according to the scheme of the first pair: the case and the counter-matching control (nontransmitted alleles to the case). In other words, the subsequent controls should be selected in pairs: for each randomly selected control, the nontransmitted alleles to this control should also be selected. This way we maximize the dissimilarity among the genotypes of these controls while still being able to yield consistent estimators of log(OR). Interestingly, this sampling strategy has a similar flavor to the 'counter-matching' proposed by Langholz et al. [1995] in the nested case-control study design. In that setting, the authors proposed to counter-match the cases in the exposure values so that the efficiency of parameter estimates can be maximized.

It is worth noting that sampling a subset of controls is useful in achieving the (nearly) maximum efficiency with fewer number of controls when the carrier frequencies are relatively common, for example, under a dominant or additive model. For recessive models or the frequencies of carriers are low, sampling only a subset of controls, whether random or countermatched, will likely lose substantial efficiency and in some cases may even result in unstable estimates.

Simulation Studies

We conducted simulation experiments to examine the finite sample properties of the proposed estimators and inference procedures and to compare the performance when data were incorporated from (1) case-parent triads only, (2) from cases and unrelated controls, or (3) combined from case-parent triads and unrelated controls. For each experiment, we considered three unlinked loci, G_1, G_2, and G_3, and one binary environmental covariate E. We varied the allele frequencies and $P(E = 1)$. While the estimates and their variances varied from one set of parameter values to another, the relative performance of the proposed methods was similar. Therefore we present results only for an allele frequency of 0.1, with the heterozygote OR equal to the homozygote OR for each variant (i.e. a dominant model), and $P(E) = 0.2$. We also compared the proposed method with the Epstein and Nagelkerke maximum likelihood (ML) approach.

We examined the performance of the proposed estimators in terms of bias and efficiency under two general models: gene-environment interaction and gene-gene interaction. The models were:

(1) $G \times E$:

$$\text{logit}\{P(D = 1|G, E)\} = \alpha + \beta_1 G_1 + \beta_2 G_2 + \beta_3 G_3 + \beta_4 E + \beta_5 G_1 \times E.$$

(2) $G \times G$:

$$\text{logit}\{P(D = 1|G, E)\} = \alpha + \beta_1 G_1 + \beta_2 G_2 + \beta_3 G_3 + \beta_4 E + \beta_5 G_1 \times G_2.$$

Under each model we simulated 1000 datasets, each consisting of an equal number of cases and controls in addition to the parents of cases. We considered three sample sizes, $n_1 = n_2 = 250$, 375, and 500, where n_1 and n_2 are the number of cases and controls, respectively. We chose these particular sample sizes for two reasons: (1) the sample sizes were comparable to our testicular cancer dataset (see below); and (2) they allow a comparison of the efficiency of parameter estimates derived from triads, case-unrelated controls, or the combined sample if the total number of individuals genotyped is fixed. For example, the number of genotyped individuals for 250 cases and their parents is equivalent to that for 375 cases and 375 unrelated controls, and the total of 250 cases and their parents and 250 unrelated controls leads to the same number of genotyped individuals as 500 cases and 500 unrelated controls.

We set $\alpha = -3$ so that the baseline disease risk was approximately 4.7%, yielding a difference between the OR and RR close to what might be considered the maximum tolerable limit, <10%. In other words, if the estimates appear to be reasonably close when $\alpha = -3$, the difference in OR and RR estimates would be even smaller with rarer diseases. The true values for β's were 0.405, 0.405, 0.405, 0.693, and 1.100, respectively, yielding ORs of 1.5 for the main effects of candidate genes, an OR of 2 for the environmental covariate, and an OR of 3 for the interaction. The corre-

Table 3. Summary statistics for estimates under a gene-environment interaction model, $\text{logit}(\Pr(D = 1 | G_1, G_2, G_3, E)) = a + b_1G_1 + b_2G_2 + b_3G_3 + b_4E + b_5G_1 \times E$

Type	$b_1 = 0.405$	$b_2 = 0.405$	$b_3 = 0.405$	$b_4 = 0.693$	$b_5 = 1.100$
$n_1 = 250, n_2 = 250$					
Triads	0.386 (0.295, 0.279, 94.1)	0.340 (0.214, 0.219, 94.3)	0.354 (0.224, 0.218, 94.4)	–	0.834 (0.498, 0.463, 88.9)
CC	0.410 (0.270, 0.263, 94.1)	0.408 (0.246, 0.231, 93.0)	0.404 (0.230, 0.230, 95.2)	0.710 (0.246, 0.247, 95.1)	1.190 (1.020, 0.559, 97.0)
Combined	0.391 (0.243, 0.231, 94.8)	0.367 (0.195, 0.189, 93.5)	0.372 (0.192, 0.189, 94.5)	0.743 (0.241, 0.240, 94.9)	0.932 (0.418, 0.400, 91.6)
$n_1 = 375, n_2 = 375$					
Triads	0.377 (0.231, 0.227, 94.4)	0.349 (0.174, 0.178, 94.2)	0.355 (0.184, 0.178, 93.0)	–	0.822 (0.382, 0.373, 87.5)
CC	0.407 (0.208, 0.214, 95.6)	0.411 (0.187, 0.187, 94.8)	0.399 (0.188, 0.187, 94.7)	0.698 (0.195, 0.200, 95.2)	1.160 (0.471, 0.451, 95.3)
Combined	0.388 (0.187, 0.188, 94.7)	0.374 (0.151, 0.154, 94.0)	0.372 (0.159, 0.154, 93.6)	0.737 (0.188, 0.196, 95.4)	0.937 (0.341, 0.326, 91.0)
$n_1 = 500, n_2 = 500$					
Triads	0.378 (0.198, 0.195, 94.5)	0.348 (0.150, 0.154, 94.8)	0.357 (0.159, 0.154, 92.6)	–	0.813 (0.331, 0.321, 82.9)
CC	0.406 (0.187, 0.185, 94.3)	0.413 (0.163, 0.162, 95.4)	0.411 (0.161, 0.162, 95.6)	0.704 (0.172, 0.173, 95.0)	1.110 (0.389, 0.384, 95.9)
Combined	0.389 (0.164, 0.163, 94.9)	0.375 (0.134, 0.133, 94.5)	0.379 (0.135, 0.133, 93.9)	0.739 (0.166, 0.169, 94.7)	0.914 (0.281, 0.280, 89.1)

The results are based on 1000 simulated datasets, each consisting of n_1 case-parent triads and n_2 unrelated controls. 'Triads' refers to using only triad data for estimation; 'CC' refers to using only cases (not parents) and unrelated controls; 'Combined' refers to using both case-parents and unrelated controls. Each entry lists the mean estimate (standard deviation of the estimates, mean of the estimated standard errors, 95% coverage probability) over the 1000 simulated datasets. – = The main effect of the environmental covariate is unestimable from triads-only.

Table 4. Summary statistics for estimates under a gene-gene interaction model, $\text{logit}(\Pr(D = 1 | G_1, G_2, G_3, E)) = a + b_1G_1 + b_2G_2 + b_3G_3 + b_4E + \beta_5G_1 \times G_2$

Type	$b_1 = 0.405$	$b_2 = 0.405$	$b_3 = 0.405$	$b_4 = 0.693$	$b_5 = 1.100$
$n_1 = 250, n_2 = 250$					
Triads	0.372 (0.255, 0.251, 94.4)	0.377 (0.264, 0.251, 94.3)	0.363 (0.214, 0.218, 95.0)	–	0.863 (0.382, 0.369, 89.1)
CC	0.409 (0.260, 0.258, 95.6)	0.411 (0.264, 0.258, 94.8)	0.421 (0.234, 0.229, 94.9)	0.701 (0.216, 0.219, 95.2)	1.210 (1.030, 0.567, 95.1)
Combined	0.393 (0.219, 0.219, 95.0)	0.397 (0.228, 0.219, 94.5)	0.384 (0.186, 0.188, 95.3)	0.694 (0.213, 0.217, 95.7)	0.926 (0.375, 0.359, 98.0)
$n_1 = 375, n_2 = 375$					
Triads	0.379 (0.200, 0.204, 95.1)	0.377 (0.213, 0.204, 94.6)	0.362 (0.179, 0.178, 95.0)	–	0.857 (0.307, 0.300, 85.7)
CC	0.402 (0.208, 0.209, 95.3)	0.397 (0.217, 0.210, 95.2)	0.412 (0.176, 0.186, 95.5)	0.702 (0.177, 0.178, 94.7)	1.160 (0.456, 0.454, 95.6)
Combined	0.395 (0.173, 0.178, 96.0)	0.392 (0.187, 0.178, 94.3)	0.381 (0.147, 0.153, 96.1)	0.696 (0.176, 0.177, 94.7)	0.926 (0.299, 0.292, 90.0)
$n_1 = 500, n_2 = 500$					
Triads	0.375 (0.170, 0.176, 95.7)	0.377 (0.181, 0.177, 94.6)	0.363 (0.157, 0.154, 93.6)	–	0.856 (0.258, 0.260, 81.0)
CC	0.410 (0.174, 0.181, 95.8)	0.402 (0.182, 0.181, 94.2)	0.407 (0.161, 0.161, 94.7)	0.693 (0.150, 0.154, 96.2)	1.130 (0.395, 0.390, 95.6)
Combined	0.399 (0.147, 0.154, 96.4)	0.395 (0.156, 0.154, 94.0)	0.381 (0.133, 0.133, 94.9)	0.688 (0.149, 0.152, 96.1)	0.918 (0.250, 0.252, 88.2)

The results are based on 1000 simulated datasets, each consisting of n_1 case-parent triads and n_2 unrelated controls. 'Triads' refers to using only triad data for estimation; 'CC' refers to using only cases (not parents) and unrelated controls; 'Combined' refers to using both case-parents and unrelated controls. Each entry lists the mean estimate (standard deviation of the estimates, mean of the estimated standard error, 95% coverage probability) over the 1000 simulated datasets. – = The main effect of the environmental covariate is unestimable from triads-only.

sponding log(RR)s are 0.378, 0.352, 0.352, 0.635, and 0.802 under the gene-environment interaction model, and 0.373, 0.373, 0.355, 0.609, and 0.842 under the gene-gene interaction model.

Performance of Estimators Under Gene-Environment and Gene-Gene Interaction Models

The estimates from case-unrelated controls under the gene-environment interaction model (table 3) and gene-gene interaction (table 4) model are essentially unbiased, and the coverage probabilities maintain the 95% nominal levels. The estimates derived from triads only are unbiased in log(RR) and as expected, differed slightly for log(OR) because log(OR) and log(RR) are generally different though the difference is negligible if the disease is rare. Using the log(OR) as the true parameter values, the estimates from triads have lower than nominal 95% coverage probabilities, especially for gene-gene and gene-environment interaction effects. The combined analysis yields an estimate that is between those from the case-unrelated controls and triads analysis, and the variance estimators are smallest as compared with those derived from either triads alone or case-unrelated controls. The observed efficiency gain in combining both case-parents and unrelated controls is consistent with earlier findings reported by Epstein et al. [2005]. The proposed variance estimators also work well in finite sample sizes, as the means of the standard error es-

Table 5. Summary statistics for parameter estimates derived from case-parent triad data with genotypes missing for one parent

	$\beta_1 = 0.405$	$\beta_2 = 0.405$	$\beta_3 = 0.405$	$\beta_5 = 1.100$
Gene-environment interaction model				
Full	0.391 (0.281, 0.280)	0.355 (0.213, 0.218)	0.350 (0.216, 0.218)	0.828 (0.469, 0.464)
Complete	0.404 (0.425, 0.455)	0.350 (0.346, 0.349)	0.350 (0.351, 0.350)	1.060 (0.751, 0.767)
Imputation	0.398 (0.303, 0.309)	0.361 (0.244, 0.240)	0.357 (0.242, 0.240)	0.838 (0.442, 0.524)
Gene-gene interaction model				
Full	0.378 (0.255, 0.251)	0.379 (0.245, 0.251)	0.354 (0.225, 0.218)	0.852 (0.383, 0.369)
Complete	0.405 (0.409, 0.404)	0.401 (0.407, 0.405)	0.355 (0.321, 0.349)	0.852 (0.383, 0.369)
Imputation	0.384 (0.276, 0.274)	0.386 (0.275, 0.276)	0.356 (0.255, 0.240)	0.850 (0.346, 0.399)

The results are based on 1000 simulated datasets, each consisting of 100 triads and 150 dyads. 'Full' data analyses are based on all 250 triads with no missing genotypes; 'Complete' data analyses use only the 100 triads with genotypes available for both parents; 'Imputation' data analyses are based on 100 triads and 150 dyads by using a multiple imputation method. Each entry lists the mean estimate (standard deviation of the estimates, mean of the estimated standard errors) over the 1000 simulated datasets.

The main effect of the environmental covariate β_4 is unestimable from triads-only and thus omitted.

timates are very close to the standard deviation of estimates over the 1000 simulated datasets.

When the number of genotyped individuals is fixed under the gene-environment interaction model, the case-unrelated control design is more efficient in estimating all main effects and interaction effects than designs with either triads alone or combined data. When estimating the gene-gene interaction effect, however, using the triads and combined data appears to offer substantially greater efficiency compared with the case-unrelated control approach. Since the estimates from triads data are lower than that from CC data and it may result in smaller variances, we also compared the coefficients of variation, a normalized measure of variance. Interestingly the gain observed in efficiency measured by the inverse of variance estimates is diminished when using the coefficient of variation estimates.

Evaluation of Multiple Imputation Procedure for Missing Genotypes in Parents

We evaluated the effectiveness of the proposed multiple imputation procedure. For each dataset, we generated 250 case-parents triads and dyads. Among these, there were 150 dyads for which parental genotypes were missing at all three loci G_1, G_2, and G_3. We compared three methods: (1) the full-data analysis, for which genotypes were available for all 250 triads; (2) complete-data analysis using the 100 families that had genotypes for both parents; and (3) imputed data analysis by using the multiple imputation method proposed earlier in this paper. Five imputations were generated for each dataset. The results were based on 1000 simulated datasets. In general, the estimates are comparable for all three sets of analysis, however, the variances are quite different (table 5). The full-data analysis is the most efficient, whereas the complete-data analysis is the least. The multiple imputation method recovers most of the efficiency lost due to missing genotypes. The standard error estimators, as measured by the means of the estimated standard error, are very close to the standard deviations of the estimates, implying that the variance estimator $\hat{\Sigma}_t$ for the multiple imputation method works well in finite sample sizes.

Comparison of Various Strategies of Choosing Pseudo-Sibling Controls in $\mathcal{L}^{fam}$

For the counter-matching selection ('counter'), we considered three different values for the number of controls selected: (a) 1, only the paired non-transmitted alleles from parents is chosen; (b) 5, that is, in addition to (a), we randomly choose 2 controls and their complements; (c) 9, which is same as (b) except now that we randomly choose 4 controls and their complements. To make the number of randomly selected controls comparable to the 'counter' strategy, we choose 1, 5, and 9 controls for the random selection strategy, respectively.

A total of 1000 simulated datasets, each consist of 500 case-parents triads, were used for comparing the bias and relative efficiencies of these strategies against the full pseudo-sibling matched set for the main effect and interaction effect. The relative efficiency is defined as the inverse of the variance estimator for a sampling strategy divided by the inverse of the variance estimator for the full pseudo-sibling matched set. The counter-matching strategy generally is much more efficient than the random-selection one under both gene-environment and gene-gene interaction models (table 6). For example, under the gene-environment interaction model, using 9 controls or much fewer in the counter-matching strategy would achieve nearly 100% of the efficiency by using the full set of pseudo-sibling controls, which, in the case of three genes, has $4^3 - 1 = 63$ controls. The efficiency gain for counter-matching strategies compared to random-matching strategies for gene-gene interaction effects is not as much for gene-environment interaction effects. This is partly due to the fact that maximizing the discordance of genotypes among pseudo sibling controls would actually to some extent penalize the number of individuals who carry both high risk alleles. As a result, the counter-matching strategy may not gain as much efficiency for gene-gene interactions as for gene-environment interactions compared with the random strategy. Note that the main effect of environmental covariate β_4 is not estimable from case-parents data.

Table 6. Relative efficiency of parameter estimates under both the gene-environment interaction and gene-gene interaction models for case-parents triads

Scheme	β_1 = 0.405	β_2 = 0.405	β_3 = 0.405	β_4 = 0.693	β_5 = 1.100
Gene-environment interaction model					
Full	1.00	1.00	1.00	–	1.00
1 counter	0.95	0.90	0.94	–	0.96
5 counter	0.98	0.96	0.98	–	0.99
9 counter	0.98	0.97	1.00	–	1.00
1 random	0.52	0.51	0.48	–	0.48
5 random	0.77	0.71	0.75	–	0.80
9 random	0.85	0.84	0.84	–	0.86
Gene-gene interaction model					
Full	1.00	1.00	1.00	–	1.00
1 counter	0.90	0.93	0.88	–	0.46
5 counter	0.97	0.97	0.94	–	0.77
9 counter	0.99	0.99	0.96	–	0.95
1 random	0.47	0.49	0.49	–	0.40
5 random	0.75	0.80	0.73	–	0.72
9 random	0.87	0.82	0.81	–	0.79

The results are based on 1000 simulated datasets, each consist of 500 case-parents triads. The main effect of the environmental covariate β_4 is unestimable from triads-only and thus omitted.

Comparison with the Epstein and Nagelkerke Estimators

We compared the performance of the proposed pseudo-likelihood to the Epstein and Nagelkerke methods. The data were generated under the logistic regression model (1) with only one locus, as the Epstein et al. and the Nagelkerke ML methods only deal with one locus at a time. We simulated a dominant locus with minor allele frequency values of 0.1, 0.2, or 0.3. The intercept α was –1 or –3, yielding the disease risk of approximately 0.27 or 0.05, respectively, among non-carriers of the high risk allele. Two values of β were considered, 0 for no effect and log(2) = 0.693 for an OR of 2. This gave a total of $12 = 3 \times 2 \times 2$ combinations of parameter values. For each combination, we generated 1000 simulated datasets, each consisting of 500 cases and their parents and 500 controls. The Epstein ML estimators were obtained using the software downloaded from http://www.genetics.emory.edu/labs/epstein/software/chaplin/.

The standard errors of $\hat{\beta}$ are essentially the same for the proposed and the Epstein methods, as are the mean point estimates for the analyses of combined data (table 7). For both methods, the combined analysis is the most efficient as compared with using either triads or case-unrelated controls in the proposed method or as compared with using parents' data in the Epstein method. When $\beta = 0$, all estimates are unbiased. When $\beta = 0.693$, the case-control likelihood for the proposed method yields an unbiased estimate for β for all situations. The triads-only data yield estimates of the RR and are therefore smaller than the true values of log(OR) because the true value of the RR is itself smaller than the log(OR). The magnitude of the difference is proportional to the disease prevalence. The estimates derived from the combined analysis are essentially a weighted average of the estimates derived from either data configuration alone. For both triads-only and combined-data analyses, the Epstein estimators and the proposed estimators give highly comparable results. The $\hat{\beta}$ derived from the parents of cases for the Epstein method, on the other hand, are biased away from the null. This is because of the violation of the underlying rare disease assumption, which leads to biased estimates of mating type frequencies and, consequently, an upward bias in $\hat{\beta}$ for the parents of cases. The extent of bias is greatly reduced when $\alpha = -3$. The Nagelkerke approach which assumes the Hardy-Weinberg equilibrium and random mating has comparable efficiency with the Epstein or the proposed approach for most scenarios except when the allele frequency is 0.5 where there appears to be a very modest gain in efficiency for the Nagelkerke approach.

An Example from a Study of Testicular Cancer

Testicular germ cell cancers (TGCC) occur most commonly in men 20–44 years of age. There is strong familial aggregation [Dieckmann and Pichlmeier, 1997], but aside from white race, a history of undescended testes, and possibly taller height [Richiardi et al., 2007], risk factors have not been established for these malignancies. Several lines of evidence point to a role for steroid hormones and/or growth factors in the somatomedin pathway [Swerdlow, 2003; Zavos et al., 2004], but to date attempts to identify lifestyle or medical characteristics as risk factors that reflect these pathways have not been successful. We therefore have initiated a research program to determine whether polymorphisms in genes from these pathways are associated with TGCC risk. A combined case-unrelated control and case-parent triad design is being employed, based on a single group of incident TGCC cases identified through the Seattle-Puget Sound Surveillance, Epidemiology, and End Results cancer registry [Hankey et al., 1999]. Population-based controls from the same region are identified and recruited using random digit telephone dialing [Harlow and Davis, 1988; Hartge et al., 1984]. Details regarding case and parent eligibility, recruitment success, data and specimen collection, and genotyping of 6 polymorphisms have been previously reported [Starr et al., 2005].

To illustrate the proposed method, we analyzed data on the 1004 non-Hispanic Caucasians in our study on whom genotyping has been completed: 228 cases, 257 parents, and 519 unrelated controls. We had genotypes for both parents of 106 cases, for only one parent of 45 cases, and for neither parent of 77 cases. The specific polymorphisms examined included three SNPs (rs274057, *CYP3A4* A-392G; rs2665802, *GH1* T1663A; rs2854744,

Table 7. Comparison of bias and efficiency between the proposed pseudo- likelihood approach, the maximum likelihood method [Epstein et al., 2005], and the maximum likelihood approach that assumes the Hardy-Weinberg equilibrium and random mating [Nagelkerke et al., 2004]

p	α	β	Proposed method			Epstein			Nagelkerke combined
			triads	case-control	combined	triads	parents	combined	
0.1	–1	0	0.006 (0.159)	0.004 (0.161)	0.003 (0.140)	0.007 (0.161)	–0.011 (0.287)	0.003 (0.140)	0.003 (0.139)
		0.693	0.457 (0.153)	0.699 (0.150)	0.572 (0.127)	0.457 (0.153)	0.851 (0.222)	0.570 (0.130)	0.560 (0.124)
	–3	0	–0.001 (0.163)	–0.001 (0.162)	–0.002 (0.140)	0.001 (0.166)	–0.012 (0.298)	–0.002 (0.141)	–0.002 (0.138)
		0.693	0.660 (0.152)	0.706 (0.159)	0.681 (0.130)	0.660 (0.152)	0.729 (0.240)	0.680 (0.131)	0.676 (0.127)
0.2	–1	0	0.005 (0.129)	0.005 (0.134)	0.004 (0.115)	0.005 (0.129)	0.002 (0.245)	0.005 (0.116)	0.004 (0.113)
		0.693	0.453 (0.130)	0.688 (0.131)	0.568 (0.113)	0.453 (0.130)	0.876 (0.207)	0.564 (0.115)	0.548 (0.108)
	–3	0	0.004 (0.126)	–0.001 (0.129)	0.001 (0.109)	0.004 (0.126)	–0.011 (0.242)	0.001 (0.109)	0.004 (0.108)
		0.693	0.655 (0.130)	0.692 (0.131)	0.673 (0.114)	0.655 (0.130)	0.722 (0.214)	0.673 (0.115)	0.669 (0.110)
0.5	–1	0	0.007 (0.136)	–0.003 (0.149)	0.003 (0.126)	0.007 (0.136)	0.001 (0.305)	0.006 (0.126)	0.004 (0.120)
		0.693	0.456 (0.148)	0.694 (0.152)	0.575 (0.136)	0.456 (0.148)	1.060 (0.353)	0.554 (0.143)	0.527 (0.129)
	–3	0	0.005 (0.144)	–0.001 (0.152)	0.003 (0.131)	0.005 (0.144)	–0.003 (0.305)	0.004 (0.131)	0.004 (0.124)
		0.693	0.661 (0.163)	0.700 (0.168)	0.681 (0.150)	0.661 (0.163)	0.762 (0.386)	0.676 (0.152)	0.670 (0.142)

Each entry lists the mean parameter estimate (standard deviation of the estimates) for over the 1,000 simulated datasets. Each dataset consists of 500 cases and 500 controls.

Table 8. Log-odds ratio (OR) estimates (95% confidence intervals, p values) of the association between TGCC risk and variant alleles at four polymorphic loci under a log-additive model, by type of subjects included in the model

Subjects in model	*CYP3A4* G (vs. A)	*IGF1* 19 repeats (vs. others)	*GH1* A (vs. T)	*IGFBP3* C (vs. A)
Triads	1.47 (0.21–2.72, 0.022)	–0.25 (–0.73 to 0.21, 0.28)	0.06 (–0.39 to 0.50, 0.81)	0.33 (–0.08 to 0.74, 0.11)
Triads and Dyads	1.26 (0.19–2.34, 0.022)	–0.12 (–0.50 to 0.27, 0.55)	0.12 (–0.28 to 0.51, 0.56)	0.27 (–0.06 to 0.60, 0.11)
Cases and Controls	0.63 (0.10–1.18, 0.024)	–0.03 (–0.26 to 0.20, 0.79)	–0.10 (–0.32 to 0.12, 0.38)	0.07 (–0.14 to 0.29, 0.50)
Triads, Cases and Controls	0.79 (0.26–1.31, 0.003)	–0.08 (–0.32 to 0.16, 0.51)	–0.07 (–0.29 to 0.16, 0.55)	0.13 (–0.09 to 0.35, 0.24)
Triads, Dyads, Cases and Controls	0.78 (0.26–1.31, 0.003	–0.06 (–0.30 to 0.18, 0.64)	–0.04 (–0.27 to 0.19, 0.73)	0.13 (–0.08 to 0.34, 0.22)

Due to missing genotypes, for *CYP3A4* there are 515 unrelated controls and 220 cases, the latter corresponding to 95 triads (16 informative in estimating log OR), 46 dyads, and 79 without any parental genotypes. For *IGF1* there are 513 controls and 216 cases, the latter corresponding to 87 triads (57 informative), 52 dyads, and 77 without parental genotypes. For *GH1*, there are 504 controls and 216 cases, the latter corresponding to 89 triads (62 informative), 49 dyads, and 78 without parental genotypes. For *IGFBP3*, there are 515 controls and 218 cases, the latter corresponding to 88 triads (69 informative), 52 dyads, and 78 without parental genotypes.

IGFBP3 A-202C) and one microsatellite polymorphism in *IGF1* ($(CA)_n$ located at approximately –940 bp relative to the transcription start site). Log-additive genetic models were used in assessing the effects of these polymorphisms on TGCC risk.

Individuals who carried at least one copy of variant *CYP3A4 G* allele were at increased risk for developing TGCC in both the case-parent and case-unrelated control analyses, though the magnitude of log-odds ratios varied approximately two-fold between these analyses (table 8). The log-odds ratio estimate for triads was 1.47 with 95% confidence interval (CI) 0.21–2.72, and for case-unrelated controls is 0.63 (95% CI 0.10–1.18). Combining triads and unrelated controls yielded $\hat{\beta} = 0.78$ and gave narrower 95% CI for an increased risk for the *CYP3A4 G* allele carriers compared with estimates from using either triads or case-unrelated controls. The p values calculated based on the Wald test statistic are 0.022, 0.024, and 0.003 for triads, case-unrelated controls, and combined, respectively. Adding dyads to both triads and unrelated controls did not greatly change the log-odds ratio estimate nor the 95% confidence interval.

For the other three polymorphisms, one each in *IGF1*, *GH1* and *IGFBP3*, we observed little evidence of associa-

Table 9. Log-odds ratio estimates (95% confidence intervals, p values) for gene-gene interactions between *GH1* and *IGF1* and between *IGFBP3* and *IGF1* under a log-additive model, by type of subjects included in the model

Subjects in model	Gene-gene interaction		
	GH1	*IGF1*	*GH1*IGF1*
Triads	–0.14 (–0.76 to 0.47, 0.64)	–0.46 (–1.09 to 0.18, 0.16)	0.22 (–0.34 to 0.77, 0.44)
Triads and Dyads	–0.10 (–0.60 to 0.40, 0.70)	–0.30 (–0.80 to 0.20, 0.24)	0.24 (–0.21 to 0.69, 0.30)
Cases and Controls	–0.01 (–0.32 to 0.31, 0.98)	0.06 (–0.28 to 0.40, 0.74)	–0.13 (–0.46 to 0.21,0.46)
Triads and Controls	–0.05 (–0.36 to 0.27, 0.76)	–0.07 (–0.41 to 0.27, 0.69)	–0.03 (–0.36 to 0.29, 0.84)
Triads, Dyads and Controls	–0.05 (–0.36 to 0.26, 0.75)	–0.07 (–0.41 to 0.26, 0.67)	0.01 (–0.32 to 0.34, 0.95)
	IGFBP3	*IGF1*	*IGFBP3*IGF1*
Triads	0.70 (0.14 to 1.26, 0.01)	0.19 (–0.59 to 0.98, 0.63)	–0.33 (–0.89 to 0.23, 0.25)
Triads and Dyads	0.41 (–0.05 to 0.87, 0.09)	0.08 (–0.58 to 0.74, 0.81)	–0.09 (–0.53 to 0.35, 0.67)
Cases and Controls	0.14 (–0.18 to 0.45, 0.39)	0.06 (–0.36 to 0.47, 0.78)	–0.09 (–0.40 to 0.24, 0.58)
Triads, Cases and Controls	0.27 (–0.04 to 0.57, 0.09)	0.10 (–0.29 to 0.50, 0.61)	–0.16 (–0.47 to 0.16, 0.32)
Triads, Dyads, Cases and Controls	0.22 (–0.09 to 0.52, 0.16)	0.07 (–0.34 to 0.48, 0.73)	–0.09 (–0.40 to 0.22, 0.55)

tions with TGCC risk, whether in analyses using the case-parent triads, the cases and unrelated controls, or both. Because the proteins encoded by *IGFBP3* and *GH1* modulate the levels and bioavailability of *IGF1*, respectively, we also examined interactions between the *IGF1* and *GH1* polymorphisms and between *IGFBP3* and *IGF1* polymorphisms (table 9). There was little evidence to suggest that *GH1* or *IGFBP3* genotypes modified the association between the *IGF1* polymorphism and TGCC risk.

Discussion

We have described a general approach to combining data from cases, parents, and unrelated controls that can easily accommodate multiple genetic loci and other risk factors. The main and interactive effects of gene-gene and gene-environment are modeled in a straightforward fashion by including genotypes, environmental covariates, and their products in the model. Additionally, in contrast to other hybrid data approaches, the proposed method is not encumbered by the need to estimate nuisance parameters, which can become quite numerous when more than one locus is considered and when the assumptions of the random mating and Hardy-Weinberg equilibrium are violated [Epstein et al., 2005]. The pseudo-likelihood approach can lose efficiency compared to the maximum likelihood approach. However, in our simulation study, we observed little efficiency loss compared with the likelihood-based approaches, except perhaps when the allele frequency is very common, say 0.5 and when the random mating and Hardy-Weinberg equilibrium are assumed.

While interpretation of the log(OR) estimates depends on an underlying rare disease assumption, this assumption is satisfied for many common diseases such as many cancers, type I diabetes, and virtually all birth defects. However, for many traits that exist on a continuum of severity or symptoms, such as mental disorders, an inclusive definition may result in a high prevalence of the disease. In this situation, α may be fixed by borrowing information from external data sources, and the RR expressed as a function of α and the OR, $\exp(\beta)$. The estimation and inference procedures follow similarly to the proposed approach. If the baseline disease risk is of interest, one can, in fact, obtain the estimates $\hat{\alpha}$ and $\hat{\beta}$ from the pseudo-likelihood $\mathcal{L}^{comb}$ in (2). Essentially β is estimated from the case-unrelated sample, whereas α is estimated from the families. We conducted a simulation study and found that $\hat{\alpha}$ and $\hat{\beta}$ are generally unbiased though with large variance, and particularly large variance for $\hat{\alpha}$ (results not shown).

One key assumption for valid estimation and inference when combining cases, case-parents, and unrelated controls is that ORs are homogeneous in case-unrelated controls and case-parents. It is possible to test whether the ORs derived from the two study components are the same. If there is no strong evidence for a difference in the ORs between the two components, the data may be com-

bined. This two-step strategy, proposed by Epstein et al. [2005], will reduce the chance of inappropriately combining data. However, true heterogeneity may be missed if there is insufficient power to detect the inequality of estimates from the two components.

On the other hand, several possible sources of genuine heterogeneity may result in differences between estimates from various data sources. First, the populations from which various data sources are derived may be different, and the effects of genes or nongenetic risk factors may not be homogeneous across populations. Secondly, it has long been known that in non-linear regression models, such as logistic regression, omission of covariates related to disease risk but unrelated to the exposure of interest (in our example genotypes) will attenuate estimates of the genotype-disease association toward the null [Diggle et al., 1994, pp 137–142]. Such a difference does not represent bias, but rather indicates a true difference in the covariate-adjusted and unadjusted effect estimates. Case-parent data are implicitly adjusted for many possible risk factors, because the case and pseudo-sibling controls differ only at the genotypes of interest. In contrast, case-unrelated control data likely do not account for some risk factors, and the estimated effects of covariates in the model may appear to be attenuated in relation to those derived from case-parent data. While the magnitude of such true difference of estimates between analyses employing case-parents and case-unrelated controls varies depending on the heterogeneity among subjects and can be larger than 10%, this source of heterogeneity of estimates is often under-appreciated.

Thirdly, possible population structure in case-unrelated control samples may yield spurious associations between genetic polymorphisms and disease risk. Thomas and Witte [2002] have summarized strategies on how to minimize the impact of population stratification in case-unrelated control studies of candidate genes. These strategies (which are not all mutually exclusive) include selecting controls properly matched to cases' race/ethnicity, collecting race information in as much detail as possible, and inferring population stratification molecularly by using markers that are not associated with disease (e.g., Devlin and Roeder [1999]; Pritchard and Rosenberg [1999]; Satten et al. [2001]; and Zhang et al. [2003]). In contrast, family-based studies are robust to such biases. However, the robustness may not hold in the presence of missing genotypes because multiple imputation draws missing genotypes based on estimated genotype frequencies which would be biased if there were population stratification in the parents of cases. This is rather undesirable as robustness to population stratification is a primary advantage for family-based studies. We can use the same strategies as that for case-unrelated controls, that is, stratifying case families based on collected race information or inferred population structure from unrelated markers. Rabinowitz [2002] developed an innovative approach to this problem for family-based genetic association that accounts for population heterogeneity and misspecified population haplotype (genotype) frequencies. Further extensions of these robust methods to estimation were considered by Whittemore [2004], Allen et al. [2005] and Allen and Satten [2007]. The robust score functions proposed by Whittemore, Allen and colleagues can be used to replace the conditional likelihood-based score function and a sandwich variance estimator similar to that in the Appendix can be derived to account for the overlapping of the cases between case-parents and case-unrelated controls.

The results of our simulation experiments suggest that the efficiency of the case-parents design is the same as or slightly better than the case-unrelated control design for main effects, and substantially better for gene-gene or gene-environment interaction effects when the same number of cases is used and there is an equal number of controls. However, the triad design requires 50% more genotyping than the case-unrelated control design. For a fixed amount of genotyping regardless of design (triads, case-unrelated controls or a combination of both), the case-unrelated control design is most efficient in all situations except for gene-gene interactions, for which the triad design appears to yield the most efficient estimators. This observation was also reported by Witte et al. [1999]. In addition to concerns about efficiency, the triad design may be sensitive to genotyping errors or stochastic variation, particularly when the number of families with informative parents (i.e. parents heterozygous for one of the loci) is small despite having a reasonable number of triads in total. This is because each family is treated as a matched set, a random genotyping error or a stochastic variation in an otherwise informative family could affect the RR estimate, and this impact can be substantial when the number of informative parents is small.

Multiple imputation is a useful tool in dealing with missing data, in particular, when multiple genes and environmental risk factors are simultaneously considered for their effects on disease risk. In our approach, a modest number of complete data sets (usually 3–10) were generated by filling in the missing data elements with values generated according to their (posterior) probability distribution. Croiseau et al. [2007] proposed a Bayesian-

based approach where a prior Dirichlet distribution was assumed for population haplotype (genotype) frequencies. This allows for the uncertainty of haplotype frequencies accounted for in imputing missing genotypes. In contrast, we used the maximum likelihood estimator (MLE) of population genotype frequencies and filled the missing genotypes with values generated from a multinomial distribution with the MLE of genotype frequencies. The proposed approach does not make any assumption for genotype frequencies, however, it also does not account for the uncertainty of the estimates. From our simulation study it seems that there are few differences between the standard deviations of the estimates of regression parameters and the means of the standard error estimates, indicating the proposed procedure worked well with sample sizes considered. The variances of regression parameter estimates, however, may be underestimated if the sample sizes are much smaller and the MLE of genotypic frequencies have much greater variances. Further investigation on the performance of both approaches in both simulated and real data sets will be needed.

In candidate gene association studies, when markers from different genes are considered simultaneously for their effects on disease risk, methods proposed here are readily applicable. However, with emerging high throughput technologies, markers become increasingly dense and many of them are in LD or closely linked. In this situation, one may either treat markers as individual covariates and fit a regression model with interaction effects, or construct haplotypes from these markers and treat haplotypes as a covariate with multiple levels. Both approaches are applicable to case-control data, though the latter requires statistical reconstruction of haplotypes from unphased genotypes (e.g. Zhao et al. [2003] and Lin et al. [2005]). For case-parent data, however, the phases of markers need to be known in order to obtain the proper genotypic distribution for pseudo-sibling controls, regardless of whether the markers are treated as individual covariates or haplotypes. A simple approach is to use only one pseudo-sibling control, whose genotypes consist of non-transmitted alleles from parents to the offspring. This approach yields an unbiased estimator of β in the absence of missing genotypes. However, our current algorithm for handling missing genotypes would require modification to account for LD among markers. Alternatively, haplotypes can be reconstructed from case-parents (e.g. Clayton and Jones [1999]; Cordell et al. [2004]; Allen and Satten [2007]) independent of or jointly with unrelated controls. The pseudo-likelihood (2) can then be modified by summing over the unknown haplotypes and weighting each individual's contribution by the posterior probability of the individual's carrying each haplotype given the unphased genotypes.

The proposed method incorporates data from cases, parents of cases, and unrelated controls. It would be useful to extend the method to allow for other types of family members, and siblings in particular, as for many late-onset diseases the parents of cases often are not living at the time that recruitment is carried out. So far little work has been done in combining other types of family members with unrelated controls. This is partly due to the difficulty of handling possible residual correlation among family members even after adjusting for all the known risk factors, as well as the non-cohort ascertainment of families. Likelihood-based methods [Kraft and Thomas, 2000] for family data together with the pseudo-likelihood approach proposed here may be used for handling such a general data structure.

Large association studies that employ both case-unrelated control and family-based designs have been conducted for some diseases. An example is the National Cancer Institute-sponsored colon cancer family registry [Newcomb et al., 2007]. This international consortium includes six clinical centers, each of which has used a different study design involving colorectal cancer cases and one or more comparison groups (including several centers with both unrelated controls and cases' relatives). Since neither design is universally superior to the other, incorporating both types of controls in the consortium permits investigators to study a broad spectrum of allele frequencies, exposures, and effects on risk for colon cancers. Now with the increasing number of SNPs being genotyped in any given study, the search for common genes with modest effects or gene-gene or gene-environment interactions is intensifying, with a concomitant demand for large sample sizes in order to achieve adequate power. Combining samples, even those that have been collected under different designs, becomes an attractive option. The proposed method provides a general framework for combining data collected under different designs. Advantages to this approach include its ability to incorporate missing genotypes and data from many loci simultaneously. As we have mentioned, however, great care needs to be taken before combining samples from different sources because of various complicating issues such as population stratification and disease heterogeneity. There remains much work to be done to meet these challenges.

The software written in R [R Development Core Team, 2006] for combined candidate gene analysis is available from the authors upon request.

Acknowledgments

This work is supported in part by the grants from the National Institutes of Health (U01ES015089, R01CA085914, PO1CA-053996, R01AG14358, P30ES07033), and Young Investigator Award from the Children's Hospital and Regional Medical Center. The authors are very grateful to Drs. Charles Kooperberg, Michael LeBlanc, and Hua Tang for many helpful discussions and suggestions.

Appendix

Consistency and Asymptotic Normality of ($\hat{\alpha}^*$, $\hat{\beta}$) as the Solution to Equation (4)

Consider I cases, J unrelated controls, and parents of cases. Some parents may be missing genotype information, and this missingness is denoted by the indicator function Δ, which is 1 if both parents are available and 0 otherwise. We assume that such missingness is completely at random, that is, whether cases have parents' genotypes available is independent of genotypic configurations of the families and other covariates. The data consist of independent and identically distributed random variables $\{D_i, \Delta_i, X_i, i = 1, \ldots, I + J\}$ where D_i and X_i are, respectively, disease status and a vector of covariates for the i-th individual. The data also consist of parental genotypes $\{G_{pi} | D_i = 1, \Delta_i = 1, i = 1, \ldots, I\}$ for cases with $\Delta_i = 1$. For simplicity of notation, let $\theta = (\alpha^*, \beta)^T$ and $h(X; \theta) = \exp(\alpha^* + \beta X)/\{1 + \exp(\alpha^* + \beta X)\}$. Below we show that for the true parameter values $\theta_0 = (\alpha_0^*, \beta_0)$: (i) $\hat{\theta} \to \theta_0$, and (ii) $(I + J)^{-1/2} (\hat{\theta} - \theta_0) \to N(0, \Sigma)$, as $I + J \to \bullet\bullet$. Define

$$r^{(m)}\left(\beta, X, G_p\right) = \sum_{G^* \in \mathcal{G}_p} X^{*(m)} \exp(\beta X^*), \quad m = 0, 1, 2,$$

where $X^{*(0)} = 1$, $X^{*(1)} = X^*$, $X^{*(2)} = X^{*T} X^*$, $\mathcal{G}_p$ is the set of offspring genotypes that are possible given the parental genotypes G_p, and X^* is a vector of covariates equal to X except that the elements involving G are replaced by $G^* \in \mathcal{G}_p$. Under the logistic regression model (1), the estimating equation (4) can be expanded and represented as

$$S^{(1)}(\theta) = \sum_{i=1}^{I+J} \left\{D_i - h\left(X_i; \theta\right)\right\},$$

$$S^{(2)}(\theta) = \sum_{i=1}^{I+J} X_i \left\{D_i - h\left(X_i; \theta\right)\right\} + \sum_{i=1}^{I+J} D_i \Delta_i \left\{X_i - \frac{r^{(1)}\left(\beta, X, G_{ip}\right)}{r^{(0)}\left(\beta, X, G_{ip}\right)}\right\}.$$

We assume that the Hessian matrix of the log(pseudo-likelihood) is negative definite. To show that $\hat{\theta} \to \theta_0$, by the Foutz theorem [Foutz, 1977] we only need to prove that $(I + J)^{-1} S(\theta) \to 0$ at θ_0. Following the law of large numbers, we show that $(I + J)^{-1} S(\theta_0) \to s(\theta_0)$, where $s(\theta_0) = \{s^{(1)}(\theta_0), s_a^{(2)}(\theta_0) + s_b^{(2)}(\theta_0)\}$ with $s^{(1)}(\theta_0) = E\{D - h(X; \theta_0)\}$, $s_a^{(2)}(\theta_0) = EX\{D - h(X; \theta_0)\}$, and $s_b^{(2)}(\theta_0) = ED\Delta X - ED\Delta r^{(1)}(\beta, X, G_p)/r^{(0)}(\beta, X, G_p)$. Prentice and Pyke [1979] showed that both $s^{(1)}(\theta_0) = 0$ and $s_a^{(2)}(\theta_0) = 0$. The following demonstrates that $s_b^{(2)}(\theta_0) = 0$:

$$\begin{aligned} E(D\Delta X) &= (E\Delta) EE\left(DX | D = 1, \mathcal{G}_p, \Delta = 1\right) \\ &= (E\Delta) E\left\{ D \frac{\sum_{G^* \in \mathcal{G}_p} X^* P\left(D = 1 | G^*\right) P\left(G^* | G_p\right)}{\sum_{G^* \in \mathcal{G}_p} P\left(D = 1 | G^*\right) P\left(G^* | G_p\right)} | \Delta = 1 \right\} \\ &= ED\Delta \frac{r^{(1)}\left(\beta, X, G_p\right)}{r^{(0)}\left(\beta, X, G_p\right)}. \end{aligned}$$

Hence by the Foutz theorem [Foutz, 1977], there exists an unique sequence $\hat{\theta}$ such that $S(\hat{\theta}) = 0$ with probability going to 1 as $I + J \to \bullet\bullet$ and that $\hat{\theta} \to \theta_0$ in probability.

Next we show that $(I + J)^{1/2}(\hat{\theta} - \theta_0) \to N(0, \Sigma)$. The Taylor expansion of $S(\hat{\theta})$ at the true parameter values θ_0 gives

$$(I + J)^{-1/2} S\left(\hat{\theta}\right) =$$

$$(I + J)^{-1/2} S\left(\theta_0\right) + (I + J)^{-1} \frac{\partial}{\partial \theta} S(\theta)|_{\theta = \theta_0} (I + J)^{1/2} \left(\hat{\theta} - \theta_0\right) + o_p(1),$$

where $o_p(1)$ means asymptotically negligible. By the definition of $\hat{\theta}$ we have $S(\hat{\theta}) = 0$. Thus $(I + J)^{1/2}(\hat{\theta} - \theta_0)$ is asymptotically equivalent to

$$-\left\{(I + J)^{-1} \frac{\partial}{\partial \theta} S(\theta)|_{\theta = \theta_0}\right\}^{-1} (I + J)^{-1/2} S\left(\theta_0\right).$$

By the law of large numbers, we show that

$$-\frac{1}{I + J} \frac{\partial}{\partial \theta} S(\theta)|_{\theta = \theta_0} \to \sum\nolimits_1 \left(\theta_0\right),$$

with

$$\begin{aligned} \sum\nolimits_1 \left(\theta_0\right) &= E \begin{pmatrix} 1 & X^T \\ X & X^{\otimes 2} \end{pmatrix} h\left(X; \theta_0\right) \left\{1 - h\left(X; \theta_0\right)\right\} \\ &\quad + E \begin{pmatrix} 0 & 0 \\ 0 & ED\Delta \left\{ \frac{r^{(2)}\left(\beta_0, X, G_p\right)}{r^{(0)}\left(\beta_0, X, G_p\right)} \right\} - \left(\frac{r^{(1)}\left(\beta_0, X, G_p\right)}{r^{(0)}\left(\beta_0, X, G_p\right)} \right)^{\otimes 2} \end{pmatrix}, \end{aligned}$$

where $X^{\otimes 2} = X^T X$. Thus the variance Σ_1 consists of two components: the variance from the case-unrelated control sample and the variance from the case-parent sample. Applying the central limit theorem shows that

$$(I + J)^{-1/2} S(\theta_0) \to N(0, \Sigma_0),$$

where $\Sigma_0 = E\{S_i(\theta)S_i(\theta)^T\}\big|_{\theta=\theta_0}$, with $S_i(\theta)^T = \{D_i - h(X_i; \theta), [X_i(D_i - h(X_i; \theta)) + D_i\Delta_i(X_i - r^{(1)}(\beta, X, G_{ip})/r^{(0)}(\beta, X, G_{ip}))]^T\}$. If a case contributes to both the case-parent and case-unrelated control likelihoods, the variance Σ_0 takes this redundancy into account by adding the covariance to each individual component. All together, we have shown that

$$(I + J)^{1/2}(\hat{\theta} - \theta_0) \rightarrow N(0, \Sigma),$$

where $\Sigma = \Sigma_1^{-1}\,\Sigma_0\,\Sigma_1^{-1}$, which can be estimated consistently by replacing the expectations with empirical averages and θ_0 with $\hat{\theta}$.

References

Allen-Brady K, Wong J, Camp NJ: PedGenie: an analysis approach for genetic association testing in extended pedigrees and genealogies of arbitrary size. BMC Bioinformatics 2006;7:209.

Allen AS, Rahouz PJ, Satten GA: Informative missingness in genetic association studies: case-parent designs. Am J Hum Genet 2003; 72:671–680.

Allen AS, Satten GA: Inference on haplotype/disease association using parent-affected-child data: The projection conditional on parental haplotypes method. Genet Epidemiol 2007;31:211–223.

Allen AS, Satten GA, Tsiatis AA: Locally-efficient robust estimation of haplotype-disease association in family-based studies. Biometrika 2005;92:559–571.

Anderson JA: Separate sample logistic discrimination. Biometrika 1972;59:19–35.

Billingsley P: Convergence of Probability Measures, ed 2. New York, NY, J. Wiley & Sons, 1999.

Breslow NE, Day NE: Statistical Methods in Cancer Research (Vol. 1): The Analysis of Case-Control Studies. World Health Organization, 1980.

Chen YH: New approach to association testing in case-parent designs under informative parental missingness. Genet Epidemiol 2004;27:131–140.

Clayton DG: A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 1999; 62:950–961.

Clayton D, Jones H: Transmission disequilibrium tests for extended marker haplotypes. Am J Hum Genet 1999;65:1161–1169.

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 geneenvironment interactions, and parent-of-origin effects. Genet Epidemiol 2004;26:167–185.

Croiseau P, Genin E, Cordell HJ: Dealing with missing data in family-based association studies: A multiple imputation approach. Hum Hered 2007;63:229–238.

Curtin K, Wong J, Allen-Brady K, Camp NJ: PedGenie: meta genetic association testing in mixed family and case-control designs. BMC Bioinformatics 2007;8:448.

Dempster AP, Laird NM, Rubin DB: Maximum Likelihood from Incomplete Data via the EM algorithm. J R Stat Soc B 1977;39:1–38.

Devlin B, Roeder K: Genomic Control for Association Studies. Biometrics 1999;55:997–1004.

Dieckmann KP, Pichlmeier U: The prevalence of familial testicular cancer. Cancer 1997;80: 1954–1960.

Dudbridge F: UNPHASED User Guide. Technical Report 2006/5, Cambridge, UK, MRC Biostatistics Unit, 2006.

Epstein M, Veal C, Trembath R, Barker J, Li C, Satten G: Genetic association analysis using data from triads and unrelated subjects. Am J Hum Genet 2005;76:592–608.

Foutz RV: On the unique consistent solution to the likelihood equations. J Am Stat Ass 1977; 72:147–148.

Hankey BF, Ries LA, Edwards BK: The surveillance, epidemiology, and end results program: a national resource. Cancer Epidemiol Biomarkers Prev 1999;8:1117–1121.

Harlow BL, Davis S: Two one-step methods for household screening and interviewing using random digit dialing. Am J Epidemiol 1988; 127:857–863.

Hartge P, Brinton LA, Rosenthal JF, Cahill JI, Hoover RN, Waksberg J: Random digit dialing in selecting a population-based control group. Am J Epidemiol 1984;120:825–833.

Hopper JL, Bishop DT, Easton DF: Population-based family studies in genetic epidemiology. Lancet 2005;366:1397–406.

Kazeem GR, Farrall M: Integrating case-control and TDT studies. Ann Hum Genet 2005;69: 329–335.

Kistner EO, Weinberg CR: A method for identifying genes related to a quantitative trait, incorporating multiple siblings and missing parents. Genet Epidemiol 2005;29:155–165.

Kraft P, Thomas DC: Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. Am J Hum Genet 2000;66: 1119–1131.

Laird NM, Lange C: Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet 2006;7:385–394.

Langholz B, Borgan Ø: Counter-matching: A stratified nested case-control sampling method. Biometrika 1995;82:69–79.

Lin DY, Zeng D, Millikan R: Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. Genet Epidemiol 2005;29: 299–312.

Little RJA, Rubin D: Statistical Analysis with Missing Data, ed 2. Chichester, John Wiley, 2002.

Nagelkerke NJ, Hoebee B, Teunis P, Kimman TG: Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression. Eur J Hum Genet 2004;12:964–970.

Newcomb PA, Baron J, Cotterchio M, Gallinger S, Grove J, Haile R, Hall D, Hopper JL, Jass J, Le Marchand L, Limburg P, Lindor N, Potter JD, Templeton AS, Thibodeau S, Seminara D: Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. Cancer Epidemiol Biomarkers Prev 2007;16:2331–2343.

Prentice RL, Pyke R: Logistic disease incidence models and case-control studies. Biometrika 1979;66:403–411.

Pritchard JK, Rosenberg NA: Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 1999;65:220–228.

R Development Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3–900051–07–0, url: http://www.R-project.org.

Rabinowitz D, Laird N: A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered 2000;50:211–223.

Rabinowitz D: Adjusting for population heterogeneity and misspecified haplotype frequencies when testing nonparametric null hypotheses in statistical genetics. J Am Stat Assoc 2002;97:742–758.

Richiardi L, Pettersson A, Akre O: Genetic and environmental risk factors for testicular cancer. Int J Androl 2007 (OnlineEarly Articles) doi:10.1111/j.1365–2605.2007.00760.x

Rubin DB: Multiple imputation for nonresponse in surveys. New York, J Wiley & Sons, 1987.

Satten GA, Flanders WD, Yang Q: Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet 2001;68:466–477.

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

Self SG, Longton G, Kopecky KJ, et al: On estimating HLA/disease association with application to a study of aplastic anemia. Biometrics 1991;47:53–61.

Starr JR, Chen C, Doody DR, Hsu L, Ricks S, Weiss NS, Schwartz SM: Risk of testicular germ cell cancer in relation to variation in maternal and offspring cytochrome p450 genes involved in catechol estrogen metabolism. Cancer Epidemiol Biomarkers Prev 2005;14:2183–2190.

Swerdlow AJ: Testicular cancer: epidemiology and molecular endocrinology; in Henderson BE, Ponder B, Ross RK (eds): Hormones, Genes, and Cancer. New York, Oxford University Press, 2003.

Thomas DC: Statistical Methods in Genetic Epidemiology. New York, Oxford University Press, 2004.

Thomas DC, Witte JS: Point: population stratification: a problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prev 2002;11:505–512.

Weinberg CR: Allowing for missing parents in genetic studies of case-parent triads. Am J Hum Genet 1999;64:1186–1193.

Weinberg CR, Umbach DM: A hybrid design for studying genetic influences on risk of diseases with onset early in life. Am J Hum Genet 2005;77:627–636.

Whittemore AS: Logistic regression of family data from case-control studies. Biometrika 1995;82:57–64.

Whittemore AS: Estimating genetic association parameters from family data. Biometrika 2004;91:219–225.

Witte JS, Gauderman WJ, Thomas D: Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: Basic family designs. Am J Epidemiol 1999;149:693–705.

Zavos C, Andreadis C, Diamantopoulos N, Mouratidou D: A hypothesis on the role of insulin-like growth factor I in testicular germ cell tumours. Med Hypotheses 2004;63:511–514.

Zhang S, Zhu X, Zhao H: On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. Genet Epidemiol 2003;24:44–56.

Zhao LP, Li S, Khalid N: Assessing haplotype-based association with multiple SNPs in case-control studies. Am J Hum Genet 2003;72:1231–1250.