Genetic Epidemiology 33: 95–104 (2009)

Two-stage Analysis for Gene-Environment Interaction Utilizing Both Case-Only and Family-Based Analysis

Yi-Hau Chen,[1]* Hui-Wen Lin,[2] and Huimei Liu[2]

[1]*Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, Republic of China*
[2]*Department of Statistics, National Chengchi University, Taipei, Taiwan, Republic of China*

The case-only study and family-based study are two popular study designs for detecting gene-environment interactions. It is well known that the case-only analysis is efficient, but its validity relies crucially on the assumption of gene-environment independence in the study population. In contrast, the family-based analysis is robust to the violation of such an assumption, but is less efficient. We propose a two-stage study design for detecting gene-environment interactions, where a case-only study is performed at the first stage, and a case-parent/case-sibling study is performed at the second stage on a random subsample of the first-stage case sample as well as their parents/unaffected siblings. Statistical inference procedures are developed for the proposed two-stage study designs, which not only preserve the robustness property of the family-based analysis, but also utilize information from the case-only analysis to enhance estimation efficiency and testing power. Simulation results reveal both the robustness and efficiency of the proposed strategies. *Genet. Epidemiol.* 33:95–104, 2009.

Key words: case-parent studies; case-sibling studies; gene-environment independence; parental missingness

Contract grant sponsor: Genomic Research Center, Academia Sinica; Contract grant number: 94B001-2; Contract grant sponsor: National Science Council Republic of China; Contract grant number: NSC 95-2118-M-001-022-MY3.
*Correspondence to: Yi-Hau Chen, Ph.D. Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, Republic of China.
E-mail: yhchen@stat.sinica.edu.tw
Received 10 March 2008; Revised 9 May 2008; Accepted 10 June 2008.
Published online 17 July 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20357

INTRODUCTION

Detecting the potential interplay between susceptibility genotypes (G) and environmental exposures (E) has been one of major goals in many epidemiologic studies for complex human diseases. This is not only because the G-E interaction itself can delineate disease etiology and hence have important impacts on disease prevention and intervention [Hunter, 2005; Olden, 2007], but also because exploiting such interaction may result in greater statistical power for locating genetic susceptibility loci [Chatterjee et al., 2005; Kraft et al., 2007].

Statistical assessment of the G-E interaction can be performed in traditional epidemiologic study designs such as the case-control and the cohort designs. However, some power studies [Hwang et al., 1994; Foppa and Spiegelman, 1997] showed that the power to detect interactions may be quite limited with reasonably large sample sizes in such studies. Piegorsch et al. [1994] noticed that the G-E interaction can be detected in the nontraditional ''case-only'' design that uses diseased subjects (cases) only. In addition to its logistical convenience, the case-only analysis can achieve remarkable power improvements over traditional case-control/cohort analysis [Khoury and Flanders, 1996; Yang et al., 1997; Umbach and Weinberg, 1997].

The validity of the case-only analysis of G-E interactions hinges on the ''G-E independence'' assumption, which assumes that the genotypes and environmental exposures of interest are independent of each other in the study population [Piegorsch et al., 1994; Albert et al., 2001; Gatto et al., 2004]. Although the G-E independence assumption is believed to be reasonable in general since the genetic factors are determined at birth while the environmental factors are acquired through life, there still remain plausible settings where such an assumption fails. One classic situation for the violation of the G-E independence is ''population stratification,'' which arises when the study population consists of a few hidden strata such as ethnic groups, and both genotype frequencies and environmental exposure distributions vary among strata. Owing to such ''population stratification,'' even though there is perfect G-E independence in each stratum, there could exist overall correlation between genotypes and environmental exposures in the study population that eventually invalidates a case-only analysis. Another important reason for the presence of genotype-exposure dependencies is the confounding of family history, which is apparently associated with both susceptibility genes and lifestyle exposures. Thomas [2000] noted that the interactions between BRCA1/2 mutations and oral contraceptive (OC) use found in a case-only study [Ursin et al., 1997] for the breast cancer might be due to G-E dependence arising from population stratification, or from confounding by family history, which could be associated with women's tendency to use OCs, and also with the carrier status of BRCA1/2 as these

mutations account for the majority of hereditary breast cancers.

All the study designs mentioned above involve only unrelated subjects and hence are "population-based." The "family-based" designs using relatives of the cases as "matched controls," such as the case-parent and case-sibling designs, can also be applied to detect G-E interactions [Schaid, 1999; Umbach and Weinberg, 2000; Witte et al., 1999; Gauderman et al., 1999; Chatterjee et al., 2005]. It is well known that family-based studies can provide protection against population stratification bias [Laird and Lange, 2006; Thomas and Witte, 2002], and may be preferred for detecting G-E interaction involving rare genetic variants [Witte et al., 1999; Gauderman, 2002]. Further, family-based analysis usually requires weaker assumptions on distributions of genetic and environmental factors than case-only analysis. For example, the case-parent study requires only the G-E independence given parental genotypes, which is a type of within-family G-E independence that is usually unaffected by population stratification and family history [Thomas, 2000]. The case-sibling study is even more robust since it requires no distributional assumptions on genetic and environmental factors [Chatterjee et al., 2005].

This report is focused on detecting G-E interactions when there is concern on the validity of the G-E independence in the study population. To ensure valid statistical analysis of G-E interactions in the presence of population-level dependencies between genotypes and exposures, we consider a two-stage study design that aims to exploit the respective advantages from the case-only and the family-based designs. In the first stage, we collect data on genetic and environmental factors for a sample of cases and perform the case-only analysis on G-E interactions. Since the case-only analysis may be biased under population G-E dependencies, in the second stage, we further recruit family controls for a random subset of the first-stage case sample and perform the family-based analysis for G-E interactions. The second-stage family-based analysis, though robust to violation of the population G-E independence assumption, may be less powerful, especially when the second-stage study is conducted on a smaller scale. We therefore propose a refined two-stage analysis for detecting G-E interactions, which is still robust to violation of the population G-E independence assumption, but also utilizes information from the first-stage case-only analysis. The rationale for our proposal is derived from the general methodology in Chen and Chen [2000] for regression analysis under two-stage study designs.

In this report we consider specifically the use of the case-parent or the case-sibling study as the second-stage study. When performing the case-parent study, genotype data for both parents of randomly selected cases are collected at stage two (exposure data for parents are not required for the G-E interaction analysis in case-parent designs); when performing the case-sibling study, genotype and exposure data for unaffected siblings of randomly selected cases are collected at stage two. Since parental genotype data are quite vulnerable to being missing (especially for late-onset diseases), when a case-parent study is undertaken at stage two, we extend the approach of Chen [2004] to incorporate the incomplete parental genotype data and hence to increase the power for detecting G-E interactions.

METHODS

BACKGROUND AND MOTIVATION

Let D denote disease status with $D = 1$ denoting affected and $D = 0$ denoting unaffected. Assume that the disease penetrance model is given by

$$\log\left\{\frac{\Pr(D=1|G,E)}{\Pr(D=0|G,E)}\right\} = \alpha + \beta^{\mathrm{T}} m(G,E), \tag{1}$$

where G denotes genotypes and E denote environmental exposures, $m(G,E)$ is a known vector of coded values of (G,E) with $m(\cdot) = 0$ for the reference genotype-exposure group, and $\beta = (\beta_g, \beta_e, \beta_{ge})$ is a vector of association parameters, where β_g, β_e, and β_{ge} assess the genotype, exposure, and genotype-by-exposure interaction effects, and α is an intercept parameter that may be family specific to account for unmeasured family-specific risk factors. Specific penetrance models such as dominant, recessive, multiplicative, and co-dominant models can be established with suitably chosen $m(\cdot)$. For example, a dominant model for a locus is given by coding G as a 0-1 variable with "1" denoting homozygous/heterozygous presence and "0" denoting homozygous absence of the risk allele; a multiplicative model is given by coding G as the number of risk alleles contained in the genotype. When the disease under study is rare so that $\Pr(D=1|G,E) \approx \exp\{\alpha + \beta^{\top} m(G,E)\}$, the model (1) is equivalent to the model for relative disease risks:

$$\log\left\{\frac{\Pr(D=1|G,E)}{\Pr(D=1|G=0,E=0)}\right\} = \beta^{\mathrm{T}} m(G,E), \tag{2}$$

where $(G=0, E=0)$ refers to the reference genotype-exposure group.

If the G-E independence assumption holds in the study population so that

$$\Pr(G,E) = \Pr(G) \times \Pr(E), \tag{3}$$

then it has been shown that the interaction parameter β_{GE} can be estimated using data from diseased subjects (cases) only [Piegorsch et al., 1994; Umbach and Weinberg, 1997]. In principle, such "case-only" analysis can be performed by fitting a regression model for $\Pr(G|E, D=1)$ to data on (G,E) from the cases. For example, when the penetrance model is given by (1) with $m(G,E) = \beta_g G + \beta_e E + \beta_{ge} GE$ and G is 0-1 coded (e.g., in a dominant or recessive penetrance model), the "case-only" estimate for β_{ge} can be obtained as the estimate $\widehat{\gamma}_e$ by fitting the logistic regression model

$$\operatorname{logit}\Pr(G=1|E,D=1) = \gamma_0 + \gamma_e E \tag{4}$$

to (G,E) data from cases [Albert et al., 2001]; when G is coded with multiple categories (e.g., in a multiplicative or co-dominant model), a multinomial logistic regression model can be applied similarly [Armstrong, 2003; Cheng, 2006].

The main advantage for the case-only analysis is that it can achieve considerable higher efficiency than the traditional case-control analysis when the G-E independence assumption (3) holds, since the case-only analysis explicitly utilizes this assumption, while the case-control

analysis does not [Khoury and Flanders, 1996; Yang et al., 1997; Umbach and Weinberg, 1997]. Also, the case-only analysis is logistically more convenient since there is no need to find appropriate controls. However, the case-only analysis can produce non-negligible bias when the G-E independence assumption (3) does not hold [Albert et al., 2001].

Alternatively, the G-E interaction can be assessed in the family-based association studies. For example, in the case-parent trio study that genotypes cases and their parents and determines environmental exposures of the cases, the G-E interaction effect β_{ge} can be estimated through the "conditional on parental genotypes" (CPG) likelihood [Schaid, 1999; Thomas, 2000], where the likelihood contribution from a single case-parent trio is given by

$$\begin{aligned}&\Pr(G|G_p, E, D=1)\\ &= \frac{\Pr(D=1|G,E)\Pr(G|G_p)}{\sum_{G^*|G_p}\Pr(D=1|G^*,E)\Pr(G^*|G_p)}\\ &= \frac{\exp\{\beta^{\mathrm{T}} m(G,E)\}\Pr(G|G_p)}{\sum_{G^*|G_p}\exp\{\beta^{\mathrm{T}} m(G^*,E)\}\Pr(G^*|G_p)}.\end{aligned} \tag{5}$$

In (5), D, G, and E are disease status, genotype and environmental exposure for the case, G_p are genotypes of both parents of the case, and the summation is over all possible offspring genotypes G^* for the given parental genotypes G_p; the term $\Pr(G|G_p)$ is determined by the Mendelian proportions. Note that, in derivation of (5) we have employed the assumption

$$\Pr(G, E|G_p) = \Pr(G|G_p) \times \Pr(E|G_p), \tag{6}$$

namely, the genotype and environmental exposure are independent of each other given parental genotypes, a form of G-E independence within family. Note that the assumption (6) is much weaker than assumption (3) required in the case-only study: unlike (3), the assumption (6) is much less affected by population stratification and family history [see detailed discussions in Thomas, 2000; Chatterjee et al., 2005]. Note also that exposure data for parents are not required for G-E interaction analysis with case-parent designs.

The case-sibling study is another popular family-based design, especially for late-onset diseases since parents of the cases are difficult to obtain. The G-E interaction analysis in the case-sibling study can be based on the conditional likelihood [Siegmund et al., 2000], which is defined by conditioning on the number of affected siblings in each family. As an example, assuming that one affected and one unaffected siblings are chosen in each family, the conditional likelihood for this case-sibling pair is of the form

$$\begin{aligned}&\Pr(D=1, D_s=0|D+D_s=1, G, E, G_s, E_s)\\ &= \frac{\Pr(D=1, D_s=0|G,E,G_s,E_s)}{\Pr(D=1, D_s=0|G,E,G_s,E_s)+\Pr(D=0, D_s=1|G,E,G_s,E_s)}\\ &= \frac{\exp\{\beta^{\mathrm{T}} m(G,E)\}}{\exp\{\beta^{\mathrm{T}} m(G,E)\}+\exp\{\beta^{\mathrm{T}} m(G_s,E_s)\}},\end{aligned} \tag{7}$$

where D, G, and E are disease status, genotype and environmental exposure for the affected sibling, and D_s, G_s, and E_s are counterparts for the unaffected sibling. Note that the case-sibling analysis of the G-E interaction does not require any assumption on the joint distribution of genotype and exposure, hence is fully robust to the violation of the G-E independence assumption.

With the respective features of the case-only and family-based (case-parent/case-sibling) studies in mind, our proposal is to make a sensible compromise between the two types of studies, so that it can enjoy the robustness property of the family-based analysis and can utilize the efficiency of the case-only analysis. To this end, we propose a two-stage study design, which performs a larger scale case-only study at stage one to obtain convenient but possibly biased information on G-E interactions from the diseased subjects and then performs a smaller-scale case-parent/case-sibling study at stage two to ascertain information on true G-E interactions. Data from the first-stage case-only study and the second-stage case-parent/case-sibling study are then integrated to yield prudent and efficient analysis for G-E interactions.

PROPOSED TWO-STAGE STUDY DESIGN AND ANALYSIS FOR G-E INTERACTION

Here, we describe the proposed two-stage study design for detecting the G-E interaction.

In the first stage of the study, we collect genotype and exposure data $\{G_i, E_i\}_{i=1}^N$ for a sample of N cases. Case-only analysis of the G-E interaction is then performed by fitting a regression model for $\Pr(G|E, D=1)$ to the first-stage data. Our proposal applies with any type of the model chosen for $\Pr(G|E, D=1)$, and, as will be seen later, the validity of our proposal does not depend on correct specification for this model. Also, numerical experiments showed that the operation characteristics (testing power and estimation efficiency) of our proposal are not sensitive to the model choice for $\Pr(G|E, D=1)$ in the first-stage case-only analysis. Suppose that $\mathscr{P}(G, E; \gamma)$ is a model chosen for $\Pr(G|E, D=1)$ and γ is the vector of parameters in this model. Let $\bar{\gamma}$ be the resulting estimate for γ by fitting the model $\mathscr{P}(G, E; \gamma)$ to the first-stage data $\{G_i, E_i\}_{i=1}^N$.

In the second stage, we randomly select a subset $\mathscr{V}$ of the first-stage case-sample, and recruit family controls for the randomly selected cases to perform a family-based study for the G-E interaction. Specifically, the second-stage study can be a case-parent or a case-sibling study. When a case-parent study is performed, genotype data $\{G_{pi}\}_{i\in\mathscr{V}}$ for both parents of randomly selected cases are collected at stage two, and the estimate $\widehat{\beta}_{\mathrm{CPG}}$ for parameters in the CPG likelihood (5) is obtained with second-stage data $\{G_i, E_i, G_{pi}\}_{i\in\mathscr{V}}$; when a case-sibling study is undertaken, genotype and exposure data $\{G_{si}, E_{si}\}_{i\in\mathscr{V}}$ for unaffected siblings of the randomly selected cases are collected at stage two, and the estimate $\widehat{\beta}_{\mathrm{CL}}$ for parameters in the conditional likelihood (7) is obtained with second-stage data $\{G_i, E_i, G_{si}, E_{si}\}_{i\in\mathscr{V}}$. Let $\widehat{\beta}$ be either $\widehat{\beta}_{\mathrm{CPG}}$ or $\widehat{\beta}_{\mathrm{CL}}$ depending on the case-parent or the case-sibling study is performed at stage two. Besides, let $\widehat{\gamma}$ be the estimate of γ obtained by fitting $\mathscr{P}(G, E; \gamma)$ to the second-stage case sample $\{G_i, E_i\}_{i\in\mathscr{V}}$ (i.e., the genotype and exposure data for the cases randomly selected at stage two).

As mentioned above, for inference on the G-E interaction parameter β_{ge}, the case-only analysis based on $\bar{\gamma}$ is presumably very efficient, but may be vulnerably biased when the G-E independence assumption (3) fails. On the other hand, the family-based analysis based on $\widehat{\beta}$ is quite robust to the violation of (3) but is less efficient, especially

when the second-stage study is performed on a small scale. We thus propose a refined estimator that is still robust but can be more efficient than the family-based estimator $\widehat{\beta}$ by utilizing information from $\bar{\gamma}$.

The proposed refined estimator for the association parameter β is given as the simple formula

$$\bar{\beta} = \widehat{\beta} - \Delta(\widehat{\gamma} - \bar{\gamma}) \tag{8}$$

with

$$\Delta = C(\widehat{\beta}, \widehat{\gamma})V(\widehat{\gamma})^{-1},$$

where $C(\widehat{\beta}, \widehat{\gamma})$ is the covariance matrix between $\widehat{\beta}$ and $\widehat{\gamma}$, and $V(\widehat{\gamma})$ is the variance matrix of $\widehat{\gamma}$; these variance-covariance matrices can be simply estimated using the sandwich estimators given in the Appendix.

The estimator (8) is derived from the general approach in Chen and Chen [2000], which is originally developed for regression analysis under general two-stage study designs where the second-stage sample is a random subsample from the first-stage sample. Intuitively, we can see that the estimator $\bar{\beta}$ is always asymptotically unbiased no matter the case-only estimates $\bar{\gamma}$ and $\widehat{\gamma}$ are unbiased or not, since the second term in the right side of (8) is asymptotically zero as long as the second-stage sample is a random subsample of the first-stage sample, so that both $\bar{\gamma}$ and $\widehat{\gamma}$ have the same asymptotic limit. This implies that the assumptions required for the validity of $\bar{\beta}$ is no more than those required for the validity of the family-based estimator $\widehat{\beta}$. Specifically, when $\bar{\beta}$ is obtained by a case-parent analysis the resulting $\bar{\beta}$ is valid when the within-family G-E independence assumption (6) holds; when $\widehat{\beta}$ is given by a case-sibling analysis, $\bar{\beta}$ is valid without any assumptions on the joint distribution of genotype and exposure.

Moreover, as shown in Chen and Chen [2000], the variance matrix of $\bar{\beta}$ is given by

$$V(\bar{\beta}) = V(\widehat{\beta}) - (1 - \rho)\Delta V(\widehat{\gamma})\Delta^{\mathrm{T}}, \tag{9}$$

where $V(\widehat{\beta})$ and $V(\widehat{\gamma})$ are variance matrices of $\widehat{\beta}$ and $\widehat{\gamma}$, and $\rho = n/N$ is the subsampling fraction, i.e., the ratio of the number n of the cases randomly selected at stage two to the number N of the total (first stage) cases. It is then observed that the refined estimator $\bar{\beta}$ is asymptotically more efficient (with smaller asymptotic variance) than the family-based estimator $\widehat{\beta}$; the relative improvement in efficiency increases as the subsampling fraction ρ decreases and the correlation between $\widehat{\beta}$ and $\widehat{\gamma}$ increases.

Let $\bar{\beta}_{ge}$ and $V(\bar{\beta}_{ge})$ be the suitable components in $\bar{\beta}$ and $V(\bar{\beta})$ corresponding to the G-E interaction parameter β_{ge}. We propose $\bar{\beta}_{ge}$ as an estimator for β_{ge}, and propose the Wald test statistic $\bar{\beta}_{ge}^{\mathrm{T}} V(\bar{\beta}_{ge})^{-1}\bar{\beta}_{ge}$ for testing the null hypothesis $H_0 : \beta_{ge} = 0$ (no G-E interaction), which follows a χ_k^2 distribution under H_0, with k the dimension of β_{ge}. According to the above discussions, the proposed two-stage analysis for the G-E interaction is expected to be more efficient and powerful than the second-stage family-based analysis and to be more robust against the violation of the G-E independence assumption than the first-stage case-only analysis. Our simulation results do confirm these expectations.

ACCOUNTING FOR PARENTAL MISSINGNESS IN THE SECOND-STAGE CASE-PARENT STUDY

When the case-parent design is adopted in the second stage of the study, the issue of parental missingness needs to be considered since it is commonly encountered, particularly for late-onset diseases, and may reduce the power dramatically if the analysis is based on complete trios only [Allen et al., 2003]. Chen [2004] proposed a conditional likelihood approach to case-parent analysis of gene-disease association with incomplete parental genotypes, which keeps the full robustness property of the traditional (complete data) case-parent analysis by Spielman et al. [1993] and Schaid and Sommer [1993], while imposing no assumptions on the mating type distribution. Moreover,this approach can allow the parental missingness to depend on the missing parental genotypes themselves and does not require models for parental missingness to be explicitly specified. We extend Chen's approach to analysis of G-E interactions and incorporate this extension into the proposed two-stage analysis. The derivation of this extension is similar to that in Chen [2004], hence here we only sketch the outlines and defer some details to the Appendix.

As in Allen et al. [2003] and Chen [2004], we adopt an assumption on the parental missingness, which essentially states that, given the parental genotypes, the parental missingness is independent of offspring's genotype [see the Appendix, Equation (A.2), for the explicit mathematical expression]. Also, as in traditional case-parent analysis, we assume that the offspring's genotype and exposure are independently distributed conditional on the parental genotypes. The likelihood proposed is based on the conditional distribution of the offspring's genotype G given the number of observed parents R ($R = 0, 1, 2$), the genotypes of the observed parents G_{op}, if any, and the offspring's exposure E and disease status ($D = 1$). It is easy to see that, for "trio" ($R = 2$) families, the conditional probability $\Pr(G|R = 2, G_{op}, E, D = 1) = \Pr(G|G_p, E, D = 1)$, which has been given in (5). Among families with incomplete parental data, including "dyad" ($R = 1$) and "monad" ($R = 0$) families, only the dyad families with one heterozygous observed parent are found to be informative for the association parameter β if no assumptions are imposed on the mating type distribution. The explicit expression of the conditional probability $\Pr(G|R = 1, G_{op}, E, D = 1)$ for a dyad with one heterozygous observed parent is given in the Appendix, Equation (A.4), which involves, in addition to the parameter of interest β, also an exposure-specific nuisance parameter η_E that depends on the offspring's exposure level E.

The proposed conditional likelihood for case-parent analysis with incomplete parental genotypes is obtained as the product of all individual likelihoods (5) from complete trios and all individual likelihoods (A.4) from dyads with one heterozygous observed parent. Accordingly, for the two-stage analysis where a case-parent study is conducted at stage two, we propose to obtain the second-stage estimator $\widehat{\beta}$ (along with estimates for the nuisance parameters η_E's) by maximizing the above conditional likelihood. Note that, since each exposure level E would induce a different nuisance parameter η_E in (A.4), such

conditional likelihood is best applicable when the environmental exposure can be suitably categorized into a small number of groups.

SIMULATION STUDIES

We examine through simulations the performance of the proposed two-stage G-E interaction analysis. In all simulations the candidate gene locus is diallelic, and the environmental exposure is binary. Data were simulated from the population consisting of two subpopulations with constituent proportions (0.5,0.5). In the two subpopulations, the frequencies of the risk allele are 20% and 40%; the frequencies of exposure are 20 and 50%; and the disease prevalence rates are 1 and 5%. Genotypes for offsprings and siblings were generated assuming that Hardy-Weinberg equilibrium holds in each subpopulation. Simulations were conducted under various choices for the penetrance mode (dominant, recessive, or multiplicative), the magnitude of the interaction effect, and sample sizes N and n (number of cases in the first- and second-stage samples). Size and power evaluations under each setting were based, respectively, on 1,000 and 500 repeated runs.

When the dominant or recessive mode of penetrance was considered, our first-stage case-only analysis was performed by fitting a logistic model (4) to the first-stage data; when the multiplicative mode was considered, a multinomial logistic regression

$$\log\left\{\frac{\Pr(G=2|E,D=1)}{\Pr(G=0|E,D=1)}\right\} = \gamma_{02} + 2\gamma_e E,$$
$$\log\left\{\frac{\Pr(G=1|E,D=1)}{\Pr(G=0|E,D=1)}\right\} = \gamma_{01} + \gamma_e E$$

was applied. We found that the model specification in the first-stage case-only analysis is not crucial in our proposal: two other choices for $\Pr(G|E,D=1)$, including a convenient linear model and a saturated multinomial logistic regression model with genotype-specific parameters yielded testing power very close to that presented in Tables I and II (results not shown). This phenomenon may be due to that, although the model specified for

TABLE I. Simulation results (mean, variance, mean of estimated variances, and size/power of the associated Wald test for testing $\beta_{ge}=0$ at 5% significance level) for various estimators, including the first-stage case-only estimator $\tilde{\gamma}$, the second-stage case-parent estimator $\tilde{\beta}$ using only trios, the second-stage case-parent estimator $\hat{\beta}$ using trios and dyads, and the two-stage estimator $\bar{\beta}$

		Multiplicative model			Recessive model			Dominant model		
	β_{ge}	0	0.5	0.8	0	0.8	1	0	0.8	1
$n/N = 200/800$										
Mean	$\tilde{\gamma}$	0.203	0.686	1.020	0.191	1.001	1.200	0.293	1.025	1.212
	$\tilde{\beta}_{ge}$	−0.014	0.532	0.857	−0.029	0.879	1.104	0.021	0.860	1.047
	$\hat{\beta}_{ge}$	−0.009	0.532	0.855	−0.025	0.851	1.090	0.016	0.851	1.040
	$\bar{\beta}_{ge}$	−0.007	0.535	0.853	−0.005	0.829	1.036	0.018	0.834	1.010
Var.	$\tilde{\gamma}$	0.045	0.042	0.057	0.034	0.032	0.032	0.021	0.029	0.030
	$\tilde{\beta}_{ge}$	0.147	0.141	0.181	0.442	0.452	0.447	0.279	0.357	0.411
	$\hat{\beta}_{ge}$	0.147	0.148	0.187	0.357	0.376	0.369	0.241	0.300	0.341
	$\bar{\beta}_{ge}$	0.150	0.152	0.187	0.303	0.260	0.238	0.168	0.214	0.247
Est. Var.	$\tilde{\gamma}$	0.043	0.049	0.059	0.035	0.031	0.031	0.022	0.029	0.032
	$\tilde{\beta}_{ge}$	0.144	0.157	0.186	0.467	0.456	0.473	0.258	0.358	0.400
	$\hat{\beta}_{ge}$	0.139	0.156	0.184	0.377	0.363	0.363	0.234	0.310	0.342
	$\bar{\beta}_{ge}$	0.122	0.135	0.167	0.281	0.241	0.222	0.162	0.219	0.244
Size/power	$\tilde{\gamma}$	0.166	0.894	0.992	0.167	1	1	0.517	1	1
	$\tilde{\beta}_{ge}$	0.054	0.272	0.536	0.040	0.247	0.344	0.049	0.293	0.394
	$\hat{\beta}_{ge}$	0.059	0.288	0.546	0.052	0.324	0.433	0.043	0.327	0.444
	$\bar{\beta}_{ge}$	0.061	0.368	0.646	0.055	0.442	0.595	0.042	0.436	0.557
$n/N = 300/900$										
Mean	$\tilde{\gamma}$	0.210	0.728	1.022	0.194	1.011	1.226	0.288	1.074	1.263
	$\tilde{\beta}_{ge}$	0.002	0.521	0.814	−0.012	0.923	1.126	−0.008	0.823	1.031
	$\hat{\beta}_{ge}$	0.003	0.533	0.812	−0.005	0.923	1.130	−0.005	0.826	1.025
	$\bar{\beta}_{ge}$	0.004	0.533	0.812	0.001	0.892	1.084	0.022	0.830	1.041
Var.	$\tilde{\gamma}$	0.030	0.028	0.035	0.031	0.028	0.031	0.019	0.026	0.025
	$\tilde{\beta}_{ge}$	0.101	0.090	0.115	0.313	0.295	0.286	0.163	0.218	0.248
	$\hat{\beta}_{ge}$	0.095	0.103	0.133	0.264	0.264	0.243	0.138	0.192	0.213
	$\bar{\beta}_{ge}$	0.098	0.108	0.132	0.223	0.192	0.197	0.103	0.141	0.156
Est. Var.	$\tilde{\gamma}$	0.028	0.032	0.039	0.031	0.029	0.028	0.019	0.025	0.027
	$\tilde{\beta}_{ge}$	0.095	0.103	0.120	0.302	0.301	0.310	0.169	0.215	0.236
	$\hat{\beta}_{ge}$	0.093	0.101	0.119	0.250	0.243	0.226	0.147	0.198	0.213
	$\bar{\beta}_{ge}$	0.085	0.091	0.101	0.198	0.183	0.181	0.105	0.147	0.159
Size/power	$\tilde{\gamma}$	0.228	0.991	1	0.187	1	1	0.430	1	1
	$\tilde{\beta}_{ge}$	0.057	0.341	0.660	0.044	0.386	0.518	0.048	0.448	0.588
	$\hat{\beta}_{ge}$	0.058	0.432	0.674	0.055	0.476	0.640	0.046	0.460	0.624
	$\bar{\beta}_{ge}$	0.061	0.470	0.750	0.060	0.600	0.752	0.052	0.610	0.748

TABLE II. Simulation results (mean, variance, mean of estimated variances, and size/power of the associated Wald test for testing $\beta_{ge} = 0$ at 5% significance level) for various estimators, including the first-stage case-only estimator $\bar{\gamma}$, the second-stage case-sibling estimator $\hat{\beta}$, and the two-stage estimator $\bar{\beta}$

		Multiplicative model			Recessive model			Dominant model		
	β_{ge}	0	0.5	0.8	0	0.8	1	0	0.8	1
$n/N = 200/800$										
Mean	$\bar{\gamma}$	0.220	0.646	0.892	0.286	1.005	1.165	0.316	1.035	1.208
	$\hat{\beta}_{ge}$	0.002	0.506	0.822	0.024	0.869	1.031	−0.015	0.792	1.034
	$\bar{\beta}_{ge}$	0.001	0.493	0.809	0.023	0.852	1.023	−0.009	0.792	0.996
Var.	$\bar{\gamma}$	0.010	0.010	0.013	0.035	0.033	0.034	0.022	0.025	0.031
	$\hat{\beta}_{ge}$	0.095	0.105	0.119	0.382	0.394	0.379	0.214	0.244	0.274
	$\bar{\beta}_{ge}$	0.064	0.069	0.086	0.249	0.263	0.261	0.140	0.150	0.182
Est. Var.	$\bar{\gamma}$	0.010	0.011	0.013	0.034	0.032	0.031	0.021	0.028	0.030
	$\hat{\beta}_{ge}$	0.094	0.106	0.124	0.348	0.343	0.339	0.202	0.250	0.273
	$\bar{\beta}_{ge}$	0.063	0.073	0.087	0.237	0.243	0.240	0.138	0.166	0.182
Sise/power	$\bar{\gamma}$	0.603	1	1	0.332	1	1	0.546	1	1
	$\hat{\beta}_{ge}$	0.053	0.332	0.664	0.050	0.298	0.430	0.045	0.368	0.518
	$\bar{\beta}_{ge}$	0.056	0.466	0.800	0.044	0.398	0.582	0.049	0.512	0.650
$n/N = 300/900$										
mean	$\bar{\gamma}$	0.221	0.649	0.891	0.282	0.989	1.163	0.315	1.038	1.207
	$\hat{\beta}_{ge}$	0.009	0.516	0.815	−0.012	0.848	1.043	0.013	0.816	0.990
	$\bar{\beta}_{ge}$	0.007	0.509	0.803	−0.011	0.818	1.014	−0.001	0.799	0.992
Var.	$\bar{\gamma}$	0.009	0.010	0.012	0.032	0.030	0.030	0.020	0.026	0.026
	$\hat{\beta}_{ge}$	0.063	0.069	0.078	0.231	0.233	0.238	0.125	0.179	0.185
	$\bar{\beta}_{ge}$	0.042	0.050	0.053	0.167	0.169	0.171	0.094	0.109	0.122
Est. Var.	$\bar{\gamma}$	0.009	0.010	0.012	0.033	0.028	0.028	0.019	0.025	0.027
	$\hat{\beta}_{ge}$	0.062	0.070	0.081	0.228	0.219	0.222	0.132	0.164	0.178
	$\bar{\beta}_{ge}$	0.043	0.050	0.059	0.157	0.161	0.164	0.092	0.114	0.125
Size/power	$\bar{\gamma}$	0.653	1	1	0.364	1	1	0.620	1	1
	$\hat{\beta}_{ge}$	0.047	0.498	0.852	0.049	0.454	0.598	0.037	0.548	0.656
	$\bar{\beta}_{ge}$	0.048	0.650	0.932	0.045	0.526	0.726	0.047	0.684	0.816

$\Pr(G|E, D = 1)$ will affect the efficiency of the case-only estimator, it can be seen from expression (9) that the efficiency of the two-stage estimator is determined by the efficiency of the family-based estimator and the correlation between the family-based and case-only estimators. That is, the efficiency and power property of the two-stage analysis depend on the case-only analysis only through the correlation between the family-based and case-only estimators.

In the first simulation study we consider the two-stage study using the case-parent design at the second stage. Nuclear families with both parents and one offspring were generated from one of two subpopulations with probabilities 0.5 and 0.5, respectively. The disease status for the offspring in each family was generated according to the model

$$\Pr(D = 1|G, E) = K \exp\{\beta_g m(G) + \beta_e E + \beta_{ge} m(G)E\},$$

where $m(G)$ denotes the coding for the genotype according to the chosen penetrance mode. The main effects β_g and β_e were fixed at 0.3, and the subpopulation-dependent constant K was chosen to yield the desired overall disease rates (1% in the first subpopulation and 5% in the second subpopulation). As the first-stage ''case-only'' sample, N affected offsprings (cases) were randomly sampled from the generated nuclear families. The second-stage sample was composed of n cases randomly selected from the first-stage sample and their parents, where the parents were allowed to be missing, with the missing probability depending on the parental genotype and subpopulation [in the first (second) subpopulation, the probability of missing was 0.1 (0.2) for parents with zero or one risk allele and was 0.3 (0.4) for parents with two risk alleles]. On average, in the second-stage case-parent sample we have 65% complete trios, 30% dyads, and 5% monads. We applied the method presented in the ''Accounting For Parental Missingness In The Second-Stage Case-parent Study'' subsection to obtain the second-stage case-parent estimate $\hat{\beta}$. Besides, we also obtained the estimate $\tilde{\beta}$ based only on complete trios by applying traditional case-parent analysis [the CPG likelihood approach of Schaid and Sommer, 1993]. The simulation results with $n/N = 200/800(0.25)$ and $300/900(0.33)$ are given in Table I.

In the second simulation study the case-sibling design was performed at the second stage. Nuclear families consisting of two siblings were generated from one of two subpopulations with probabilities 0.5 and 0.5, respectively. The disease status for the siblings was simulated by the model

$$\text{logit}\Pr(D = 1|G, E) = \alpha + \beta_g m(G) + \beta_e E + \beta_{ge} m(G)E,$$

with $m(G)$ denoting the chosen coding for the genotype. The values for main effects (β_g, β_e) were set to 0.3, and the value of α was given by a normal random variable with unit variance and a mean chosen to yield the desired

overall disease rate (1% in the first subpopulation and 5% in the second subpopulation). The exposures of each sib-pair were generated by dichotomizing bivariate normal random variables (marginal means = 0, marginal variances = 1, correlation coefficient = 0.3), with the cut-point chosen to yield the desired frequency of exposure in each subpopulation. As the first-stage "case-only" sample, N affected siblings (cases) were randomly sampled from the generated sib-pairs with one affected and one unaffected sibling. The second-stage sample was obtained by randomly selecting n cases from the first-stage sample and then incorporating their unaffected siblings. The results with $n/N = 200/800(0.25)$ and $300/900(0.33)$ are given in Table II.

We have the following observations from the results shown in Tables I and II. First, in the considered setting where population stratification exits, the proposed estimator $\bar{\beta}$, like the family-based estimator $\hat{\beta}$, is essentially unbiased and so is its variance estimator. The Wald test based on $\bar{\beta}$ for testing $\beta_{ge} = 0$ has correct type-I error rates. On the contrary, the case-only estimator $\bar{\gamma}$ is biased for β_{ge}, and the associated Wald test for testing $\beta_{ge} = 0$ has substantially inflated type-I error rates. These results highlight the potential drawback for the case-only design and the importance of more robust study designs such as the family-based and the proposed two-stage designs. Second, the proposed estimator $\bar{\beta}$ is usually more efficient (with smaller variability) than the family-based (case-parent and case-sibling) estimators $\hat{\beta}$ using only the second-stage data, and the Wald test based on $\bar{\beta}$ for testing $\beta_{ge} = 0$ is more powerful than that based on $\hat{\beta}$. The relative improvements in power for our proposal over the second stage family-based test can be as high as 40% when the subsampling fraction = 0.25 and 30% when the subsampling fraction = 0.33. The power improvements are more significant for the recessive and dominant penetrance models, and are less significant for the multiplicative model. Also, comparing results between Tables I and II we see that the power improvements are similar regardless of which study design (case-parent/case-sibling) is undertaken at stage two. Third, we see from Table I that, except when the penetrance mode is multiplicative, the estimator $\tilde{\beta}$ based only on complete trios is less efficient than the estimator $\hat{\beta}$ we proposed for further exploiting incomplete trios and the two-stage estimator $\bar{\beta}$ we proposed for exploiting both the incomplete trios and the first-stage case-only sample.

To gain more insights into the ability of utilizing the "case-only" information for the proposed two-stage analysis, we further conducted simulations under the settings described above with dominant penetrance and $\beta_{ge} = 1$ (assuming all parents were available when the case-parent study was performed at stage two). We fixed the number of families sampled in the second stage at $n = 200$ but varied the number of cases in the first stage sample: $N = 400, 500, 600, 800, 1,000$, or $2,000$ (corresponding to $\rho = 0.5, 0.4, 0.33, 0.25, 0.2$, or 0.1). We can see from the power curves in Figure 1 that, as the size of the first-stage case sample increases, the proposed two-stage analysis adaptively incorporates the information wherein and increases its power to detect the G-E interaction. The results highlight the fact that the proposed two-stage analysis, not only can be as robust as the family-based analysis, but also can borrow efficiency from the case-only analysis.

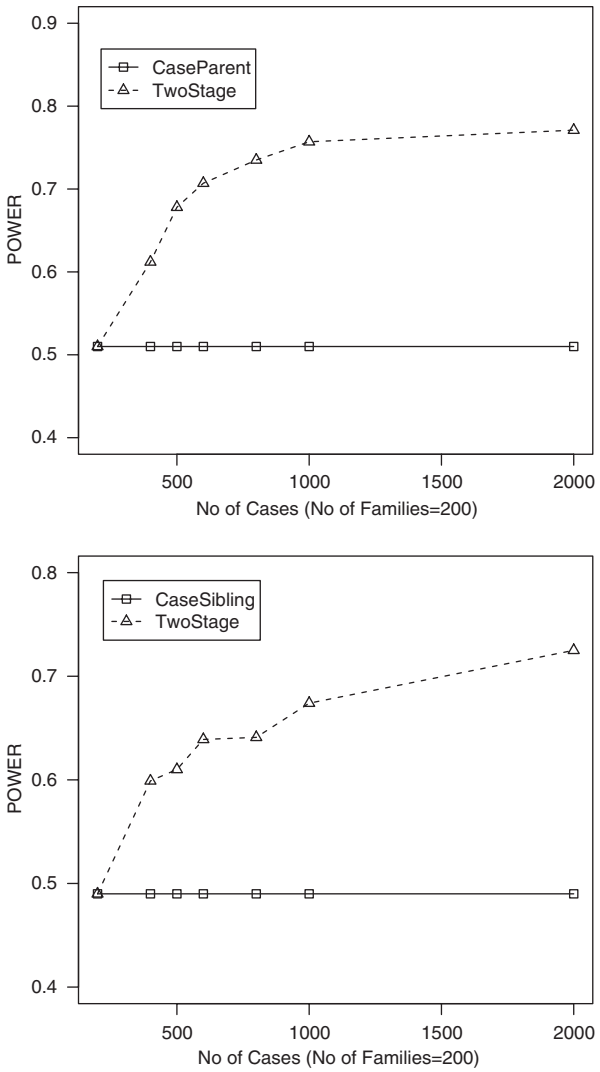

Fig. 1. Power (for detecting gene-environment interaction at 5% significance level) of the proposed two-stage analysis using case-parent (upper panel) and case-sibling design (lower panel) at stage two. The number of trios (upper panel) or sibships (lower panel) is 200. The number of cases at stage one increases from 400 to 2,000. Data were simulated under the dominant model with the value of the interaction parameter fixed at 1.

DISCUSSION

We have proposed a two-stage study design for detecting gene-environment interactions, which consists of a larger-scale first-stage case-only study and a smaller-scale second-stage case-parent/case-sibling study. The motivation is based on the consideration that the case-only design is usually convenient and efficient but vulnerable to bias due to gene-environment dependencies in the study population, while the case-parent/case-sibling deign is more robust to such bias but less efficient. The proposed two-stage analysis for G-E interactions is intended to utilize the strengths from the two different type of studies: it always yields asymptotically unbiased parameter estimate and test size, while also gaining

efficiency from the case-only analysis, even in presence of population-level gene-environment dependencies. The proposed estimator $\bar{\beta}$ and its variance estimator are very easy to implement, involving only quantities that are available from existing approaches such as the conditional and unconditional logistic regressions.

Since a case-only study may be biased under G-E dependencies, a prudent strategy would be to perform a case-only design as a main study, subsequently followed by an independent family-based study as a confirmatory study. When results from the two independent studies are consistent, the case-only analysis may be reported as the final analysis. When results from the two studies are inconsistent, suggesting that the G-E independence may fail so that the case-only analysis may be subject to bias, one may base the final analysis only on the family-based study, which is less likely to be biased but may be of limited power due to its limited sample size. Alternatively, by regarding the cases (affected offspring/sibling) in the family-based sample as a subsample of the pooled case sample combining all the cases in case-only and family-based samples, the proposed joint analysis can then be applied to yield the refined analysis utilizing both the case-only and family-based information on the G-E interaction, provided that the case-only and the family-based samples are essentially from the same underlying population.

An assumption required for the proposed approach is that the stage-two sample is a random subsample of the stage-one (case-only) sample. This assumption may be violated when the participation of the parents/siblings for a second-stage study depends on a variety of selection factors. The mechanism of selecting the parents/siblings at stage two can be equivalently regarded as a missing data mechanism (the family member not participating the second-stage study is ''missing''). Recall that the case-parent trio analysis (and also the case-parent trio and dyad analysis we proposed above) is in fact valid under the missing data mechanism expressed in (A.2), which is a type of ''informative missing'' mechanism [Allen et al., 2003; Chen, 2004]. Hence, if the case-parent design is used at stage two, the proposed analysis is valid even under a non-random selection mechanism, so long as such a mechanism satisfies (A.2). On the other hand, when the case-sibling design is adopted at stage two, our proposal does rely on the randomness assumption. When this assumption fails, techniques for allowing non-random missing-data mechanism in conditional logistic regression analysis may be applied, e.g., Sinha and Maiti [2008]; this is an issue deserving further investigation.

To examine the potential bias caused by the non-random selection of the stage-two sample, we conducted a small simulation study with the same setting as in Table II ($n/N = 200/800$) except that here the selection of stage-two unaffected siblings depends on the subpopulation membership: the siblings in one subpopulation have a lower likelihood of participating the second-stage study than the siblings in the other subpopulation. When the likelihood of participating the second-stage study in the two subpopulations is 20 and 30%, the type-I error rates for testing $\beta_{ge} = 0$, evaluated by 1,000 simulations, are 5.4, 5.0, and 5.6% under the multiplicative, recessive, and dominant models, respectively. It is seen that under the mild deviation from random selection, the bias of the proposed two-stage analysis is ignorable. In the setting with more severe deviation from random selection, i.e., the likelihood of participating the second-stage study in the two subpopulations is 15 and 35%, the type-I error rates under the multiplicative, recessive, and dominant models are, respectively, 8.8, 6.0, and 6.2%, which are still acceptable except for the result under the multiplicative model.

When the population G-E independence assumption does hold, our proposal may be less powerful than the case-only analysis since the proposed two-stage analysis, to achieve robustness in general, does not exploit this assumption while the the case-only analysis does and achieves full efficiency [Umbach and Weinberg, 1997]. To assess the potential loss of efficiency for the proposed two-stage analysis as compared to the case-only analysis when the G-E independence holds, we performed a simulation study with the same setting as in Table I (using the case-parent design at stage two and assuming all parents are available), but with all the data collected from the first subpopulation, i.e., the study population is homogeneous and has no substructure, and hence the G-E independence holds. Table III demonstrates the results on testing power based on 500 replications. The proposed two-stage analysis, though more powerful than the second-stage family-based (case-parent) analysis, is less powerful than the first-stage case-only analysis. We note that the loss of power for the two-stage analysis relative to the case-only analysis can be substantial or modest, depending on the second-stage subsampling fraction (n/N) is smaller or larger. Given these results and the fact that the case-only

TABLE III. Power of the Wald tests for testing $\beta_{ge} = 0$ at 5% significance level based on the first-stage case-only estimator $\tilde{\gamma}$, the second-stage case-parent estimator $\hat{\beta}$, and the two-stage estimator $\bar{\beta}$ under the setting where the G-E independence holds

	Multiplicative model		Recessive model		Dominant model	
β_{ge}	0.5	0.8	0.8	1	0.8	1
$n/N = 200/800$						
$\hat{\beta}_{ge}$	0.318	0.531	0.361	0.545	0.370	0.468
$\bar{\beta}_{ge}$	0.452	0.697	0.563	0.749	0.587	0.690
$\tilde{\gamma}$	0.990	1	0.997	1	0.984	1
$n/N = 400/800$						
$\hat{\beta}_{ge}$	0.490	0.730	0.602	0.808	0.572	0.734
$\bar{\beta}_{ge}$	0.606	0.824	0.762	0.909	0.732	0.850
$\tilde{\gamma}$	0.984	1	0.992	1	0.982	0.992

analysis may result in substantial bias even under mild G-E dependencies [Albert et al., 2001], the proposed two-stage analysis for G-E interactions is thus best suited for the setting when there is some concern that the G-E independence may not hold in the study population, e.g., when there has been substantial evidence that the population stratification exists.

ACKNOWLEDGMENTS

The authors thank the two reviewers for their many helpful comments.

REFERENCES

Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. 2001. Limitations of the case-only design for identifying gene-environment interactions. Am J Epidemiol 154:687–693.

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

Armstrong BG. 2003. Fixed factors that modify the effects of time-varying factors: applying the case-only approach. Epidemiol 14:467–472.

Chatterjee N, Kalaylioglu Z, Carroll RJ. 2005. Exploiting gene-environment independence in family-based case-control studies: increased power for detecting associations, interactions and joint effects. Genet Epidemiol 28:138–156.

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

Chen YH, Chen H. 2000. A unified approach to regression analysis under double-sampling designs. J R Stat Soc B 62:449–460.

Cheng KF. 2006. A maximum likelihood method for studying gene-environment interactions under conditional independence of genotype and exposure. Stat Med 25:3093–3109.

Foppa I, Spiegelman D. 1997. Power and sample size calculations for case-control studies of gene-environment interactions with a polytomous exposure variable. Am J Epidemiol 146:596–604.

Gatto NM, Campbell UB, Rundle AG, Ahsan H. 2004. Further development of the case-only design for assessing gene-environment interaction: evaluation of and adjustment for bias. Int J Epidemiol 33:1014–1024.

Gauderman WJ. 2002. Sample size requirements for matched case-control studies of gene-environment interaction. Stat Med 21:35–50.

Gauderman WJ, Witte JS, Thomas DC. 1999. Family-based association studies. J N Cancer Inst 26:31–37.

Hunter DJ. 2005. Gene-environment interactions in human diseases. Nat Rev Genet 6:287–298.

Hwang SJ, Beaty TH, Liang KY, Coresh J, Khoury MJ. 1994. Minimum sample size estimation to detect gene-environment interaction in case-control designs. Am J Epidemiol 140:1029–1037.

Khoury MJ, Flanders WD. 1996. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls. Am J Epidemiol 144:207–213.

Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. 2007. Exploiting gene-environment interaction to detect genetic associations. Hum Hered 63:111–119.

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

Olden K. 2007. Commentary: From phenotype, to genotype, to gene-environment interaction and risk for complex diseases. Int J Epidemiol 36:18–20.

Piegorsch WW, Weinberg CR, Taylor JA. 1994. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med 13:153–162.

Schaid DJ. 1999. Case-parents design for gene-environment interaction. Genet Epidemiol 16:261–273.

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

Siegmund KD, Langholtz B, Kraft P, Thomas DC. 2000. Testing linkage disequilibrium in sibships. Am J Hum Genet 67:244–248.

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

Sinha S, Maiti T. 2008. Analysis of matched case-control data in presence of nonignorable missing exposure. Biometrics 64:106–114.

Thomas DC. 2000. Case-parents design for gene-environment interaction by Schaid. Genet Epidemiol 19:461–463.

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

Umbach DM, Weinberg CR. 1997. Designing and analysing case-control studies to exploit independence of genotype and exposure. Stat Med 16:1731–1743.

Umbach DM, Weinberg CR. 2000. The use of case-parent triads to study joint effects of genotype and exposure. Am J Hum Genet 66:251–261.

Ursin G, Henderson BE, Haile RW, Zhou N, Diep A, Bernstein L. 1997. Is oral contraceptive use more common in women with BRCA1/BRCA2 mutations than in other women with breast cancer? Cancer Res 57:3678–3681.

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

Yang Q, Khoury MJ, Flanders WD. 1997. Sample size requirements in case-only designs to detect gene-environment interaction. Am J Epidemiol 146:713–720.

APPENDIX

SANDWICH ESTIMATORS FOR THE VARIANCE-COVARIANCE MATRICES USED IN (8)

We first consider the setting where the case-parent study is performed at stage two and all parents are available. For the second-stage case i ($i \in \mathscr{V}$), let $S_i(\beta) = \partial \log \Pr(G_i|G_{pi}, E_i, D_i = 1)/\partial\beta$ and $I(\beta) = -\sum_{i\in\mathscr{V}} \partial S_i(\beta)/\partial\beta$ be the score function and observed information matrix for the CPG likelihood (5). Also, let $\mathscr{P}(G, E; \gamma)$ be a model specified for $\Pr(G|E, D = 1)$, $U_i(\gamma) = \partial \log \mathscr{P}(G_i, E_i; \gamma)/\partial\gamma$ and $J(\gamma) = -\sum_{i\in\mathscr{V}} \partial U_i(\gamma)/\partial\gamma$ be the score function and information matrix for the case-only analysis based on the second-stage case sample. Then the sandwich estimators for $V(\widehat{\gamma})$ and $C(\widehat{\beta}, \widehat{\gamma})$ are given as

$$V(\widehat{\gamma}) = J(\widehat{\gamma})^{-1}\left\{\sum_{i\in\mathscr{V}} U_i(\widehat{\gamma})U_i^{\mathrm{T}}(\widehat{\gamma})\right\}J(\widehat{\gamma})^{-1},$$

$$C(\widehat{\beta}, \widehat{\gamma}) = I(\widehat{\beta})^{-1}\left\{\sum_{i\in\mathscr{V}} S_i(\widehat{\beta})U_i^{\mathrm{T}}(\widehat{\gamma})\right\}J(\widehat{\gamma})^{-1}. \quad \text{(A.1)}$$

Note that we use a sandwich-type robust estimate, rather than the inverse of information matrix, for estimating the variance of $\widehat{\gamma}$ since we allow the model used in case-only

analysis to be misspecified. The variance matrix of $\widehat{\beta}$, $V(\widehat{\beta})$, can be obtained as $I(\widehat{\beta})^{-1}$ or the sandwich-type estimate. If the case-sibling study is performed at stage two, $S_i(\beta)$ and $I(\beta)$ are, respectively, replaced by the score function and information matrix from the conditional likelihood (7).

Now consider the setting where the case-parent study is performed at stage two and there are missing parents. Recall that the proposed conditional likelihood, consisting of both trios and dyads with one heterozygous observed parent, involves β and a set of exposure-specific nuisance parameters. Denote the set of nuisance parameters as η and the proposed conditional likelihood by $L(\beta, \eta)$. If both parents of the second-stage case i are available (i.e. the family is a trio), then $S_i(\widehat{\beta})$ is still given by the score function of the CPG likelihood (5). If only one of the parents of the second-stage case i is available and is heterozygous, the conditional likelihood for this family, whose form is given in Equation (A.4), involves β and an exposure-specific nuisance parameter, say η_e. For this family, $S_i(\widehat{\beta})$ in (A.1) is replaced by

$$S_{\beta,i}(\widehat{\beta}, \widehat{\eta}) - I_{\beta\eta_e} I_{\eta_e\eta_e}^{-1} S_{\eta_e,i}(\widehat{\beta}, \widehat{\eta}),$$

where $S_{\beta,i}(\beta, \eta)$ and $S_{\eta_e,i}(\beta, \eta)$ are the score functions of family i's likelihood with respect to β and η_e, respectively, and $I_{\beta\eta_e} = -\partial^2 \log L(\beta, \eta)/\partial\beta\partial\eta_e$ and $I_{\eta_e\eta_e} = -\partial^2 \log L(\beta, \eta)/\partial\eta_e^2$ are submatrices from the full information matrix. Also, in this setting, the matrix $I(\beta)$ in (A.1) is replaced by $I_{\beta\beta} = -\partial^2 \log L(\beta, \eta)/\partial\beta^2$, and the variance matrix $V(\widehat{\beta})$ is replaced by $(I_{\beta\beta} - I_{\beta\eta} I_{\eta\eta}^{-1} I_{\beta\eta}^{\mathrm{T}})^{-1}$. All these matrices are evaluated at $(\widehat{\beta}, \widehat{\eta})$, the parameter estimates obtained by maximizing $L(\beta, \eta)$.

CONDITIONAL LIKELIHOODS FOR FAMILIES WITH MISSING PARENTS

Without loss of generality, we assume that the gene locus under study has two alleles, and the associated genotypes are labeled by $G = 0$, 1, and 2. Let Z index the subpopulations. Let $\mathscr{R} = (\mathscr{R}_f, \mathscr{R}_m)$ denote the parental missing pattern, where $\mathscr{R}_f$ ($\mathscr{R}_m$) equals one if the father (mother) is observed, and equals zero if the father (mother) is missing. Similar to Allen et al. [2003] and Chen [2004], we make the following probabilistic assumption on the parental missingness:

$$\begin{aligned} &\Pr(\mathscr{R}|G_p, G, E, D = 1, Z) \\ &\quad = \Pr(\mathscr{R}|G_p, E, D = 1, Z), \end{aligned} \tag{A.2}$$

that is, given the subpopulation Z, the offspring's environmental exposure E and disease status $D = 1$, and the parental genotypes G_p, the parental missingness R is independent of offspring's genotype G.

The likelihood proposed for each family is based on the conditional distribution of the offspring's genotype G given the number of observed parents R ($R = 0, 1, 2$), the genotypes of the observed parents G_{op}, if any, and the offspring's exposure E and disease status ($D = 1$). In particular, we consider this conditional probability for families with $R = 1$ (''dyad'') and $R = 2$ (''trio''), since, based on arguments in Chen [2004], families with $R = 0$ (''monad'') contain no information on the association parameter β if we do not impose any assumptions on the mating-type distribution. By assumption (A.2) we immediately have $\Pr(G|R = 2, G_{op}, E, D = 1) = \Pr(G|G_p, E, D = 1)$, which has been given in (5). Let $G_p = (G_f, G_m)$ be the complete parental genotypes including the paternal (G_f) and maternal (G_m) genotypes. For dyad families with one heterozygous observed parent, using the penetrance model $\Pr(D = 1|G, E, Z) = K_Z \exp\{\beta^{\mathrm{T}} m(G, E)\}$ where $K_Z = \Pr(D = 1|G = 0, E = 0, Z)$, and the assumption (A.2) we have

$$\begin{aligned} &\Pr(G, G_{op} = 1, R = 1, E, D = 1) \\ &\quad = \exp\{\beta^{\mathrm{T}} m(G, E)\} K_{GE}, \quad G = 0, 1, 2, \end{aligned}$$

where

$$\begin{aligned} K_{GE} = &\sum_{G_m} q_E(G_f = 1, G_m) \Pr(G|G_f = 1, G_m, E, Z) \\ &+ \sum_{G_f} q'_E(G_f, G_m = 1) \Pr(G|G_f, G_m = 1, E, Z), \end{aligned} \tag{A.3}$$

$$\begin{aligned} q_E(G_f, G_m) = &\sum_Z \Pr[\mathscr{R} = (1, 0)|G_f, G_m, D = 1, E, Z] \\ &\times K_Z \Pr(G_f, G_m|E, Z) \Pr(E|Z) \Pr(Z), \end{aligned}$$

and $q'_E(G_f, G_m)$ is defined as $q_E(G_f, G_m)$ with the first term replaced by the probability for $R = (0, 1)$. Therefore,

$$\begin{aligned} &\Pr(G|G_{op} = 1, R = 1, E, D = 1) \\ &\quad = \frac{\exp\{\beta^{\mathrm{T}} m(G, E)\} K_{GE}}{\sum_{g=0,1,2} \exp\{\beta^{\mathrm{T}} m(g, E)\} K_{gE}}, \quad G = 0, 1, 2. \end{aligned}$$

In (A.3), if we further assume that G and E are independent given parental genotypes, and the Mendelian proportions hold in each subpopulation, we have, similar to Chen [2004], the identity $K_{0E} + K_{2E} = K_{1E}$ for each exposure level E. This identify then leads to

$$\begin{aligned} &\Pr(G|G_{op} = 1, R = 1, E, D = 1) \\ &\quad = \frac{\exp\{\beta^{\mathrm{T}} m(G, E)\} \theta(G, E)}{\sum_{g=0,1,2} \exp\{\beta^{\mathrm{T}} m(g, E)\} \theta(g, E)}, \quad G = 0, 1, 2, \end{aligned} \tag{A.4}$$

where

$$\theta(G, E) = \begin{cases} \exp(\eta_E) & \text{for } G = 0 \\ 1 & \text{for } G = 1 \\ 1 - \exp(\eta_E) & \text{for } G = 2 \end{cases}$$

and $\eta_E (= \log\{K_{0E}/K_{1E}\})$ is an exposure-specific nuisance parameter depending on the off-springs exposure level E. The conditional probabilities for dyad families with one homozygous observed parent ($R = 1$ and $G_{op} = 0$ or 2) can be derived in a similar way and they would involve two nuisance parameters for a given value of E. Recall that the genotype G is trinomial and hence the degrees of freedom is 2. Thus, after accounting for the nuisance parameters, the dyad families with one homozygous observed parent cannot provide additional information on β and hence can be excluded from analysis. The proposed conditional likelihood for case-parent analysis with incomplete parental genotypes is then obtained as the product of all individual likelihoods (5) from trios and all individual likelihoods (A.4) from dyads with one heterozygous observed parent. This conditional likelihood generalizes that proposed in Chen [2004] to incorporate inference on the G-E interaction.