Likelihood-Based Inference on Haplotype Effects in Genetic Association Studies

D. Y. LIN and D. ZENG

A haplotype is a specific sequence of nucleotides on a single chromosome. The population associations between haplotypes and disease phenotypes provide critical information about the genetic basis of complex human diseases. Standard genotyping techniques cannot distinguish the two homologous chromosomes of an individual, so only the unphased genotype (i.e., the combination of the two homologous haplotypes) is directly observable. Statistical inference about haplotype–phenotype associations based on unphased genotype data presents an intriguing missing-data problem, especially when the sampling depends on the disease status. The objective of this article is to provide a systematic and rigorous treatment of this problem. All commonly used study designs, including cross-sectional, case-control, and cohort studies, are considered. The phenotype can be a disease indicator, a quantitative trait, or a potentially censored time-to-disease variable. The effects of haplotypes on the phenotype are formulated through flexible regression models, which can accommodate various genetic mechanisms and gene–environment interactions. Appropriate likelihoods are constructed that may involve high-dimensional parameters. The identifiability of the parameters and the consistency, asymptotic normality, and efficiency of the maximum likelihood estimators are established. Efficient and reliable numerical algorithms are developed. Simulation studies show that the likelihood-based procedures perform well in practical settings. An application to the Finland–United States Investigation of NIDDM Genetics Study is provided. Areas in need of further development are discussed.

KEY WORDS: Case-control study; Gene–environment interaction; Hardy–Weinberg equilibrium; Missing data; Single nucleotide polymorphism; Unphased genotype.

1. INTRODUCTION

In the early 1900s there was a fierce debate between Gregor Mendel's followers and the biometrical school led by Francis Galton and Karl Pearson as to whether the patterns of inheritance were consistent with Mendel's law of segregation or with a "blending"-type theory. Fisher (1918) reconciled the two conflicting schools by recognizing the difference in the genetic basis for the variation in the trait being studied. For the traits that the Mendelists studied, the observed variation was due to a simple difference at a single gene; for the traits studied by the biometrical school, individual differences were attributed to many different genes, with no particular gene having a singly large effect.

Like the traits studied by Mendel, many genetic disorders, such as Huntington disease and cystic fibrosis, are caused by mutations of single genes. The genes underlying a number of these Mendelian syndromes have been discovered over the last 20 years through linkage analysis and positional cloning (Risch 2000). The same approach, however, is failing to unravel the genetic basis of complex human diseases (e.g., hypertension, bipolar disorder, diabetes, schizophrenia), which are influenced by a variety of genetic and environmental factors, just like the traits studied by the biometrical school a century ago. It is widely recognized that genetic dissection of complex human disorders requires large-scale association studies, which relate disease phenotypes to genetic variants, especially single nucleotide polymorphisms (SNPs) (Risch 2000; Botstein and Risch 2003).

SNPs are DNA sequence variations that occur when a single nucleotide in the genome sequence is altered. SNPs make up about 90% of all human genetic variation and are believed to have a major impact on disease susceptibility. Aided by the sequencing of the human genome (International Human Genome Sequencing Consortium 2001; Venter et al. 2001), geneticists have identified several million SNPs (International SNP Map Working Group 2001). With current technology, it is economically feasible to genotype thousands of subjects for thousands of SNPs. These remarkable scientific and technological advances offer unprecedented opportunities to conduct SNPs-based association studies aimed at unraveling the genetic basis of complex diseases.

There is one of three possible genotypes at each SNP site: homozygous with allele A, homozygous with allele a, or heterozygous with one allele A and one allele a. Thus assessing the association between a SNP and a disease phenotype is a trivial three-sample problem. It is, however, desirable to deal with multiple SNPs simultaneously. One appealing approach is to consider the haplotypes for multiple SNPs within candidate genes (Hallman, Groenemeijer, Jukema, and Boerwinkle 1999; International SNP Map Working Group 2001; Patil et al. 2001; Stephens, Smith, and Donnelly 2001).

The haplotype (i.e., a specific combination of nucleotides at a series of closely linked SNPs on the same chromosome of an individual) contains information about the protein products. Because the actual number of haplotypes within a candidate gene is much smaller than the number of all possible haplotypes, haplotyping serves as an effective data-reduction strategy. Using SNP-based haplotypes may yield more powerful tests of genetic associations than using individual, unorganized SNPs, especially when the causal variants are not measured directly or when there are strong interactions of multiple mutations on the same chromosome (Akey, Jin, and Xiong 2001; Fallin et al. 2001; Li 2001; Morris and Kaplan 2002; Schaid, Rowland, Tines, Jacobson, and Poland 2002; Zaykin et al. 2002; Schaid 2004).

Determining haplotype requires the parental origin or gametic phase information, which cannot be easily obtained with the current genotyping technology. As a result, only the unphased genotype (i.e., the combination of the two homologous

D. Y. Lin is Dennis Gillings Distinguished Professor (E-mail: *lin@bios.unc.edu*) and D. Zeng is Assistant Professor (E-mail: *dzeng@bios.unc.edu*), Department of Biostatistics, CB#7420, University of North Carolina, Chapel Hill, NC 27599. The authors are grateful to the FUSION Study Group for sharing their data and to Michael Boehnke and Laura Scott for transmitting the data. They also thank the editor, an associate editor, and two referees for their comments. This work was supported by the National Institutes of Health.

Journal of the American Statistical Association
March 2006, Vol. 101, No. 473, Theory and Methods
DOI 10.1198/016214505000000808

haplotypes) can be determined. Statistically speaking, this is a missing-data problem in which the variable of interest pertains to two ordered sequences of 0's and 1's but only the summation of the two sequences is observed. This type of missing-data problem has not been studied in the statistical literature.

Many authors (e.g., Clark 1990; Excoffier and Slatkin 1995; Stephens et al. 2001; Zhang, Pakstis, Kidd, and Zhao 2001; Niu, Qin, Xu, and Liu 2002; Qin, Niu, and Liu 2002) have proposed methods to infer haplotypes or estimate haplotype frequencies from unphased genotype data. To make inference about haplotype effects, one may then relate the probabilistically inferred haplotypes to the phenotype through a regression model (e.g., Zaykin et al. 2002). This approach does not account for the variation due to haplotype estimation, and does not yield consistent estimators of regression parameters.

A growing number of articles have been published in genetic journals on making proper inference about the effects of haplotypes on disease phenotypes. Most of these articles have dealt with case-control studies. Specifically, Zhao, Li, and Khalid (2003) developed an estimating function that approximates the expectation of the complete-data prospective-likelihood score function given the observable data. This method assumes that the disease is rare and that haplotypes are independent of environmental variables, and it is not statistically efficient. Epstein and Satten (2003) derived a retrospective likelihood for the relative risk that does not accommodate environmental variables. Stram et al. (2003) proposed a conditional likelihood for the odds ratio assuming that cases and controls are chosen randomly with known probabilities from the target population, but did not consider environmental variables or investigate the properties of the estimator. Building on the earlier work of Schaid et al. (2002), Lake et al. (2003) discussed likelihood-based inference for cross-sectional studies under generalized linear models. Seltman, Roeder, and Devlin (2003) provided a similar discussion based on the cladistic approach. Recently, Lin (2004) showed how to perform Cox's (1972) regression when potentially censored age at onset of the disease observations are collected in cohort studies. All of the aforementioned work assumes Hardy–Weinberg equilibrium (Weir 1996, p. 40). Simulation studies (Epstein and Satten 2003; Lake et al. 2003; Satten and Epstein 2004) revealed that violation of this assumption can adversely affect the validity of the inference.

The aim of this article is to address statistical issues in estimating haplotype effects in a systematic and rigorous manner. For case-control studies, we allow environmental variables and derive efficient inference procedures. For cross-sectional and cohort studies, we consider more versatile models than those in the existing literature. For all study designs, we accommodate Hardy–Weinberg disequilibrium. We construct appropriate likelihoods for a variety of models. Under case-control sampling, the likelihood pertains to the distribution of genotypes and environmental variables conditional on the case-control status, which involves infinite-dimensional nuisance parameters if environmental variables are continuous. In cohort studies, it is desirable to not parameterize the distribution of time to disease, so that the likelihood also involves infinite-dimensional parameters. The presence of infinite-dimensional parameters entails considerable theoretical and computational challenges. We establish the theoretical properties of the maximum likelihood estimators (MLEs) by appealing to modern asymptotic techniques, and develop efficient and stable algorithms to implement the corresponding inference procedures. We assess the performance of the proposed methods through simulation studies, and provide an application to a major genetic study of type 2 diabetes mellitus.

2. INFERENCE PROCEDURES

2.1 Preliminaries

We consider SNP-based association studies of unrelated individuals. Suppose that each individual is genotyped at M biallelic SNPs within a candidate gene. At each SNP site, we indicate the two possible alleles by the values 0 and 1. Thus each haplotype h is a unique sequence of M numbers from $\{0, 1\}$. The total number of possible haplotypes is $K \equiv 2^M$; the actual number of haplotypes consistent with the data is usually much smaller. For $k = 1, \ldots, K$, let h_k denote the kth possible haplotype. Figure 1 shows the eight possible haplotypes for three SNPs.

Our human chromosomes come in pairs, one member of each pair inherited from our mother and the other member inherited from our father. These pairs are called homologous chromosomes. Thus each individual has a pair of homologous haplotypes that may or may not be identical. Routine genotyping procedures cannot separate the two homologous chromosomes, so only the (unphased) genotypes (i.e., the combinations of the two homologous haplotypes) are directly observable. For each individual, the multi-SNP genotype is an ordered sequence of M numbers from $\{0, 1, 2\}$.

Let H and G denote the pair of haplotypes and the genotype for an individual. We write $H = (h_k, h_l)$ if the individual's haplotypes are h_k and h_l, in which case $G = h_k + h_l$. The ordering of the two homologous haplotypes within an individual is considered arbitrary. By allowing genotypes to include missing SNP information, we may assume that G is known for each individual. Given G, the value of H is unknown if the individual is heterozygous at more than one SNP or if any SNP genotype is missing. For the case of $M = 3$ shown in Figure 1, if $G = (0, 2, 1)$, then $H = (h_3, h_4)$, and if $G = (0, 1, 1)$, then $H = (h_1, h_4)$ or $H = (h_2, h_3)$.

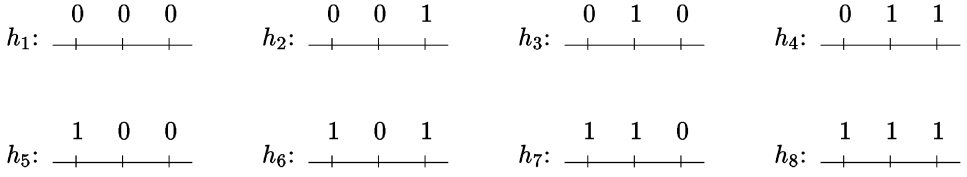

Figure 1. Possible Haplotype Configurations With Three SNPs.

The goal of the association studies is to relate the pair of haplotypes to disease phenotypes or traits. The simplest phenotype is the binary indicator for the disease status, which takes the value 1 if the individual is diseased and 0 otherwise. The diseased individuals may be further classified into several categories corresponding to different types of disease or varying degrees of disease severity. If the age of onset is likely to be genetically mediated, then it is desirable to use the age of onset as the phenotype. One may also be interested in disease-related traits, such as blood pressure.

The data on the disease phenotype may be gathered in various ways. The simplest approach is to obtain a random sample from the target population and measure the phenotype of interest on every individual in the sample. Such studies are referred to as cross-sectional studies, which are feasible if the disease is relatively frequent or if one is interested only in some readily measured traits that are related to the disease. If one is interested in the age at the onset of a disease, however, then it is necessary to follow a cohort of individuals forward in time, in which case the phenotype (i.e., time to disease occurrence) may be censored. When the disease is relatively rare, it is more cost-effective to use the case-control design, which collects data retrospectively on a sample of diseased individuals and on a separate sample of disease-free individuals. It is often desirable to collect data on environmental variables or covariates so as to investigate gene–environment interactions.

Let $\mathbf{Y}$ be the phenotype of interest, and let $\mathbf{X}$ be the covariates. For cross-sectional and case-control studies, the association between $\mathbf{Y}$ and $(\mathbf{X}, H)$ is characterized by the conditional density of $\mathbf{Y} = \mathbf{y}$ given $H = (h_k, h_l)$ and $\mathbf{X} = \mathbf{x}$, denoted by $P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(\mathbf{y}|\mathbf{x}, (h_k, h_l))$, where $\boldsymbol{\alpha}$ denotes the intercept(s), $\boldsymbol{\beta}$ denotes the regression effects, and $\boldsymbol{\xi}$ denotes the nuisance parameters (e.g., variance and overdispersion parameters). There is considerable flexibility in specifying the regression relationship. Suppose that h^* is the target haplotype of interest and that there are no covariates. Then a linear predictor in the form of $\alpha + \beta I(h_k = h_l = h^*)$ pertains to a recessive model, $\alpha + \beta\{I(h_k = h^*) + I(h_l = h^*) - I(h_k = h_l = h^*)\}$ pertains to a dominant model, $\alpha + \beta\{I(h_k = h^*) + I(h_l = h^*)\}$ pertains to an additive model, and $\alpha + \beta_1\{I(h_k = h^*) + I(h_l = h^*)\} + \beta_2 I(h_k = h_l = h^*)$ pertains to a codominant model, where $I(\cdot)$ is the indicator function. Clearly, the codominant model contains the other three models as special cases. A codominant model with gene–environment interactions has the following linear predictor:

$$\begin{aligned}\alpha &+ \beta_1\{I(h_k = h^*) + I(h_l = h^*)\} + \beta_2 I(h_k = h_l = h^*) \\ &+ \boldsymbol{\beta}_3^T\mathbf{x} + \boldsymbol{\beta}_4^T\{I(h_k = h^*) + I(h_l = h^*)\}\mathbf{x} \\ &+ \boldsymbol{\beta}_5^T I(h_k = h_l = h^*)\mathbf{x}. \qquad (1)\end{aligned}$$

Additional terms may be included so as to examine the effects of several haplotype configurations or to investigate the joint effects of multiple candidate genes.

Although we are interested in the effects of H and $\mathbf{X}$ on $\mathbf{Y}$, we observe G instead of H. As mentioned earlier, G is the summation of the paired sequences in H. Thus we have a regression problem with missing data in which the primary explanatory variable pertains to two ordered sequences of numbers from $\{0, 1\}$, but only the summation of the two sequences is observed. We assume that $\mathbf{X}$ is independent of H conditional on G and that $(1, \mathbf{X}^T)$ is linearly independent with positive probability.

Write $\pi_{kl} = P\{H = (h_k, h_l)\}$ and $\pi_k = P(h = h_k)$, $k, l = 1, \ldots, K$. As we demonstrate in this article, it is sometimes possible to make inference about haplotype effects without imposing any structures on $\{\pi_{kl}\}$, although estimating $\{\pi_k\}$ and testing for no haplotype effects require some restrictions on $\{\pi_{kl}\}$. Under Hardy–Weinberg equilibrium,

$$\pi_{kl} = \pi_k\pi_l, \qquad k, l = 1, \ldots, K. \qquad (2)$$

We consider two specific forms of departure from Hardy–Weinberg equilibrium,

$$\pi_{kl} = (1 - \rho)\pi_k\pi_l + \delta_{kl}\rho\pi_k \qquad (3)$$

and

$$\pi_{kl} = \frac{(1 - \rho + \delta_{kl}\rho)\pi_k\pi_l}{1 - \rho + \rho\sum_{j=1}^K \pi_j^2}, \qquad (4)$$

where $0 \le \pi_k \le 1$, $\sum_{k=1}^K \pi_k = 1$, $\delta_{kk} = 1$, and $\delta_{kl} = 0$ $(k \ne l)$. In (3), ρ is called the inbreeding coefficient or fixation index (Weir 1996, p. 93) and corresponds to Cohen's (1960) kappa measure of agreement. Equation (4) creates disequilibrium by giving different fitness values to the homozygous and heterozygous pairs (Niu et al. 2002). The denominator is a normalizing constant. Both (3) and (4) reduce to (2) if $\rho = 0$. Excess homozygosity (i.e., $\pi_{kk} > \pi_k^2, k = 1, \ldots, K$) arises when $\rho > 0$, and excess heterozygosity (i.e., $\pi_{kk} < \pi_k^2, k = 1, \ldots, K$) arises when $\rho < 0$. Recently, Satten and Epstein (2004) considered (3) for the control population under the case-control design. We abuse the notation slightly in that the $\{\pi_k\}$ in (4) do not pertain to the marginal distribution of H unless $\rho = 0$.

Let $\widetilde{h}$ denote a haplotype that differs from h at only one SNP. Write $\nabla_{\mathbf{u}} f(\mathbf{u}, \mathbf{v}) = \partial f(\mathbf{u}, \mathbf{v})/\partial\mathbf{u}$. The following lemma states that under (3) or (4), $\{\pi_k\}$ and ρ are identifiable from the data on G, and the data on G provide positive information about these parameters.

Lemma 1. Assume that either (3) or (4) holds. The parameters $\{\pi_k\}$ and ρ are uniquely determined by the distribution of G. For nondegenerate distribution $\{\pi_k\}$, if there exist a constant μ and a vector $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)^T$ such that $\sum_{k=1}^K \nu_k = 0$ and $\mu\nabla_\rho \log P(G = g) + \sum_{k=1}^K \nu_k \nabla_{\pi_k} \log P(G = g) = 0$ for $g = 2h$, then $\mu = 0$ and $\boldsymbol{\nu} = \mathbf{0}$.

In the sequel, $\mathcal{G}$ denotes the set of all possible genotypes and $\mathcal{S}(G)$ denotes the set of haplotype pairs that are consistent with genotype G. We suppose that $\pi_k > 0$ for all $k = 1, \ldots, K$, where K is now interpreted as the total number of haplotypes that exist in the population. For any parameter $\boldsymbol{\theta}$, we use $\boldsymbol{\theta}_0$ to denote its true value if the distinction is necessary. We assume that the true value of any Euclidean parameter $\boldsymbol{\theta}$ belongs to the interior of a known compact set within the domain of $\boldsymbol{\theta}$. Proofs of Lemma 1 and all of the theorems are provided in the Appendix.

2.2 Cross-Sectional Studies

There is a random sample of n individuals from the underlying population. The observable data consist of $(\mathbf{Y}_i, \mathbf{X}_i, G_i)$, $i = 1, \ldots, n$. The trait $\mathbf{Y}$ can be discrete or continuous, univariate or multivariate. As stated in Section 2.1, the conditional density of $\mathbf{Y}$ given $\mathbf{X}$ and H is given by $P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(\mathbf{Y}|\mathbf{X}, H)$. For a univariate trait, this regression model may take the form of a generalized linear model (McCullagh and Nelder 1989) with the linear predictor given in (1). If the trait is measured repeatedly in a longitudinal study, then generalized linear mixed models (Diggle, Heagerty, Liang, and Zeger 2002, chap. 9) may be used. The following conditions are required for estimating $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi})$.

Condition 1. If $P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(\mathbf{Y}|\mathbf{X}, H) = P_{\widetilde{\boldsymbol{\alpha}},\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\xi}}}(\mathbf{Y}|\mathbf{X}, H)$ for any $H = (h_k, h_k)$ and $H = (h_k, \widetilde{h}_k)$, $k = 1, \ldots, K$, then $\boldsymbol{\alpha} = \widetilde{\boldsymbol{\alpha}}$, $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}$, and $\boldsymbol{\xi} = \widetilde{\boldsymbol{\xi}}$.

Condition 2. If there exists a constant vector $\boldsymbol{\nu}$ such that $\boldsymbol{\nu}^T \nabla_{\underset{\sim}{\boldsymbol{\alpha}},\boldsymbol{\beta},\boldsymbol{\xi}} \log P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(\mathbf{Y}|\mathbf{X}, H) = 0$ for $H = (h_k, h_k)$ and $H = (h_k, \widetilde{h}_k)$, then $\boldsymbol{\nu} = \mathbf{0}$.

Remark 1. Condition 1 ensures that the parameters of interest are identifiable from the genotype data. The linear independence of the score function stated in Condition 2 ensures nonsingularity of the information matrix. The reason for considering $H = (h_k, h_k)$ and $H = (h_k, \widetilde{h}_k)$ is that these haplotype pairs can be inferred with certainty because of the unique decompositions of the corresponding genotypes $g = 2h_k$ and $g = h_k + \widetilde{h}_k$. All of the commonly used regression models, particularly generalized linear (mixed) models with linear predictors in the form of (1), satisfy Conditions 1 and 2.

We show in Section A.2.1 that it is possible to estimate the regression parameters without imposing any structure on the joint distribution of H. But this estimation requires knowledge of whether or not the dominant effects exist. Specifically, if there are no dominant effects, then only $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi})$ and $P(G = g)$ are identifiable; otherwise, $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi})$, $P(G = g)$, and $P(H = (h^*, g - h^*))$ are identifiable. If either (3) or (4) holds, then it follows from Lemma 1 and Condition 1 that all of the parameters are identifiable regardless of the genetic mechanism. Denote the joint distribution of H by $P_{\boldsymbol{\gamma}}(H = (h_k, h_l))$, where $\boldsymbol{\gamma}$ consists of the identifiable parameters in the distribution of H. Under (3) or (4), $\boldsymbol{\gamma} = (\rho, \pi_1, \ldots, \pi_K)^T$. When the distribution of H is unspecified, $\boldsymbol{\gamma}$ pertains to the aspects of the distribution of H that are identifiable.

Write $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi})$. The likelihood for $\boldsymbol{\theta}$ based on the cross-sectional data is proportional to

$$L_n(\boldsymbol{\theta}) \equiv \prod_{i=1}^{n} \prod_{g \in \mathcal{G}} \{m_g(\mathbf{Y}_i, \mathbf{X}_i; \boldsymbol{\theta})\}^{I(G_i = g)}, \tag{5}$$

where

$$m_g(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}) = \sum_{(h_k, h_l) \in \mathcal{S}(g)} P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}\big(\mathbf{y}|\mathbf{x}, (h_k, h_l)\big) P_{\boldsymbol{\gamma}}(h_k, h_l).$$

The MLE $\widehat{\boldsymbol{\theta}}$ can be obtained by maximizing (5) via the Newton–Raphson algorithm or an optimization algorithm. It is generally more efficient to use the expectation–maximization (EM) algorithm (Dempster, Laird, and Rubin 1977), especially when the distribution of H satisfies (3) with $\rho \geq 0$; see Section A.2.2 for details.

By the classical likelihood theory, we can show that $\widehat{\boldsymbol{\theta}}$ is consistent, asymptotically normal, and asymptotically efficient under Conditions 1 and 2 and the following condition.

Condition 3. If there exists a constant vector $\boldsymbol{\nu}$ such that $\boldsymbol{\nu}^T \nabla_{\boldsymbol{\theta}} \log m_g(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}_0) = 0$, then $\boldsymbol{\nu} = \mathbf{0}$.

Remark 2. Condition 3 ensures the nonsingularity of the information matrix. This condition can be easily verified when the joint distribution of H is unspecified and is implied by Lemma 1 and Condition 2 when the distribution satisfies (3) or (4).

2.3 Case-Control Studies With Known Population Totals

We consider case-control data supplemented by information on population totals (Scott and Wild 1997). There is a finite population of N individuals that is considered a random sample from the joint distribution of $(Y, \mathbf{X}, H)$, where Y is a categorical response variable. All that is known about this finite population is the total number of individuals in each category of $Y = y$. A sample of size n stratified on the disease status is drawn from the finite population, and the values of $\mathbf{X}$ and G are recorded for each sampled individual. The supplementary information on population totals is often available from hospital records, cancer registries, and official statistics. If a case-control sample is drawn from a cohort study, then the cohort serves as the finite population. The observable data consist of $(Y_i, R_i, R_i\mathbf{X}_i, R_iG_i)$, $i = 1, \ldots, N$, where R_i indicates, by the values 1 versus 0, whether or not the ith individual in the finite population is selected into the case-control sample.

The association between Y and $(\mathbf{X}, H)$ is characterized by $P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(Y|\mathbf{X}, H)$, where $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\xi}$ pertain to the intercept(s), regression effects, and overdispersion parameters (McCullagh and Nelder 1989). In the case of a binary response variable, important examples of $P_{\alpha,\boldsymbol{\beta},\xi}(Y|\mathbf{X}, H)$ include the logistic, probit, and complementary log–log regression models. When there are more than two categories, examples include the proportional odds, multivariate probit, and multivariate logistic regression models. Because the data associated with $R_i = 1$ yield the same form of likelihood as that of a cross-sectional study and the data associated $R_i = 0$ yield a missing-data likelihood, all of the identifiability results stated in Section 2.2 apply to the current setting. We again write $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ consists of the identifiable parameters in the distribution of H.

Let $F_g(\cdot)$ be the cumulative distribution function of $\mathbf{X}$ given $G = g$, and let $f_g(\mathbf{x})$ be the density of $F_g(\mathbf{x})$ with respect to a dominating measure $\mu(\mathbf{x})$. Note that $F_g(\cdot)$ is infinite-dimensional if $\mathbf{X}$ has continuous components. The joint density of $(Y = y, G = g, \mathbf{X} = \mathbf{x})$ is $m_g(y, \mathbf{x}; \boldsymbol{\theta}) f_g(\mathbf{x})$. The likelihood concerning $\boldsymbol{\theta}$ and $\{F_g\}$ takes the form

$$L_n(\boldsymbol{\theta}, \{F_g\}) = \prod_{i=1}^{N} \left[\prod_{g \in \mathcal{G}} \{m_g(Y_i, \mathbf{X}_i; \boldsymbol{\theta}) f_g(\mathbf{X}_i)\}^{I(G_i = g)} \right]^{R_i} \times \left[\sum_{g \in \mathcal{G}} \int m_g(Y_i, \mathbf{x}; \boldsymbol{\theta})\, dF_g(\mathbf{x}) \right]^{1 - R_i}. \tag{6}$$

Unlike the likelihood for the cross-sectional design given in (5), the density functions of $\mathbf{X}$ given G cannot be factored out of the likelihood given in (6) and thus cannot be omitted from the likelihood.

We maximize (6) to obtain the MLEs $\widehat{\boldsymbol{\theta}}$ and $\{\widehat{F}_g(\cdot)\}$. The latter is an empirical function with point masses at the observed $\mathbf{X}_i$ such that $G_i = g$ and $R_i = 1$. The maximization can be carried out via the Newton–Raphson, profile-likelihood, or large-scale optimization methods. An alternative way to calculate the MLEs is through the EM algorithm described in Section A.3.1.

We impose the following regularity condition, and then state the asymptotic results in Theorem 1.

Condition 4. For any $g \in \mathcal{G}$, $f_g(\mathbf{x})$ is positive in its support and continuously differentiable with respect to a suitable measure.

Theorem 1. Under Conditions 1–4, $\widehat{\boldsymbol{\theta}}$ and $\{\widehat{F}_g(\cdot)\}$ are consistent in that $|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| + \sup_{\mathbf{x},g} |\widehat{F}_g(\mathbf{x}) - F_g(\mathbf{x})| \to 0$ almost surely. In addition, $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a mean 0 normal random vector whose covariance matrix attains the semiparametric efficiency bound.

Let $pl_n(\boldsymbol{\theta})$ be the profile log-likelihood for $\boldsymbol{\theta}$, that is, $pl_n(\boldsymbol{\theta}) = \max_{\{F_g\}} \log L_n(\boldsymbol{\theta}, \{F_g\})$. Then the (s,t)th element of the inverse covariance matrix of $\widehat{\boldsymbol{\theta}}$ can be estimated by $-\epsilon_n^{-2}\{pl_n(\widehat{\boldsymbol{\theta}} + \epsilon_n\mathbf{e}_s + \epsilon_n\mathbf{e}_t) - pl_n(\widehat{\boldsymbol{\theta}} + \epsilon_n\mathbf{e}_s - \epsilon_n\mathbf{e}_t) - pl_n(\widehat{\boldsymbol{\theta}} - \epsilon_n\mathbf{e}_s + \epsilon_n\mathbf{e}_t) + pl_n(\widehat{\boldsymbol{\theta}})\}$, where ϵ_n is a constant of the order $n^{-1/2}$ and $\mathbf{e}_s$ and $\mathbf{e}_t$ are the sth and tth canonical vectors. The function $pl_n(\boldsymbol{\theta})$ can be calculated via the EM algorithm by holding $\boldsymbol{\theta}$ constant in both the E-step and the M-step.

Remark 3. If N is much larger than n or if the population frequencies rather than the totals are known, then we maximize $\prod_{i=1}^n \prod_{g\in\mathcal{G}}\{m_g(Y_i, \mathbf{X}_i; \boldsymbol{\theta}) f_g(\mathbf{X}_i)\}^{I(G_i=g)}$ subject to the constraints that $\sum_{g\in\mathcal{G}} \int m_g(y, \mathbf{x}; \boldsymbol{\theta})\, dF_g(\mathbf{x}) = p_y$, where p_y is the population frequency of $Y = y$. The resultant estimator of $\boldsymbol{\theta}_0$ is consistent, asymptotically normal, and asymptotically efficient. The results in this section can be extended straightforwardly to accommodate stratifications on covariates.

2.4 Case-Control Studies With Unknown Population Totals

We consider the classical case-control design, which measures $\mathbf{X}$ and G on n_1 cases ($Y = 1$) and n_0 controls ($Y = 0$) and requires no knowledge about the finite population. With the notation introduced in the previous section, the likelihood contribution from one individual takes the form

$$RL(\boldsymbol{\theta}, \{F_g\}) = \frac{\prod_{g\in\mathcal{G}}\{m_g(y, \mathbf{X}; \boldsymbol{\theta}) f_g(\mathbf{X})\}^{I(G=g)}}{\sum_{g\in\mathcal{G}} \int m_g(y, \mathbf{x}; \boldsymbol{\theta})\, dF_g(\mathbf{x})}, \qquad (7)$$

where we use y instead of Y to emphasize that y is not random.

Define

$$f_g^\dagger(\mathbf{x}) = \frac{m_g(0, \mathbf{x}; \boldsymbol{\theta}) f_g(\mathbf{x})}{\int m_g(0, \mathbf{x}; \boldsymbol{\theta})\, dF_g(\mathbf{x})},$$

$$q_g = \frac{\int m_g(0, \mathbf{x}; \boldsymbol{\theta})\, dF_g(\mathbf{x})}{\sum_{\tilde{g}\in\mathcal{G}} \int m_{\tilde{g}}(0, \mathbf{x}; \boldsymbol{\theta})\, dF_{\tilde{g}}(\mathbf{x})}.$$

Clearly, $f_g^\dagger(\mathbf{x})$ is the conditional density of $\mathbf{X}$ given $G = g$ and $Y = 0$, and q_g is the conditional probability of $G = g$ given $Y = 0$. Let g_0 and $\mathbf{x}_0$ be some specific values of G and $\mathbf{X}$. Write $F_g^\dagger(\mathbf{x}) = \int_0^{\mathbf{x}} f_g^\dagger(\mathbf{s})\, d\mu(\mathbf{s})$. We can express (7) as

$$RL(\boldsymbol{\theta}, \{F_g^\dagger\}, \{q_g\}) = \frac{\prod_{g\in\mathcal{G}}\{\eta(y, \mathbf{X}, g; \boldsymbol{\theta}) f_g^\dagger(\mathbf{x}) q_g\}^{I(G=g)}}{\sum_{g\in\mathcal{G}} q_g\{\int \eta(y, \mathbf{x}, g; \boldsymbol{\theta})\, dF_g^\dagger(\mathbf{x})\}}, \qquad (8)$$

where

$$\eta(y, \mathbf{x}, g; \boldsymbol{\theta}) = \frac{m_g(y, \mathbf{x}; \boldsymbol{\theta}) m_{g_0}(0, \mathbf{x}_0; \boldsymbol{\theta})}{m_g(0, \mathbf{x}; \boldsymbol{\theta}) m_{g_0}(y, \mathbf{x}_0; \boldsymbol{\theta})}.$$

We call η the generalized odds ratio (Liang and Qin 2000), which reduces to the ordinary odds ratio when $\mathcal{S}(g)$ is a singleton.

Remark 4. The parameter q_g is a functional of $f_g^\dagger$ and $\boldsymbol{\theta}$ because $\int m_g(0, \mathbf{x}; \boldsymbol{\theta})\, dF_g(\mathbf{x}) = \{\int m_g^{-1}(0, \mathbf{x}; \boldsymbol{\theta})\, dF_g^\dagger(\mathbf{x})\}^{-1}$. This constraint makes it very difficult to study the identifiability of the parameters. Thus we treat q_g as a free parameter in our development.

For traditional case-control data, the odds ratio is identifiable (whereas the intercept is not), and its MLE can be obtained by maximizing the prospective likelihood (Prentice and Pyke 1979). Similar results hold when the exposure is measured with error (Roeder, Carroll, and Lindsay 1996); however, the distribution of the measurement error needs to be estimated from a validation set or an external source. With unphased genotype data, identifiability is much more delicate. We show in Section A.4.1 that the components of $\boldsymbol{\theta}$ that are identifiable from the retrospective likelihood are exactly those that are identifiable from the generalized odds ratio. Thus we assume that the generalized odds ratio depends only on a set of identifiable parameters, still denoted by $\boldsymbol{\theta}$; otherwise, the inference is not tractable. For the logistic link function with linear predictor (1), we show in Section A.4.2 that if there are no dominant effects, then $\boldsymbol{\theta}$ consists only of $\boldsymbol{\beta}$; if there are no covariate effects but there exists a dominant main effect, then $\boldsymbol{\beta}$ is identifiable and $P(H = (h^*, g - h^*))/P(G = g)$ is identifiable up to a scalar constant; and if the dominant effect depends on a continuous covariate or if the dominant main effect and the main effect of a continuous covariate are nonzero, then $\boldsymbol{\theta}$ consists of α, $\boldsymbol{\beta}$, and $P(H = (h^*, g - h^*))/P(G = g)$. For the probit and complementary log–log link functions, we show in Section A.4.3 that if there are dominant effects and at least one continuous covariate has an effect, then $\boldsymbol{\theta}$ consists of α, $\boldsymbol{\beta}$, and $P(H = (h^*, g - h^*))/P(G = g)$.

We maximize the product of (8) over the $n \equiv n_1 + n_0$ individuals in the case-control sample to produce the MLEs $\widehat{\boldsymbol{\theta}}$, $\{\widehat{F}_g^\dagger(\cdot)\}$, and $\{\widehat{q}_g\}$. Although the $\{F_g^\dagger(\cdot)\}$ are high-dimensional, we show in Section A.4.4 that $\widehat{\boldsymbol{\theta}}$ can be obtained by profiling a likelihood function over a scalar nuisance parameter.

To state the asymptotic properties of the MLEs, we impose the following conditions.

Condition 5. If there exists a vector $\mathbf{v}$ such that $\mathbf{v}^T \nabla_{\boldsymbol{\theta}} \log \eta(1, \mathbf{x}, g; \boldsymbol{\theta})$ is a constant with probability 1, then $\mathbf{v} = \mathbf{0}$.

Condition 6. The function $f_g^\dagger$ is positive in its support and continuously differentiable.

Condition 7. The fraction $n_1/n \to \varrho \in (0, 1)$.

Remark 5. Condition 5 implies nonsingularity of the information matrix for $\boldsymbol{\theta}_0$ and can be shown to hold for the logistic, probit, and complementary log–log link functions. Condition 7 ensures that there are both cases and controls in the sample.

Theorem 2. Under Conditions 5–7, $|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| + \sup_g |\widehat{q}_g - q_g| + \sup_{\mathbf{x},g} |\widehat{F}_g^{\dagger}(\mathbf{x}) - F_g^{\dagger}(\mathbf{x})| \to 0$ almost surely. In addition, $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a normal random vector whose covariance matrix attains the semiparametric efficiency bound.

In most case-control studies, the disease is (relatively) rare. When the disease is rare, considerable simplicity arises because of the following approximation for the logistic regression model:

$$P_{\alpha,\boldsymbol{\beta}}(Y|\mathbf{X}, H) \approx \exp\{Y(\alpha + \boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X}, H))\},$$

where $\mathcal{Z}(\mathbf{X}, H)$ is a specific function of $\mathbf{X}$ and H. We assume that either (3) or (4) holds. The likelihood based on $(\mathbf{X}_i, G_i, y_i)$, $i = 1, \ldots, n$, can be approximated by

$$\begin{aligned}\widetilde{L}_n(\boldsymbol{\theta}, \{F_g\}) &= \prod_{i=1}^{n} \left(\frac{\prod_{g\in\mathcal{G}} [f_g(\mathbf{X}_i) \sum_{(h_k,h_l)\in\mathcal{S}(g)} e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X}_i,h_k,h_l)} P_{\boldsymbol{\gamma}}(h_k,h_l)]^{I(G_i=g)}}{\sum_{g\in\mathcal{G}} \int_{\mathbf{x}} \sum_{(h_k,h_l)\in\mathcal{S}(g)} e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{x},h_k,h_l)} P_{\boldsymbol{\gamma}}(h_k,h_l)\, dF_g(\mathbf{x})} \right)^{y_i} \\ &\quad \times \left[\prod_{g\in\mathcal{G}} \left\{ f_g(\mathbf{X}_i) \sum_{(h_k,h_l)\in\mathcal{S}(g)} P_{\boldsymbol{\gamma}}(h_k,h_l) \right\}^{I(G_i=g)} \right]^{1-y_i}. \end{aligned} \tag{9}$$

We impose the following condition.

Condition 8. If $\alpha + \boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X}, H) = \widetilde{\alpha} + \widetilde{\boldsymbol{\beta}}^T \mathcal{Z}(\mathbf{X}, H)$ for $H = (h_k, h_k)$ and $H = (h_k, \widetilde{h}_k)$, then $\alpha = \widetilde{\alpha}$ and $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}$.

This condition is similar to Condition 1 stated in Section 2.2, and it holds for the codominant model. Under this condition, it follows from Lemma 1 that no two sets of parameters can give the same likelihood with probability 1. Thus the maximizer of (9), denoted by $(\widehat{\boldsymbol{\theta}}, \{\widehat{F}_g\})$, is locally unique. We show in Section A.4.5 that $\widehat{\boldsymbol{\theta}}$ can be easily obtained by profiling over a small number of parameters.

To derive the asymptotic properties, we provide a mathematical definition of rare disease.

Condition 9. For $i = 1, \ldots, n$, the conditional distribution of Y_i given $(\mathbf{X}_i, H_i)$ satisfies that $P(Y_i = 1|\mathbf{X}_i, H_i) = a_n \exp\{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X}_i, H_i)\}/[1 + a_n \exp\{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X}_i, H_i)\}]$, where $a_n = o(n^{-1/2})$.

Theorem 3. Under Conditions 6–9, $|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| + \sup_{\mathbf{x},g} |\widehat{F}_g(\mathbf{x}) - F_g(\mathbf{x})| \to^{P_n} 0$, where P_n is the probability measure given by Condition 9. Furthermore, $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a normal random vector whose covariance matrix achieves the semiparametric efficiency bound.

2.5 Cohort Studies

In a cohort study, Y represents the time to disease occurrence, which is subject to right-censorship by C. The data consist of $(\widetilde{Y}_i, \Delta_i, \mathbf{X}_i, G_i)$, $i = 1, \ldots, n$, where $\widetilde{Y}_i = \min(Y_i, C_i)$ and $\Delta_i = I(Y_i \le C_i)$. We relate Y_i to $(\mathbf{X}_i, H_i)$ through a class of semiparametric linear transformation models,

$$\Gamma(Y_i) = -\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X}_i, H_i) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{10}$$

where Γ is an unknown increasing function, $\mathcal{Z}(\mathbf{X}, H)$ is a known function of $\mathbf{X}$ and H, and the ϵ_i's are independent errors with known distribution function F. We may rewrite (10) as

$$P(Y_i \le t|\mathbf{X}_i, H_i) = Q\big(\Lambda(t) e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X}_i,H_i)}\big),$$

where $\Lambda(t) = e^{\Gamma(t)}$ and $Q(x) = F(\log x)$ $(x > 0)$. The choices of the extreme-value and standard logistic distributions for F, or equivalently, $Q(x) = 1 - e^{-x}$ and $Q(x) = 1 - (1+x)^{-1}$, yield the proportional hazards model and the proportional odds model (Pettitt 1984).

We impose Condition 8. Under this condition, $\boldsymbol{\beta}$ and $\Lambda(\cdot)$ are identifiable from the observable data. The identifiability of the distribution of H is the same here as in the case of cross-sectional studies. Under (3) or (4) and Condition 8, all of the parameters, including $\boldsymbol{\beta}$, $\Lambda(\cdot)$, and $\boldsymbol{\gamma}$, are identifiable. This is shown in Section A.5.1.

The following assumption on censoring is required in constructing the likelihood.

Condition 10. Conditional on $\mathbf{X}$ and G, the censoring time C is independent of Y and H.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$. The likelihood concerning $\boldsymbol{\theta}$ and Λ takes the form

$$\begin{aligned} L_n(\boldsymbol{\theta}, \Lambda) &= \prod_{i=1}^{n} \Bigg[\sum_{(h_k,h_l)\in\mathcal{S}(G_i)} \{\dot{\Lambda}(\widetilde{Y}_i) e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X}_i,(h_k,h_l))} \\ &\qquad \times \dot{Q}\big(\Lambda(\widetilde{Y}_i) e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X}_i,(h_k,h_l))}\big)\}^{\Delta_i} \\ &\quad \times \{1 - Q\big(\Lambda(\widetilde{Y}_i) e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X}_i,(h_k,h_l))}\big)\}^{1-\Delta_i} P_{\boldsymbol{\gamma}}(h_k,h_l) \Bigg]. \end{aligned} \tag{11}$$

Here and in the sequel, $\dot{f}(x) = df(x)/dx$ and $\ddot{f}(x) = d^2 f(x)/dx^2$. Like (6), (8), and (9), this likelihood involves infinite-dimensional parameters. If Λ is restricted to be absolutely continuous, then, as in the case of density estimation, there is no maximizer of this likelihood. Thus we relax Λ to be right-continuous and replace $\dot{\Lambda}(\widetilde{Y}_i)$ in (11) by the jump size of Λ at $\widetilde{Y}_i$. By the arguments of Zeng, Lin, and Lin (2005), the resultant MLE, denoted by $(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})$, exists, and $\widehat{\Lambda}$ is a step function with jumps only at the observed $\widetilde{Y}_i$ for which $\Delta_i = 1$. The maximization can be carried out through an optimization algorithm. Furthermore, the covariance matrix of $\widehat{\boldsymbol{\theta}}$ can be estimated by the profile likelihood method, as discussed by Zeng et al. (2005).

Lin (2004) considered the special case of the proportional hazards model under condition (2) and provided an EM algorithm for obtaining the MLEs. We can modify that algorithm to accommodate Hardy–Weinberg disequilibrium along the lines of Section A.2.2. In addition, the EM algorithm can be used to evaluate the profile likelihood.

We assume the following regularity conditions for the asymptotic results.

Condition 11. There exists some positive constant δ_0 such that $P(C_i \geq \tau | \mathbf{X}_i, G_i) = P(C_i = \tau | \mathbf{X}_i, G_i) \geq \delta_0$ almost surely, where τ corresponds to the end of the study.

Condition 12. The true value $\Lambda_0(t)$ of $\Lambda(t)$ is a strictly increasing function in $[0, \tau]$ and is continuously differentiable. In addition, $\Lambda_0(0) = 0$, $\Lambda_0(\tau) < \infty$, and $\dot{\Lambda}_0(0) > 0$.

Theorem 4. Under Conditions 8 and 10–12, $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, \widehat{\Lambda} - \Lambda_0)$ converges weakly to a Gaussian process in $\mathbb{R}^d \times l^\infty([0, \tau])$, where d is the dimension of $\boldsymbol{\theta}_0$, and $l^\infty([0, \tau])$ is the space of all bounded functions on $[0, \tau]$ equipped with the supremum norm. Furthermore, $\widehat{\boldsymbol{\theta}}$ is asymptotically efficient.

3. SIMULATION STUDIES

We used Monte Carlo simulation to evaluate the proposed methods in realistic settings. We considered the five SNPs on chromosome 22 from the Finland–United States Investigation of NIDDM Genetics (FUSION) Study described in the next section. We obtained the π_k's from the frequencies shown in Table 1 by assuming a 7% disease rate, and generated haplotypes under (3) with $\rho = .05$. The R_h^2 in Table 1 is the measure of haplotype certainty of Stram et al. (2003). We focused on $h^* = (0, 1, 1, 0, 0)$ and considered case-control and cohort studies.

For the cohort studies, we generated ages of onset from the proportional hazards model,

$$\lambda\{t|x, (h_k, h_l)\} = 2t \exp\big[\beta_1\{I(h_k = h^*) + I(h_l = h^*)\} + \beta_2 x + \beta_3\{I(h_k = h^*) + I(h_l = h^*)\}x\big],$$

where X is a Bernoulli variable with $P(X = 1) = .2$ that is independent of H. We generated the censoring times from the uniform $(0, \tau)$ distribution, where τ was chosen to yield approximately 250, 500, and 1,000 cases under $n = 5{,}000$. We let $\beta_1 = \beta_2 = .25$ and varied β_3 from $-.5$ to .5.

As shown in Table 2, the MLE is virtually unbiased, the likelihood ratio test has proper type I error, and the confidence interval has reasonable coverage. Additional simulation studies revealed that the proposed methods also perform well for making inference about other parameters and under other genetic models.

Table 1. Observed Haplotype Frequencies in the FUSION Study

	Frequencies		
Haplotype	Controls	Cases	R_h^2
00011	.0042	.0066	.388
00100	.0035	.0034	.336
00110	.0018	.0007	.377
01011	.1292	.1344	.592
01100	.2514	.3183	.738
01101	.0012	$<10^{-4}$	.450
01110	$<10^{-4}$	.0045	.499
01111	.0019	$<10^{-4}$	.325
10000	.0136	.0114	.456
10010	$<10^{-4}$	.0012	.500
10011	.3573	.2883	.727
10100	.0521	.0597	.402
10110	.0317	.0318	.554
11011	.1392	.1290	.560
11100	.0109	.0092	.266
11110	$<10^{-4}$	.0014	$<10^{-4}$
11111	.0020	$<10^{-4}$	.338

Table 2. Simulation Results for the Haplotype–Environment Interactions in Cohort Studies

β_3	Cases	Bias	SE	CP	Power
0	250	−.010	.232	.949	.051
	500	−.005	.157	.953	.047
	1,000	−.003	.114	.954	.046
−.25	250	−.014	.256	.950	.190
	500	−.008	.172	.949	.334
	1,000	−.004	.122	.952	.554
−.5	250	−.022	.281	.950	.505
	500	−.011	.190	.950	.806
	1,000	−.006	.132	.952	.976
.25	250	−.007	.216	.947	.207
	500	−.002	.146	.953	.395
	1,000	−.001	.109	.954	.614
.5	250	−.003	.204	.943	.693
	500	−.001	.140	.951	.940
	1,000	−.001	.105	.952	.998

NOTE: Bias and SE are the bias and standard error of $\widehat{\beta}_3$. CP is the coverage probability of the 95% confidence interval for β_3. Power pertains to the .05-level likelihood ratio test of $H_0 : \beta_3 = 0$. Each entry is based on 5,000 replicates.

For the case-control studies, we used the same distributions of H and X and considered the same h^* as in the cohort studies. We generated disease incidence from the logistic regression model,

$$\begin{aligned}\operatorname{logit} P\{Y = 1|x, (h_k, h_l)\} &= \alpha + \beta_1\{I(h_k = h^*) + I(h_l = h^*)\} \\ &\quad + \beta_2 x + \beta_3\{I(h_k = h^*) + I(h_l = h^*)\}x. \qquad (12)\end{aligned}$$

For making inference on β_1, we set $\beta_2 = \beta_3 = .25$ and varied β_1 from $-.5$ to .5; for making inference on β_3, we set $\beta_1 = \beta_2 = .25$ and varied β_3 from $-.5$ to .5. We chose $\alpha = -3$ or -4, yielding disease rates between 1.6% and 7%. We let $n_1 = n_0 = 500$ or 1,000. We considered the situations of known and unknown population totals, with N being 15 and 30 times of n under $\alpha = -3$ and -4. For known population totals, we used the EM algorithm described in Section A.3.1 and evaluated the inference procedures based on the likelihood ratio statistic. For unknown population totals, we used the profile-likelihood method for rare diseases described in Section A.4.5 and set the $\widehat{\pi}_k$ less than $2/n$ to 0 to improve numerical stability. The results for β_1 and β_3 are displayed in Tables 3 and 4.

For known population totals, the proposed estimators are virtually unbiased, and the likelihood ratio statistics yield proper tests and confidence intervals. For unknown population totals, $\widehat{\beta}_1$ has little bias, especially for large n, whereas $\widehat{\beta}_3$ tends to be slightly biased downward; the variance estimators are fairly accurate, and the corresponding confidence intervals have reasonable coverage probabilities except for $\{\alpha = -3, \beta_3 = .5\}$. The method with known population totals yields slightly higher power than the method with unknown population totals.

All the aforementioned results pertain to haplotype 01100, which has a relatively high frequency and a large value of R_h^2; the covariate is binary, and ρ is .05, which is relatively large. Additional simulation studies showed that the foregoing conclusions continue to hold for other haplotypes, other values of ρ, and continuous covariates. Table 5 reports some results for haplotype 10100, which has a frequency of about 5% and

Table 3. Simulation Results for the Main Effects of the Haplotype in Case-Control Studies

			Known totals				Unknown totals				
$n_1 = n_0$	α	β_1	Bias	SE	CP	Power	Bias	SE	SEE	CP	Power
500	−3	−.5	−.003	.117	.952	.987	.019	.121	.124	.951	.979
		−.25	−.002	.109	.954	.587	.014	.112	.117	.960	.525
		0	−.001	.104	.951	.049	.009	.109	.112	.955	.045
		.25	−.001	.102	.950	.641	.002	.105	.108	.961	.646
		.5	.000	.099	.948	.996	−.005	.103	.106	.958	.998
	−4	−.5	.001	.112	.954	.987	.022	.119	.124	.951	.977
		−.25	−.002	.104	.955	.574	.013	.114	.117	.952	.529
		0	−.002	.100	.953	.047	.004	.109	.112	.953	.047
		.25	−.001	.095	.956	.636	−.003	.103	.108	.959	.640
		.5	−.000	.094	.950	.999	−.009	.102	.105	.956	.997
1,000	−3	−.5	−.003	.082	.953	1.00	.005	.087	.087	.949	1.00
		−.25	−.002	.076	.952	.874	.005	.081	.082	.948	.853
		0	−.001	.073	.951	.049	.005	.077	.077	.954	.046
		.25	−.001	.071	.953	.898	.004	.075	.076	.948	.920
		.5	−.001	.070	.953	1.00	.003	.075	.075	.946	1.00
	−4	−.5	.000	.079	.952	1.00	.005	.087	.088	.947	1.00
		−.25	−.000	.074	.959	.867	.005	.081	.083	.954	.847
		0	−.001	.070	.955	.045	.002	.079	.079	.949	.051
		.25	−.001	.067	.956	.904	.000	.074	.076	.955	.909
		.5	−.001	.066	.956	1.00	−.002	.073	.074	.954	1.00

NOTE: Bias and SE are the bias and standard error of $\widehat{\beta}_1$. SEE is the mean of the standard error estimator for $\widehat{\beta}_1$. CP is the coverage probability of the 95% confidence interval for β_1. Power pertains to the .05-level test of $H_0 : \beta_1 = 0$. Each entry is based on 5,000 replicates.

an R_h^2 of .4. We generated disease incidence from the logistic regression model

$$\begin{aligned}&\operatorname{logit} P\{Y = 1|X_1, X_2, (h_k, h_l)\}\\ &\quad = \alpha + \beta_h\{I(h_k = h^*) + I(h_l = h^*)\}\\ &\qquad + \beta_{x_1}X_1 + \beta_{x_2}X_2 + \beta_{hx_2}\{I(h_k = h^*) + I(h_l = h^*)\}X_2,\end{aligned}$$

where $h^* = (10100)$, X_1 is Bernoulli with .2 success probability, and X_2 is uniform(0, 1). We set $\rho = .01$, $\alpha = -3.7$, $\beta_h = 0$, and $\beta_{x_1} = \beta_{x_2} = -\beta_{x_2h} = .5$, yielding an overall disease rate of 7%. We assumed unknown population totals and used the profile-likelihood method for rare diseases described in Section A.4.5. The method performed remarkably well.

4. APPLICATION TO THE FUSION STUDY

Type 2 diabetes mellitus or non–insulin-dependent diabetes mellitus is a complex disease characterized by resistance of peripheral tissues to insulin and a deficiency of insulin secretion. Approximately 7% of adults in developed countries suffer from the disease. The FUSION study is a major effort to map and clone genetic variants that predispose to type 2 diabetes (Valle et al. 1998). We consider a subset of data from this study.

Table 4. Simulation Results for the Haplotype–Environment Interactions in Case-Control Studies

			Known totals				Unknown totals				
$n_1 = n_0$	α	β_3	Bias	SE	CP	Power	Bias	SE	SEE	CP	Power
500	−3	−.5	−.008	.205	.949	.729	.030	.187	.195	.953	.692
		−.25	−.002	.186	.949	.271	.016	.169	.176	.961	.244
		0	−.001	.173	.946	.054	−.006	.155	.162	.963	.037
		.25	.002	.165	.949	.334	−.038	.144	.151	.958	.287
		.5	.006	.161	.947	.885	−.088	.138	.143	.915	.831
	−4	−.5	−.009	.198	.950	.763	.012	.194	.195	.950	.720
		−.25	−.005	.181	.949	.309	.006	.172	.176	.953	.264
		0	−.002	.168	.945	.055	−.007	.156	.161	.956	.044
		.25	−.001	.157	.944	.370	−.022	.146	.149	.948	.333
		.5	.001	.148	.945	.926	−.047	.136	.141	.945	.904
1,000	−3	−.5	−.004	.147	.943	.953	.027	.134	.136	.950	.953
		−.25	−.003	.133	.946	.493	.013	.122	.123	.949	.477
		0	−.001	.123	.951	.049	−.005	.114	.113	.948	.052
		.25	.000	.119	.945	.580	−.034	.107	.106	.934	.535
		.5	.002	.117	.947	.994	−.080	.102	.101	.870	.986
	−4	−.5	−.005	.140	.945	.965	.010	.137	.136	.949	.965
		−.25	−.002	.128	.945	.529	.005	.124	.123	.951	.505
		0	−.001	.119	.946	.054	−.004	.113	.113	.947	.053
		.25	−.000	.110	.947	.633	−.016	.104	.105	.952	.601
		.5	.002	.105	.949	.998	−.037	.099	.099	.937	.995

NOTE: Bias and SE are the bias and standard error of $\widehat{\beta}_3$. SEE is the mean of the standard error estimator for $\widehat{\beta}_3$. CP is the coverage probability of the 95% confidence interval for β_3. Power pertains to the .05-level test of $H_0 : \beta_3 = 0$. Each entry is based on 5,000 replicates.

Table 5. Simulation Results for Haplotype 10100 in Case-Control Studies

$n_1 = n_0$	Parameter	True value	Bias	SE	SEE	CP	Power
500	β_h	0	−.030	.400	.401	.957	.043
	β_{x_1}	.5	.002	.151	.152	.951	.917
	β_{x_2}	.5	.001	.228	.230	.953	.584
	β_{hx_2}	−.5	.015	.641	.644	.956	.118
1,000	β_h	0	−.017	.275	.277	.953	.047
	β_{x_1}	.5	.002	.107	.107	.954	.997
	β_{x_2}	.5	.000	.162	.161	.950	.871
	β_{hx_2}	−.5	.012	.441	.443	.950	.198

NOTE: Bias and SE are the bias and standard error of the parameter estimator. SEE is the mean of the standard error estimator. CP is the coverage probability of the 95% confidence interval. Power pertains to the .05-level test of zero parameter value. Each entry is based on 5,000 replicates.

A total of 796 cases and 415 controls were genotyped at 5 SNPs in a putative susceptibility region on chromosome 22, with 131 cases and 82 controls having missing genotype information for at least one SNP. If G_i is missing, then the set $\mathcal{S}(G_i)$ is enlarged accordingly in the analysis. Table 1 displays the estimated haplotype frequencies under (3) separated by the cases and controls, along with the values of R_h^2 (Stram et al. 2003) for the controls. We estimated ρ at .000 for controls and .002 for cases.

We use the method based on (9) to estimate the effects of the haplotypes whose observed frequencies in the controls are greater than 2%. As shown in Table 6, the results are significant for the two most common haplotypes; haplotype 01100 increases the risk of disease, whereas haplotype 10011 is protective against diabetes. Epstein and Satten (2003) also reported the estimates for these two haplotypes, which agree with our numbers. Although they did not report standard error estimates, their confidence intervals are similar to those based on Table 6. The results under the codominant model as well as the calculations of the Akaike information criterion (AIC) (Akaike 1985) suggest that the additive model fits the data the best for both haplotypes 01100 and 10011.

The FUSION investigators are currently exploring gene–environment interactions on chromosome 22, so the covariate information is confidential at this stage. To illustrate our method for detecting gene–environment interactions, we artificially created a binary covariate X by setting $X = 1$ for the first 600 individuals in the dataset. Under the additive genetic model for haplotype 01100, the estimate of the interaction is .043 with an estimated standard error of .110. For further illustration, we generated a binary covariate from the conditional distribution of X given Y and G under model (12) with $\alpha = -3.7$, $\beta_1 = .32$, and $\beta_2 = .25$. Based on 5,000 replicates, the power for testing $H_0 : \beta_3 = 0$ is estimated at .053, .479, or .974 under $\beta_3 = 0, .25$, or .5.

5. DISCUSSION

Inferring haplotype–disease associations is an interesting and difficult statistical problem. The presence of infinite-dimensional nuisance parameters in the likelihoods for case-control and cohort studies entails considerable theoretical and computational challenges. Although we have conducted a systematic and rigorous investigation, providing powerful new methods, there remain substantial open problems. Here we discuss some directions for future research.

Case-Control Studies. It is numerically difficult to maximize (6) when N is much larger than n, and algorithms for implementing the constrained maximization mentioned in Remark 3 have yet to be developed. For case-control studies with unknown population totals, identifiability is a thorny issue. We have provided a simple and efficient method under the rare disease assumption, which appears to work well even when the disease is not rare. But can we do better?

Model Selection and Model Assessment. Because our approach is built on likelihood, we can apply likelihood-based model selection criteria, such as the AIC used in Section 4. Lin (2004) showed that the AIC performs well for the proportional hazards model. It is unclear how to apply the traditional residual-based methods for assessing model adequacy, because the haplotypes are not directly observable.

Other Genetic Variants. We have focused on SNPs-based haplotypes. The proposed inference procedures are potentially applicable to microsatellite loci and other genotype data, although the identifiability of parameters needs to be verified for each kind of genotype data.

Other Study Designs. It is sometimes desirable to use the matched case-control design in which one or more controls are individually matched to each case. In large cohort studies with rare diseases, it is cost-effective to adopt the case-cohort design or nested case-control design, so that only a subset of the cohort members needs to be genotyped. We are currently developing efficient inference procedures for such designs.

Population Substructure. The presence of latent population substructure can cause bias in association studies of unrelated individuals. There exist several statistical methods to adjust for the effects of population substructure with the aid of genomic markers. It should be possible to extend the proposed methods so as to accommodate potential population substructure.

Table 6. Estimates of Haplotype Effects Under Various Genetic Models for the FUSION Study

Haplotype	Recessive model	Dominant model	Additive model	Codominant model	
				Additive	Recessive
01011	$.327_{(.270)}$	$-.027_{(.140)}$	$.049_{(.135)}$	$.005_{(.143)}$	$.331_{(.289)}$
01100	$.316_{(.146)}$	$.274_{(.109)}$	$.355_{(.099)}$	$.334_{(.114)}$	$.063_{(.167)}$
10011	$-.206_{(.155)}$	$-.323_{(.112)}$	$-.320_{(.095)}$	$-.344_{(.111)}$	$.076_{(.183)}$
10100	$-1.019_{(1.020)}$	$.196_{(.219)}$	$.116_{(.213)}$	$.169_{(.217)}$	$-1.131_{(1.029)}$
10110	$.903_{(.746)}$	$-.007_{(.248)}$	$.063_{(.249)}$	$.016_{(.254)}$	$.892_{(.765)}$
11011	$-.222_{(.328)}$	$-.096_{(.140)}$	$-.127_{(.133)}$	$-.108_{(.140)}$	$-.143_{(.344)}$

NOTE: Standard error estimates are shown in parentheses.

Studies of Related Individuals. This article is concerned with studies of unrelated individuals. Many genetic studies involve multiple family members or relatives. Haplotype ambiguity possibly can be reduced by using the genotype information from related individuals. Inference on haplotype effects needs to account for the intraclass correlation.

Genotyping Error and DNA Pooling. Laboratory genotyping is prone to error. It is sometimes necessary to pool DNA samples rather than genotyping individual samples (Wang, Kidd, and Zhao 2003). Such data create additional complexity in haplotype analysis (Zeng and Lin 2005).

Many SNPs. The traditional EM algorithm works well for a small number of SNPs. When the number of SNPs is large, the partition–ligation method of Niu et al. (2002) and Qin et al. (2002) and other modifications potentially can be adapted to reduce the computational burden. However, the haplotype analysis may not be very useful if the SNPs are weakly linked.

Many Haplotypes and Rare Haplotypes. The approach taken in this article assumes that we are interested in a small number of haplotype configurations that are relatively frequent. If there are many haplotypes, then we are confronted with the problem of multiple comparisons and sparse data. Schaid (2004) discussed some possible solutions.

Large-Scale Studies. There is an increasing interest in genome-wide association studies. With a large number of SNPs, one possible approach is to use sliding windows of 5–10 SNPs and test for the haplotype–disease association in each window. Because most of the SNPs are common between adjacent windows, the test statistics tend to be highly correlated, so that the Bonferroni-type correction for multiple comparisons would be extremely conservative. To properly adjust for multiple comparisons, one needs to ascertain the joint distribution of the test statistics. This can be done by permuting the data or by evaluating the asymptotic joint normal distribution of the test statistics (Lin 2005).

We hope that other statisticians will join us in tackling the foregoing problems and other challenges in genetic association studies.

APPENDIX: TECHNICAL AND COMPUTATIONAL DETAILS

A.1 Proof of Lemma 1

We provide a proof under (3); the proof under (4) is simpler and is omitted here. To prove the first part of the lemma, we suppose that two sets of parameters, $(\{\pi_k\}, \rho)$ and $(\{\widetilde{\pi}_k\}, \widetilde{\rho})$, yield the same distribution of G. We wish to show that these two sets are identical. Consider $g = 2h_k$. For such a choice of g, the set $\mathcal{S}(g)$ is a singleton. Clearly, $(1-\rho)\pi_k^2 + \rho\pi_k = (1-\widetilde{\rho})\widetilde{\pi}_k^2 + \widetilde{\rho}\widetilde{\pi}_k$. We denote this constant by c_k. Then $0 \le c_k \le 1$ for all k, and $0 < c_k < 1$ for at least one k. Because $\pi_k \ge 0$, we have $\pi_k = [-\rho + \{\rho^2 + 4c_k(1-\rho)\}^{1/2}]/2(1-\rho)$. Thus $(1-\rho)^{-1}$ satisfies the equation $\sum_k[(1-x) + \{(x-1)^2 + 4c_k x\}^{1/2}] = 2$, and $(1-\widetilde{\rho})^{-1}$ satisfies the same equation. It can be shown that the first derivative of $(1-x) + \{(x-1)^2 + 4c_k x\}^{1/2}$ is nonpositive and is strictly negative for at least one k. Thus the foregoing equation has a unique solution for $x > 1$, which implies that $\rho = \widetilde{\rho}$. It follows immediately that $\pi_k = \widetilde{\pi}_k$ for all k. To prove the second part of the lemma, we choose $g = 2h_k$ to obtain $\nu_k\{2\pi_k(1-\rho) + \rho\} + \mu\pi_k(1-\pi_k) = 0$. Because $\sum_k \nu_k = 0$, we have $\sum_k\{\mu\pi_k(1-\pi_k)\}/\{2\pi_k(1-\rho) + \rho\} = 0$. Therefore, $\mu = 0$ and $\boldsymbol{\nu} = \mathbf{0}$.

A.2 Cross-Sectional Studies

A.2.1 Identifiability Under Arbitrary Distributions of H. Under Condition 1, $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi})$ is identifiable. The identifiability of the distribution of H depends on the structure of $P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}$. For concreteness, we consider the codominant logistic regression model for a binary trait. We divide $\mathcal{G}$ into three categories: $\mathcal{G}_1 = \{g \in \mathcal{G} : g = h + h \text{ or } g = h + \widetilde{h}\}$, $\mathcal{G}_2 = \{g \in \mathcal{G} - \mathcal{G}_1 : g \text{ is not } \ge h^*\}$, and $\mathcal{G}_3 = \mathcal{G} - \mathcal{G}_1 - \mathcal{G}_2$. We derive the expression for $m_g(y, \mathbf{x}; \boldsymbol{\theta})$ when g belongs to each of the three categories.

For $g \in \mathcal{G}_1$, $\mathcal{S}(g) = \{(h,h)\}$ or $\{(h,\widetilde{h})\}$, so that $m_g(y, \mathbf{x}; \boldsymbol{\theta}) = P_{\alpha,\boldsymbol{\beta},\xi}(Y = y|\mathbf{X} = \mathbf{x}, H = (h,h))P(H = (h,h))$ or $m_g(y, \mathbf{x}; \boldsymbol{\theta}) = P_{\alpha,\boldsymbol{\beta},\xi}(Y = y|\mathbf{X} = \mathbf{x}, H = (h,\widetilde{h}))P(H = (h,\widetilde{h}))$. For $g \in \mathcal{G}_2$, $P_{\alpha,\boldsymbol{\beta},\xi}(Y = y|\mathbf{X} = \mathbf{x}, H = (h_k, h_l))$ does not depend on $(h_k, h_l) \in \mathcal{S}(g)$, so that $m_g(y, \mathbf{x}; \boldsymbol{\theta}) = P_{\alpha,\boldsymbol{\beta},\xi}(Y = y|\mathbf{X} = \mathbf{x}, H = (h_k, h_l))P(G = g)$, where $(h_k, h_l) \in \mathcal{S}(g)$. For $g \in \mathcal{G}_3$,

$$m_g(y, \mathbf{x}; \boldsymbol{\theta}) = \frac{\exp\{y(\alpha + \beta_1 + \boldsymbol{\beta}_3^T\mathbf{x} + \boldsymbol{\beta}_4^T\mathbf{x})\}}{1 + \exp(\alpha + \beta_1 + \boldsymbol{\beta}_3^T\mathbf{x} + \boldsymbol{\beta}_4^T\mathbf{x})}\pi_1(g) + \frac{\exp\{y(\alpha + \boldsymbol{\beta}_3^T\mathbf{x})\}}{1 + \exp(\alpha + \boldsymbol{\beta}_3^T\mathbf{x})}\pi_2(g),$$

where $\pi_1(g) = 2P(H = (h^*, g - h^*))$ and $\pi_2(g) = P(H = (h_k, h_l) : h_k + h_l = g, h_k \ne h^*, h_l \ne h^*)$.

Let $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$, $P_0(G = g)$ denote the true value of $P(G = g)$, and $\pi_{0j}(g)$ denote the true values $\pi_j(g)$, $j = 1, 2$. We then can draw the following conclusions: (1) When $\beta_{01} = 0$ and $\boldsymbol{\beta}_{04} = \mathbf{0}$, $m_g(y, \mathbf{x}; \boldsymbol{\theta}) = m_g(y, \mathbf{x}; \boldsymbol{\theta}_0)$ if and only if $\alpha = \alpha_0$, $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, and $P(G = g) = P_0(G = g)$ for any $g \in \mathcal{G}$; and (2) when either β_{01} or $\boldsymbol{\beta}_{04}$ is nonzero, $m_g(y, \mathbf{x}; \boldsymbol{\theta}) = m_g(y, \mathbf{x}; \boldsymbol{\theta}_0)$ if and only if $\alpha = \alpha_0$, $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, $P(G = g) = P_0(G = g)$ for $g \in \mathcal{G}_1 \cup \mathcal{G}_2$, and $\pi_j(g) = \pi_{0j}(g)$ for $g \in \mathcal{G}_3$ and $j = 1, 2$. These conclusions hold for any generalized linear model with the linear predictor given in (1).

A.2.2 EM Algorithm. The complete-data likelihood is proportional to $\prod_{i=1}^n\{P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(\mathbf{Y}_i|\mathbf{X}_i, H_i)P_{\boldsymbol{\gamma}}(H_i)\}$. The expectation of the logarithm of this function conditional on the observable data $(\mathbf{Y}_i, \mathbf{X}_i, G_i)$, $i = 1, \ldots, n$, is

$$\sum_{i=1}^n \sum_{(h_k,h_l)\in\mathcal{S}(G_i)} p_{ikl}(\boldsymbol{\theta})\{\log P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(\mathbf{Y}_i|\mathbf{X}_i, (h_k,h_l)) + \log P_{\boldsymbol{\gamma}}(h_k,h_l)\},$$

where

$$p_{ikl}(\boldsymbol{\theta}) = \frac{P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(\mathbf{Y}_i|\mathbf{X}_i, (h_k,h_l))P_{\boldsymbol{\gamma}}(h_k,h_l)}{\sum_{(h_k,h_l)\in\mathcal{S}(G_i)} P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(\mathbf{Y}_i|\mathbf{X}_i, (h_k,h_l))P_{\boldsymbol{\gamma}}(h_k,h_l)}.$$

Thus, in the $(m+1)$st iteration of the EM algorithm, we evaluate $p_{ikl}(\boldsymbol{\theta})$ at the current estimate $\widehat{\boldsymbol{\theta}}^{(m)}$, and obtain $\widehat{\boldsymbol{\theta}}^{(m+1)}$ by solving the following equations through the Newton–Raphson algorithm:

$$\sum_{i=1}^n \sum_{(h_k,h_l)\in\mathcal{S}(G_i)} p_{ikl}\big(\widehat{\boldsymbol{\theta}}^{(m)}\big) \times \nabla_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}} \log P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}\big(\mathbf{Y}_i|\mathbf{X}_i, (h_k,h_l)\big) = \mathbf{0},$$

$$\sum_{i=1}^n \sum_{(h_k,h_l)\in\mathcal{S}(G_i)} p_{ikl}\big(\widehat{\boldsymbol{\theta}}^{(m)}\big)\nabla_{\boldsymbol{\gamma}} \log P_{\boldsymbol{\gamma}}(h_k,h_l) = \mathbf{0}. \tag{A.1}$$

Under (3) with $\rho \ge 0$, the estimate of $\boldsymbol{\gamma} \equiv (\rho, \{\pi_k\})$ can be obtained in a closed form rather than by solving (A.1). Let B be a Bernoulli variable with success probability ρ, let Q_1 be a discrete random variable taking values in H with $P(Q_1 = (h_k, h_l)) = \delta_{kl}\pi_k$, and let Q_2 be another discrete random variable taking values in H

with $P(Q_2 = (h_k, h_l)) = \pi_k\pi_l$. Then H has the same distribution as $BQ_1 + (1-B)Q_2$. The complete-data likelihood can be represented by

$$\prod_{i=1}^{n}\Big\{P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(\mathbf{Y}_i|\mathbf{X}_i, H_i)\prod_k \pi_k^{I(Q_{1i}=(h_k,h_k))B_i} \times \prod_{k,l}(\pi_k\pi_l)^{I(Q_{2i}=(h_k,h_l))(1-B_i)}\rho^{B_i}(1-\rho)^{1-B_i}\Big\}.$$

The corresponding score equations for $\{\pi_k\}$ and ρ satisfy

$$\pi_k = c^{-1}\left[\sum_{i=1}^{n} B_i I\big(Q_{1i} = (h_k, h_k)\big) + \sum_{i=1}^{n}\sum_{l=1}^{K}(1-B_i)\big\{I\big(Q_{2i} = (h_k, h_l)\big) + I\big(Q_{2i} = (h_l, h_k)\big)\big\}\right]$$

and $\rho = n^{-1}\sum_{i=1}^{n} B_i$, where c is a normalizing constant such that $\sum_k \pi_k = 1$. Define

$$\begin{aligned}
&E\{\omega(B_i, Q_{1i}, Q_{2i})|\mathbf{Y}_i, \mathbf{X}_i, G_i\}\\
&\quad= \sum_{bq_1+(1-b)q_2\in\mathcal{S}(G_i)} P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}\big(\mathbf{Y}_i|\mathbf{X}_i, bq_1 + (1-b)q_2\big) \times p(b, q_1, q_2)\omega(b, q_1, q_2)\\
&\qquad\times\left[\sum_{bq_1+(1-b)q_2\in\mathcal{S}(G_i)} P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}\big(\mathbf{Y}_i|\mathbf{X}_i, bq_1 + (1-b)q_2\big) \times p(b, q_1, q_2)\right]^{-1},
\end{aligned}$$

where $\omega(B, Q_1, Q_2) = BI(Q_1 = (h_k, h_k))$, $(1-B)I(Q_2 = (h_k, h_l))$ or B, and

$$p(b, q_1, q_2) = \prod_k \pi_k^{bI(q_1=(h_k,h_k))} \times \prod_{k,l}(\pi_k\pi_l)^{(1-b)I(q_2=(h_k,h_l))}\rho^b(1-\rho)^{1-b}.$$

In the $(m+1)$st iteration, the estimates of π_k and ρ are obtained in closed form,

$$\pi_k^{(m+1)} = \frac{1}{c^{(m+1)}}\left[\sum_{i=1}^{n} E^{(m)}\big\{B_i I\big(Q_{1i} = (h_k, h_k)\big)\big\} + 2\sum_{i=1}^{n}\sum_{l=1}^{K} E^{(m)}\big\{(1-B_i)I\big(Q_{2i} = (h_k, h_l)\big)\big\}\right],$$

and $\rho^{(m+1)} = n^{-1}\sum_{i=1}^{n} E^{(m)}(B_i)$, where $E^{(m)}\{\omega(B_i, Q_{1i}, Q_{2i})\}$ is $E\{\omega(B_i, Q_{1i}, Q_{2i})|\mathbf{Y}_i, \mathbf{X}_i, G_i\}$ evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(m)}$ and $c^{(m+1)}$ is the constant such that $\sum_k \pi_k^{(m+1)} = 1$.

A.3 Case-Control Studies With Known Population Totals

A.3.1 EM Algorithm. This is similar to the EM algorithm for cross-sectional studies, except that in addition to unknown H on all individuals, $\mathbf{X}$ is missing for the individuals not selected into the case-control sample and there are nonparametric components $\{F_g(\cdot)\}$. The complete-data likelihood is

$$\prod_{i=1}^{N} P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(Y_i|\mathbf{X}_i, H_i)P_{\boldsymbol{\gamma}}(H_i)\prod_g\{f_g(\mathbf{X}_i)\}^{I(G_i=g)}.$$

The M-step solves the following equations for $\boldsymbol{\theta}$:

$$\begin{aligned}
&\sum_{i=1}^{N} I(R_i = 1)E\big\{\nabla_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}\log P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(Y_i|\mathbf{X}_i, H_i)|Y_i, \mathbf{X}_i, G_i\big\}\\
&\qquad+\sum_{i=1}^{N} I(R_i = 0)E\big\{\nabla_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}\log P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(Y_i|\mathbf{X}_i, H_i)|Y_i\big\} = \mathbf{0},\\
&\sum_{i=1}^{N} I(R_i = 1)E\{\nabla_{\boldsymbol{\gamma}}\log P_{\boldsymbol{\gamma}}(H_i)|Y_i, \mathbf{X}_i, G_i\}\\
&\qquad+\sum_{i=1}^{N} I(R_i = 0)E\{\nabla_{\boldsymbol{\gamma}}\log P_{\boldsymbol{\gamma}}(H_i)|Y_i\} = \mathbf{0}, \quad \text{(A.2)}
\end{aligned}$$

and also estimates F_g by an empirical function with the following point mass at the $\mathbf{X}_i$ for which $(G_i = g, R_i = 1)$:

$$\begin{aligned}
&F_g\{\mathbf{X}_i\}\\
&\quad=\left[\sum_{j=1}^{N} I(\mathbf{X}_j = \mathbf{X}_i, G_j = g, R_j = 1) + \sum_{j=1}^{N} I(R_j = 0)E\{I(\mathbf{X}_j = \mathbf{X}_i, G_j = g)|Y_j\}\right]\\
&\qquad\times\left[\sum_{j=1}^{N} I(G_j = g, R_j = 1) + \sum_{j=1}^{N} I(R_j = 0)E\{I(G_j = g)|Y_j\}\right]^{-1},
\end{aligned}$$

where the conditional expectations are evaluated at the current estimates of $\boldsymbol{\theta}$ and $\{F_g\}$ in the E-step. For a random function $\omega(Y_i, \mathbf{X}_i, H_i)$, the conditional expectation takes the form

$$\frac{\sum_{(h_k,h_l)\in\mathcal{S}(G_i)} w(Y_i, \mathbf{X}_i, (h_k, h_l))P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(Y_i|\mathbf{X}_i, (h_k, h_l))P_{\boldsymbol{\gamma}}(h_k, h_l)}{\sum_{(h_k,h_l)\in\mathcal{S}(G_i)} P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(Y_i|\mathbf{X}_i, (h_k, h_l))P_{\boldsymbol{\gamma}}(h_k, h_l)}$$

for $R_i = 1$ and

$$\begin{aligned}
&\sum_{g\in\mathcal{G}}\sum_{\mathbf{x}\in\{\mathbf{X}_i:\, G_i=g, R_i=1\}}\sum_{(h_k,h_l)\in\mathcal{S}(g)} \omega(Y_i, \mathbf{x}, (h_k, h_l)) \times P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(Y_i|\mathbf{x}, (h_k, h_l)) \times P_{\boldsymbol{\gamma}}(h_k, h_l)F_g\{\mathbf{x}\}\\
&\quad\times\left(\sum_{g\in\mathcal{G}}\sum_{\mathbf{x}\in\{\mathbf{X}_i:\, G_i=g, R_i=1\}}\sum_{(h_k,h_l)\in\mathcal{S}(g)} P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(Y_i|\mathbf{x}, (h_k, h_l)) \times P_{\boldsymbol{\gamma}}(h_k, h_l)F_g\{\mathbf{x}\}\right)^{-1}
\end{aligned}$$

for $R_i = 0$. Under (3) with $\rho \geq 0$, the idea described in Section A.2.2 can be applied to (A.2) to obtain a closed-form estimate of $\boldsymbol{\gamma}$.

A.3.2 Proof of Theorem 1. The case-control design with known population totals is a special case of the two-phase designs studied by Breslow, McNeney, and Wellner (2003). The likelihood given in (6) resembles (2.3) of Breslow et al. The key difference is that the former involves several nonparametric components $\{F_g(\cdot)\}$, whereas the latter involves only a single nonparametric function. Despite this difference, the arguments of Breslow et al. can be used to prove Theorem 1 with minor modifications. Specifically, the regularity conditions of Breslow et al. hold under our Conditions 1–4. Thus, the consistency of $(\widehat{\boldsymbol{\theta}}, \{\widehat{F}_g(\cdot)\})$ follows from the results of van der Vaart and Wellner (2001), whereas the weak convergence and asymptotic efficiency can be established by applying the results of Murphy and van der Vaart (2000) through a least favorable submodel, which can be constructed as done by Breslow et al. (2003, sec. 3).

A.4 Case-Control Studies With Unknown Population Totals

A.4.1 Equivalence Class. Suppose that two sets of parameters, $(\boldsymbol{\theta}, \{F_g^\dagger\}, \{q_g\})$ and $(\widetilde{\boldsymbol{\theta}}, \{\widetilde{F}_g^\dagger\}, \{\widetilde{q}_g\})$, yield the same likelihood,

$$RL(\boldsymbol{\theta}, \{F_g^\dagger\}, \{q_g\}) = RL(\widetilde{\boldsymbol{\theta}}, \{\widetilde{F}_g^\dagger\}, \{\widetilde{q}_g\}). \tag{A.3}$$

Because $\eta(0, \mathbf{x}, g; \boldsymbol{\theta}) = 1$, (A.3) with $y = 0$ implies that $f_g^\dagger(\mathbf{x}) q_g / \sum_{\widetilde{g}\in\mathcal{G}} q_{\widetilde{g}} = \widetilde{f}_g^\dagger(\mathbf{x}) \widetilde{q}_g / \sum_{\widetilde{g}\in\mathcal{G}} \widetilde{q}_{\widetilde{g}}$. Thus $f_g^\dagger(\mathbf{x}) = \widetilde{f}_g^\dagger(\mathbf{x})$ and $q_g = \widetilde{q}_g$. It then follows from (A.3) that

$$\eta(y, \mathbf{x}, g; \boldsymbol{\theta}) = C(y)\eta(y, \mathbf{x}, g; \widetilde{\boldsymbol{\theta}}), \tag{A.4}$$

where $C(y)$ depends only on y. By setting $\mathbf{x} = \mathbf{x}_0$ and $g = g_0$ in (A.4) and noting that $\eta(y, \mathbf{x}_0, g_0; \boldsymbol{\theta}) = 1$, we conclude that $C(y) = 1$. Hence the equivalence class for $(\boldsymbol{\theta}, \{F_g^\dagger\}, \{q_g\})$ is $\{(\widetilde{\boldsymbol{\theta}}, \{F_g^\dagger\}, \{q_g\}) : \eta(y, \mathbf{x}, g; \widetilde{\boldsymbol{\theta}}) = \eta(y, \mathbf{x}, g; \boldsymbol{\theta})\}$.

A.4.2 Identifiability for the Logistic Link Function. Suppose that

$$\eta(y, \mathbf{x}, g; \widetilde{\boldsymbol{\theta}}) = \eta(y, \mathbf{x}, g; \boldsymbol{\theta}) \tag{A.5}$$

for two sets of parameters $\widetilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$. Let $g_0 = 0$. As in Section A.2.1, we partition $\mathcal{G}$ into $(\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3)$. For $g \in \mathcal{G}_1$, $\mathcal{S}(g)$ is a singleton, so the generalized odds ratio reduces to the ordinary odds ratio of Y given $\mathbf{X}$ and H. Thus (A.5) is equivalent to $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}$ under Condition 8. For $g \in \mathcal{G}_2$, $P(Y = 0|\mathbf{X} = \mathbf{x}, H = (h_k, h_l)) = \{1 + \exp(\alpha + \boldsymbol{\beta}_3^T \mathbf{x})\}^{-1}$. Thus (A.5) holds if and only if $\widetilde{\boldsymbol{\beta}}_3 = \boldsymbol{\beta}_3$. For $g \in \mathcal{G}_3$, both $\pi_1(g)$ and $\pi_2(g)$ are nonzero. Then (A.5) becomes

$$\frac{\widetilde{\pi}_1(g)(1 + e^{\widetilde{\alpha}+\psi_2(\mathbf{x})})/\widetilde{\pi}_2(g)(1 + e^{\widetilde{\alpha}+\psi_1(\mathbf{x})}) + e^{\psi_2(\mathbf{x})-\psi_1(\mathbf{x})}}{\widetilde{\pi}_1(g)(1 + e^{\widetilde{\alpha}+\psi_2(\mathbf{x})})/\widetilde{\pi}_2(g)(1 + e^{\widetilde{\alpha}+\psi_1(\mathbf{x})}) + 1}$$
$$= \frac{\pi_1(g)(1 + e^{\alpha+\psi_2(\mathbf{x})})/\pi_2(g)(1 + e^{\alpha+\psi_1(\mathbf{x})}) + e^{\psi_2(\mathbf{x})-\psi_1(\mathbf{x})}}{\pi_1(g)(1 + e^{\alpha+\psi_2(\mathbf{x})})/\pi_2(g)(1 + e^{\alpha+\psi_1(\mathbf{x})}) + 1}, \tag{A.6}$$

where $\psi_1(\mathbf{x}) = \beta_1 + \boldsymbol{\beta}_3^T \mathbf{x} + \boldsymbol{\beta}_4^T \mathbf{x}$ and $\psi_2(\mathbf{x}) = \boldsymbol{\beta}_3^T \mathbf{x}$.

Without loss of generality, assume that $\mathbf{0}$ is in the support of $\mathbf{X}$. We then have the following results:

1. $\beta_1 = 0$ and $\boldsymbol{\beta}_4 = \mathbf{0}$. Then (A.6) holds naturally.
2. $\beta_1 \neq 0, \boldsymbol{\beta}_4 = \mathbf{0}$, and $\boldsymbol{\beta}_3 = \mathbf{0}$. Then, because the function $(\lambda + c)/(\lambda + 1)$ is strictly monotone in λ for $c \neq 1$, (A.6) yields

$$\frac{\widetilde{\pi}_1(g)}{\widetilde{\pi}_2(g)} \frac{1 + e^{\widetilde{\alpha}}}{1 + e^{\widetilde{\alpha}+\beta_1}} = \frac{\pi_1(g)}{\pi_2(g)} \frac{1 + e^{\alpha}}{1 + e^{\alpha+\beta_1}}.$$

   Thus (A.6) is equivalent to

$$\frac{\widetilde{\pi}_1(g)/\widetilde{\pi}_2(g)}{\widetilde{\pi}_1(\widetilde{g})/\widetilde{\pi}_2(\widetilde{g})} = \frac{\pi_1(g)/\pi_2(g)}{\pi_1(\widetilde{g})/\pi_2(\widetilde{g})} \quad \text{for all } g, \widetilde{g} \in \mathcal{G}_3.$$

3. $\beta_1 \neq 0, \boldsymbol{\beta}_4 = \mathbf{0}$, and $\beta_{3,z} \neq 0$, where $\beta_{3,z}$ is the component of $\boldsymbol{\beta}_3$ associated with a continuous covariate Z. For $\mathbf{x}$ such that $\beta_{3,z} z \neq 0$, (A.6) yields

$$\frac{\widetilde{\pi}_1(g)}{\widetilde{\pi}_2(g)} \frac{1 + e^{\widetilde{\alpha}+\beta_{3,z}z}}{1 + e^{\widetilde{\alpha}+\beta_1+\beta_{3,z}z}} = \frac{\pi_1(g)}{\pi_2(g)} \frac{1 + e^{\alpha+\beta_{3,z}z}}{1 + e^{\alpha+\beta_1+\beta_{3,z}z}}.$$

   The foregoing equation holds for any $z \in (-\infty, \infty)$ because the functions on the two sides are analytic in z and z is continuous. Without loss of generality, assume that $\beta_{3,z} > 0$. By letting $z = -\infty$, we have $\widetilde{\pi}_1(g)/\widetilde{\pi}_2(g) = \pi_1(g)/\pi_2(g)$. Then by letting $z = 0$, we have $\widetilde{\alpha} = \alpha$. Thus (A.6) is equivalent to $\{\widetilde{\alpha} = \alpha, \widetilde{\pi}_1(g)/\widetilde{\pi}_2(g) = \pi_1(g)/\pi_2(g)\}$.
4. $\beta_{4,z} \neq 0$, where $\beta_{4,z}$ is the component of $\boldsymbol{\beta}_4$ pertaining to z. Then (A.6) is equivalent to

$$\frac{\widetilde{\pi}_1(g)}{\widetilde{\pi}_2(g)} \frac{1 + e^{\widetilde{\alpha}+\psi_2(\mathbf{x})}}{1 + e^{\widetilde{\alpha}+\psi_1(\mathbf{x})}} = \frac{\pi_1(g)}{\pi_2(g)} \frac{1 + e^{\alpha+\psi_2(\mathbf{x})}}{1 + e^{\alpha+\psi_1(\mathbf{x})}} \tag{A.7}$$

   for any $\mathbf{x}$ such that $\beta_1 + \boldsymbol{\beta}_4^T \mathbf{x} \neq 0$. We set $\mathbf{x}$ except the component z to $\mathbf{0}$. By letting $z \to -\beta_1/\beta_{4,z}$, we have $\widetilde{\pi}_1(g)/\widetilde{\pi}_2(g) = \pi_1(g)/\pi_2(g)$. Then by differentiating both sides of (A.7) with respect to z and letting $z \to -\beta_1/\beta_{4,z}$, we obtain $\alpha = \widetilde{\alpha}$. Thus (A.6) is equivalent to $\{\widetilde{\alpha} = \alpha, \widetilde{\pi}_1(g)/\widetilde{\pi}_2(g) = \pi_1(g)/\pi_2(g)\}$.

A.4.3 Identifiability for Probit and Complementary Log–Log Link Functions. Assume that $|\beta_1| + |\boldsymbol{\beta}_4| \neq 0$. Also assume that there exists a continuous covariate in $\mathbf{X}$, denoted by Z, such that the corresponding regression parameter β_z is nonzero. Let $\mathbf{x}_0 = \mathbf{0}$ and $g_0 = 0$. We claim that under the probit and complementary log–log regression models, $\eta(1, \mathbf{x}, g; \boldsymbol{\theta}) = \eta(1, \mathbf{x}, g; \widetilde{\boldsymbol{\theta}})$ for two sets of parameters $\boldsymbol{\theta}$ and $\widetilde{\boldsymbol{\theta}}$ if and only if $\alpha = \widetilde{\alpha}$, $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}$, and $\pi_1(g)/\pi_2(g) = \widetilde{\pi}_1(g)/\widetilde{\pi}_2(g)$ for $g \in \mathcal{G}_3$.

We first prove the foregoing claim for the probit model. Suppose that $\eta(1, \mathbf{x}, g; \boldsymbol{\theta}) = \eta(1, \mathbf{x}, g; \widetilde{\boldsymbol{\theta}})$. Without loss of generality, assume that h^* is a nonzero sequence. Let $g = 2h^*, h^* + \widetilde{h}^*$, and 0 in turn. Because $\mathcal{S}(g)$ has a single element for such g, we obtain

$$\frac{\Phi(\alpha)}{1 - \Phi(\alpha)} \left\{ \frac{1}{\Phi(\alpha + 2\beta_1 + \beta_2 + \boldsymbol{\beta}_3^T \mathbf{x} + 2\boldsymbol{\beta}_4^T \mathbf{x} + \boldsymbol{\beta}_5^T \mathbf{x})} - 1 \right\}$$
$$= \frac{\Phi(\widetilde{\alpha})}{1 - \Phi(\widetilde{\alpha})} \left\{ \frac{1}{\Phi(\widetilde{\alpha} + 2\widetilde{\beta}_1 + \widetilde{\beta}_2 + \widetilde{\boldsymbol{\beta}}_3^T \mathbf{x} + 2\widetilde{\boldsymbol{\beta}}_4^T \mathbf{x} + \widetilde{\boldsymbol{\beta}}_5^T \mathbf{x})} - 1 \right\}, \tag{A.8}$$

$$\frac{\Phi(\alpha)}{1 - \Phi(\alpha)} \left\{ \frac{1}{\Phi(\alpha + \beta_1 + \boldsymbol{\beta}_3^T \mathbf{x} + \boldsymbol{\beta}_4^T \mathbf{x})} - 1 \right\}$$
$$= \frac{\Phi(\widetilde{\alpha})}{1 - \Phi(\widetilde{\alpha})} \left\{ \frac{1}{\Phi(\widetilde{\alpha} + \widetilde{\beta}_1 + \widetilde{\boldsymbol{\beta}}_3^T \mathbf{x} + \widetilde{\boldsymbol{\beta}}_4^T \mathbf{x})} - 1 \right\}, \tag{A.9}$$

and

$$\frac{\Phi(\alpha)}{1 - \Phi(\alpha)} \left\{ \frac{1}{\Phi(\alpha + \boldsymbol{\beta}_3^T \mathbf{x})} - 1 \right\}$$
$$= \frac{\Phi(\widetilde{\alpha})}{1 - \Phi(\widetilde{\alpha})} \left\{ \frac{1}{\Phi(\widetilde{\alpha} + \widetilde{\boldsymbol{\beta}}_3^T \mathbf{x})} - 1 \right\}, \tag{A.10}$$

where Φ is the standard normal distribution function. In (A.10), we let $\mathbf{x}$ except the component z be $\mathbf{0}$. Then

$$\frac{\Phi(\alpha)}{1 - \Phi(\alpha)} \left\{ \frac{1}{\Phi(\alpha + \beta_z z)} - 1 \right\} = \frac{\Phi(\widetilde{\alpha})}{1 - \Phi(\widetilde{\alpha})} \left\{ \frac{1}{\Phi(\widetilde{\alpha} + \widetilde{\beta}_z z)} - 1 \right\}.$$

By letting $z \to \infty$ or $-\infty$, we conclude that β_z and $\widetilde{\beta}_z$ must have the same sign. Without loss of generality, assume that $\beta_z > \widetilde{\beta}_z > 0$. Then the left side divided by the right side goes to 0 as $z \to \infty$. This is a contradiction. Therefore, $\beta_z = \widetilde{\beta}_z$. We differentiate both sides to obtain

$$\frac{\Phi(\alpha)}{1 - \Phi(\alpha)} \frac{\phi(\alpha + \beta_z z)}{\Phi(\alpha + \beta_z z)^2} = \frac{\Phi(\widetilde{\alpha})}{1 - \Phi(\widetilde{\alpha})} \frac{\phi(\widetilde{\alpha} + \beta_z z)}{\Phi(\widetilde{\alpha} + \beta_z z)^2}.$$

By taking the ratio of the two sides and letting $z \to \operatorname{sgn}(\beta_z)\infty$, we immediately conclude that $\alpha = \widetilde{\alpha}$. Applying this result to (A.8)–(A.10), we obtain $2\beta_1 + \beta_2 + \boldsymbol{\beta}_3^T \mathbf{x} + 2\boldsymbol{\beta}_4^T \mathbf{x} + \boldsymbol{\beta}_5^T \mathbf{x} = 2\widetilde{\beta}_1 + \widetilde{\beta}_2 + \widetilde{\boldsymbol{\beta}}_3^T \mathbf{x} + 2\widetilde{\boldsymbol{\beta}}_4^T \mathbf{x} + \widetilde{\boldsymbol{\beta}}_5^T \mathbf{x}$, $\beta_1 + \boldsymbol{\beta}_3^T \mathbf{x} + \boldsymbol{\beta}_4^T \mathbf{x} = \widetilde{\beta}_1 + \widetilde{\boldsymbol{\beta}}_3^T \mathbf{x} + \widetilde{\boldsymbol{\beta}}_4^T \mathbf{x}$, and $\boldsymbol{\beta}_3^T \mathbf{x} = \widetilde{\boldsymbol{\beta}}_3^T \mathbf{x}$. Therefore, $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}$. For $g \in \mathcal{G}_3$,

$$\begin{aligned}
&\eta(1, \mathbf{x}, g; \boldsymbol{\theta}) \\
&\quad = \frac{1 - \Phi(\alpha)}{\Phi(\alpha)} \{\Phi(\alpha + \beta_1 + \boldsymbol{\beta}_3^T \mathbf{x} + \boldsymbol{\beta}_4^T \mathbf{x})\pi_1(g)/\pi_2(g) \\
&\qquad + \Phi(\alpha + \boldsymbol{\beta}_3^T \mathbf{x})\} \\
&\qquad \times \left[\{1 - \Phi(\alpha + \beta_1 + \boldsymbol{\beta}_3^T \mathbf{x} + \boldsymbol{\beta}_4^T \mathbf{x})\}\pi_1(g)/\pi_2(g) \right. \\
&\qquad \left. + 1 - \Phi(\alpha + \boldsymbol{\beta}_3^T \mathbf{x})\right]^{-1}.
\end{aligned} \tag{A.11}$$

It follows that $\pi_1(g)/\pi_2(g) = \widetilde{\pi}_1(g)/\widetilde{\pi}_2(g)$. The other direction of the claim is obvious in view of (A.11) and the expressions of $\eta(1, \mathbf{x}, g)$ for $g \in \mathcal{G}_1$ and $g \in \mathcal{G}_2$.

For the complementary log–log model, we obtain the same equations as (A.8)–(A.11) with $\Phi(x)$ replaced by $1 - \exp(-e^x)$. In particular, $e^{-e^\alpha}(e^{e^{\alpha+\beta_z z}} - 1)/(1 - e^{-e^\alpha}) = e^{-e^{\widetilde{\alpha}}}(e^{e^{\widetilde{\alpha}+\widetilde{\beta}_z z}} - 1)/(1 - e^{-e^{\widetilde{\alpha}}})$. Taking the first and second derivatives of the two sides with respect to z and forming the ratio of them, we obtain $\beta_z(e^{\alpha+\beta_z z} + 1) = \widetilde{\beta}_z(e^{\widetilde{\alpha}+\widetilde{\beta}_z z} + 1)$. Thus $\alpha = \widetilde{\alpha}$ and $\beta_z = \widetilde{\beta}_z$. The rest of the proof is the same as that of the probit model.

A.4.4 Profile Likelihood of $\boldsymbol{\theta}$ Based on (8). Suppose that there are J distinct observed values of $(\mathbf{X}, G)$, denoted by $(\mathbf{x}_1, g_1), \ldots, (\mathbf{x}_J, g_J)$. Let n_{1j} and n_{0j} be the number of times that $(\mathbf{x}_j, g_j)$ is observed in the cases and controls, and let δ_j be the jump size of the estimated distribution of $(\mathbf{X}, G)$ at $(\mathbf{x}_j, g_j)$. Then the log-likelihood based on (8) can be written as

$$l_n(\boldsymbol{\theta}, \{\delta_j\}) = \sum_{j=1}^{J} n_{1j} \log \eta(1, \mathbf{x}_j, g_j; \boldsymbol{\theta}) - n_1 \log\left\{\sum_{j=1}^{J} \eta(1, \mathbf{x}_j, g_j; \boldsymbol{\theta})\delta_j\right\} + \sum_{j=1}^{J} n_{+j} \log \delta_j,$$

where $n_{+j} = n_{0j} + n_{1j}$. Following Scott and Wild (1997), we introduce a Lagrange multiplier λ for the constraint $\sum_j \delta_j = 1$ and set the derivative with respect to δ_j to 0. We then obtain

$$\frac{n_{+j}}{\delta_j} - \frac{n_1 \eta(1, \mathbf{x}_j, g_j; \boldsymbol{\theta})}{\sum_{j=1}^{J} \eta(1, \mathbf{x}_j, g_j; \boldsymbol{\theta})\delta_j} + \lambda = 0.$$

Multiplying both sides by δ_j and summing over j, we see that $\lambda = n_1 - n$. Thus

$$\delta_j = \frac{n_{+j}}{n - n_1 + n_1 \eta(1, \mathbf{x}_j, g_j; \boldsymbol{\theta})/\mu}, \tag{A.12}$$

where $\mu = \sum_{j=1}^{J} \eta(1, \mathbf{x}_j, g_j; \boldsymbol{\theta})\delta_j$. Plugging (A.12) into $l_n(\boldsymbol{\theta}, \{\delta_j\})$, we see that the objective function to be maximized is, up to a constant C_n, equal to

$$l_n^*(\boldsymbol{\theta}, \mu) = \sum_{j=1}^{J} n_{1j} \log \eta(1, \mathbf{x}_j, g_j; \boldsymbol{\theta}) - \sum_{j=1}^{J} n_{+j} \log\left\{\frac{n_1}{n}\eta(1, \mathbf{x}_j, g_j; \boldsymbol{\theta}) + \left(1 - \frac{n_1}{n}\right)\mu\right\} + (n - n_1)\log \mu.$$

Thus $\max_{\{\delta_j\}} l_n(\boldsymbol{\theta}, \{\delta_j\}) \le \max_\mu l_n^*(\boldsymbol{\theta}, \mu) + C_n$. If μ maximizes $l_n^*(\boldsymbol{\theta}, \mu)$, then $\partial l_n^*(\boldsymbol{\theta}, \mu)/\partial \mu = 0$, and the δ_j given in (A.12) satisfy $\sum_{j=1}^{J} \delta_j = 1$. Thus $\max_\mu l_n^*(\boldsymbol{\theta}, \mu) + C_n \le \max_{\{\delta_j\}} l_n(\boldsymbol{\theta}, \{\delta_j\})$. Therefore, the profile log-likelihood function for $\boldsymbol{\theta}$ based on $l_n(\boldsymbol{\theta}, \{\delta_j\})$ equals the profile function based on $l_n^*(\boldsymbol{\theta}, \mu)$, up to a constant C_n. We maximize $l_n^*(\boldsymbol{\theta}, \mu)$ via Newton–Raphson to yield $\widehat{\boldsymbol{\theta}}$ and $\widehat{\mu}$, where $\widehat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$. It can be shown that up to a constant, $l_n^*(\boldsymbol{\theta}, \mu)$ is the log-likelihood based on a random sample of size n from a conditional distribution of Y given $\mathbf{X}$ and G. Hence the covariance matrix of $(\widehat{\boldsymbol{\theta}}, \widehat{\mu})$ can be estimated by the inverse information matrix of $l_n^*(\boldsymbol{\theta}, \mu)$.

A.4.5 Profile Likelihood of $\boldsymbol{\theta}$ Based on (9). Suppose that (3) holds. Write $\boldsymbol{\theta} = (\boldsymbol{\beta}, \{\pi_k\}, \rho)$. Also define

$$\zeta_1(\mathbf{x}, g; \boldsymbol{\theta}) = \sum_{(h_k, h_l) \in \mathcal{S}(g)} e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{x}, h_k, h_l)}\{\rho \pi_k \delta_{kl} + (1 - \rho)\pi_k \pi_l\},$$

$$\zeta_0(g; \boldsymbol{\theta}) = \sum_{(h_k, h_l) \in \mathcal{S}(g)} \{\rho \pi_k \delta_{kl} + (1 - \rho)\pi_k \pi_l\}.$$

By a derivation similar to that of Section A.4.4, profiling (9) over $\{F_g(\cdot)\}$ is equivalent to profiling the following function over $\{\mu_g\}$:

$$\widetilde{l}_n^*(\boldsymbol{\theta}, \{\mu_g\}) = \sum_{i=1}^{n}\{y_i \log \zeta_1(\mathbf{X}_i, G_i; \boldsymbol{\theta}) + (1 - y_i)\log \zeta_0(G_i; \boldsymbol{\theta})\} - \sum_{i=1}^{n}\sum_{g} I(G_i = g)\log\left\{\zeta_1(\mathbf{X}_i, G_i; \boldsymbol{\theta}) + n_1^{-1}\widetilde{n}_g \sum_{\widetilde{g}} \mu_{\widetilde{g}} - \mu_g\right\} + \sum_{i=1}^{n}(1 - y_i)\log\left\{\sum_g \mu_g\right\},$$

where $\widetilde{n}_g$ is the number of times $G = g$ in the sample. The covariance matrix of $\widehat{\boldsymbol{\theta}}$ can be estimated by the sandwich estimator or the profile likelihood method.

If $\mathbf{X}$ is independent of G, then we obtain the MLE $\widehat{\boldsymbol{\theta}}$ by maximizing the following function:

$$\widetilde{l}_n^*(\boldsymbol{\theta}, \mu) = \sum_{i=1}^{n} y_i \log \zeta_1(\mathbf{X}_i, G_i; \boldsymbol{\theta}) + \sum_{i=1}^{n}(1 - y_i)\log \zeta_0(G_i; \boldsymbol{\theta}) + \sum_{i=1}^{n}(1 - y_i)\log \mu - \sum_{i=1}^{n}\log\left\{(1 - r)\mu + r\sum_g \zeta_1(\mathbf{X}_i, g; \boldsymbol{\theta})\right\},$$

where $r = n_1/n$. Let $H = BQ_1 + (1 - B)Q_2$, where B is a Bernoulli variable, Q_1 takes values in $\{(h_k, h_k); k = 1, \ldots, K\}$, and Q_2 takes values in $\{(h_k, h_l); k, l = 1, \ldots, K\}$. Suppose that Y is a binary variable and that the conditional distribution of (B, Q_1, Q_2, Y) given $\mathbf{X}$ is characterized by

$$P(B, Q_1, Q_2, Y|\mathbf{X}) = \frac{\exp\{\boldsymbol{\vartheta}^T \mathcal{W}(B, Q_1, Q_2, Y, \mathbf{X})\}}{\sum_{B, Q_1, Q_2, Y} \exp\{\boldsymbol{\vartheta}^T \mathcal{W}(B, Q_1, Q_2, Y, \mathbf{X})\}},$$

where $\boldsymbol{\vartheta} = (-\log \mu + \log r/(1 - r), \boldsymbol{\beta}^T, \log \pi_1 - \log \rho/(1 - \rho), \ldots, \log \pi_K - \log \rho/(1 - \rho))^T$ and $\mathcal{W}(B, Q_1, Q_2, Y, \mathbf{X}) = (Y, Y\mathcal{Z}^T(\mathbf{X}, H), BI(Q_1 = (h_1, h_1)) + (1 - B)\sum_l\{I(Q_2 = (h_1, h_l)) + I(Q_2 = (h_l, h_1))\}, \ldots, BI(Q_1 = (h_K, h_K)) + (1 - B)\sum_l\{I(Q_2 = (h_K, h_l)) + I(Q_2 = (h_l, h_K))\}^T$. We can show that $\widetilde{l}_n^*(\boldsymbol{\theta}, \mu)$ is equivalent to the log-likelihood

$$\widetilde{l}_n^*(\boldsymbol{\vartheta}) = \sum_{i=1}^{n} \log\left[\sum_{BQ_1 + (1-B)Q_2 \in \mathcal{S}(G_i)} \frac{e^{\boldsymbol{\vartheta}^T \mathcal{W}(B, Q_1, Q_2, Y_i, \mathbf{X}_i)}}{\sum_{b, q_1, q_2, y} e^{\boldsymbol{\vartheta}^T \mathcal{W}(b, q_1, q_2, y, \mathbf{X}_i)}}\right].$$

We maximize $\widetilde{l}_n^*(\boldsymbol{\vartheta})$ through the EM algorithm, in which (B, Q_1, Q_2) is treated as missing. The estimation of the covariance matrix of $\widehat{\boldsymbol{\theta}}$ is based on the information matrix of $\widetilde{l}_n^*(\boldsymbol{\vartheta})$.

The complete-data score function is

$$\sum_{i=1}^{n}\left[\mathcal{W}(B_i, Q_{1i}, Q_{2i}, Y_i, \mathbf{X}_i) - \frac{\sum_{b, q_1, q_2, y} \mathcal{W}(b, q_1, q_2, y, \mathbf{X}_i)\exp\{\boldsymbol{\vartheta}^T \mathcal{W}(b, q_1, q_2, y, \mathbf{X}_i)\}}{\sum_{b, q_1, q_2, y} \exp\{\boldsymbol{\vartheta}^T \mathcal{W}(b, q_1, q_2, y, \mathbf{X}_i)\}}\right].$$

Thus in the E-step we calculate the conditional expectation of $\mathcal{W}(B_i, Q_{1i}, Q_{2i}, Y_i, \mathbf{X}_i)$ given $(Y_i, \mathbf{X}_i, G_i)$ and the current parameter estimates,

$$E[\mathcal{W}(B_i, Q_{1i}, Q_{2i}, Y_i, \mathbf{X}_i)|Y_i, \mathbf{X}_i, G_i] = \frac{\sum_{b, q_1, q_2} I(bq_1 + (1-b)q_2 \in \mathcal{S}(G_i))e^{\boldsymbol{\vartheta}^T \mathcal{W}(b, q_1, q_2, Y_i, \mathbf{X}_i)}\mathcal{W}(b, q_1, q_2, Y_i, \mathbf{X}_i)}{\sum_{b, q_1, q_2} I(bq_1 + (1-b)q_2 \in \mathcal{S}(G_i))e^{\boldsymbol{\vartheta}^T \mathcal{W}(b, q_1, q_2, Y_i, \mathbf{X}_i)}}.$$

In the M-step we use the one-step Newton–Raphson iteration to update the parameter estimates,

$$\begin{aligned}\boldsymbol{\vartheta}^{(k+1)} &= \boldsymbol{\vartheta}^{(k)} - \boldsymbol{\Sigma}^{-1} \times \sum_{i=1}^{n}\Bigg[E[\mathcal{W}(B, Q_1, Q_2, Y_i, \mathbf{X}_i)|Y_i, \mathbf{X}_i, G_i] \\ &\quad - \frac{\sum_{b,q_1,q_2,y} \mathcal{W}(b, q_1, q_2, y, \mathbf{X}_i)\exp\{\boldsymbol{\vartheta}^T\mathcal{W}(b, q_1, q_2, y, \mathbf{X}_i)\}}{\sum_{b,q_1,q_2,y}\exp\{\boldsymbol{\vartheta}^T\mathcal{W}(b, q_1, q_2, y, \mathbf{X}_i)\}}\Bigg],\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{\Sigma} &= -\Bigg[\sum_{i=1}^{n}\frac{\sum_{b,q_1,q_2,y}\mathcal{W}^{\otimes 2}(b, q_1, q_2, y, \mathbf{X}_i)e^{\boldsymbol{\vartheta}^T\mathcal{W}(b,q_1,q_2,y,\mathbf{X}_i)}}{\sum_{b,q_1,q_2,y}e^{\boldsymbol{\vartheta}^T\mathcal{W}(b,q_1,q_2,y,\mathbf{X}_i)}}\Bigg] \\ &\quad + \sum_{i=1}^{n}\Bigg[\frac{\{\sum_{b,q_1,q_2,y}\mathcal{W}(b, q_1, q_2, y, \mathbf{X}_i)e^{\boldsymbol{\vartheta}^T\mathcal{W}(b,q_1,q_2,y,\mathbf{X}_i)}\}^{\otimes 2}}{\{\sum_{b,q_1,q_2,y}e^{\boldsymbol{\vartheta}^T\mathcal{W}(b,q_1,q_2,y,\mathbf{X}_i)}\}^2}\Bigg]\end{aligned}$$

and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$.

A.4.6 Proof of Theorem 2. Write $F_{\mathbf{x},g}(\mathbf{x}, g) = F_g^\dagger(\mathbf{x})q_g$ and $\widehat{F}_{\mathbf{x},g}(\mathbf{x}, g) = \widehat{F}_g^\dagger(\mathbf{x})\widehat{q}_g$. Because $\widehat{\boldsymbol{\theta}}$ is bounded and $\widehat{F}_{\mathbf{x},g}$ is a probability distribution, we can choose a subsequence such that $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^*$ and $\widehat{F}_{\mathbf{x},g}(\mathbf{x}, g) \to F^*_{\mathbf{x},g}(\mathbf{x}, g) \equiv F^*_g(\mathbf{x})q^*_g$, where $q^*_g > 0$ for any g.

Because $\widehat{F}_g^\dagger$ maximizes the likelihood, there exists some Lagrange multiplier $\widehat{\lambda}_g$ such that

$$\frac{I(G_i = g)}{\widehat{F}_g^\dagger\{\mathbf{X}_i\}} - \frac{n_1\eta(1, \mathbf{X}_i, g; \widehat{\boldsymbol{\theta}})\widehat{q}_g}{\int_{\mathbf{x},\widetilde{g}}\eta(1, \mathbf{x}, \widetilde{g}; \widehat{\boldsymbol{\theta}})\,d\widehat{F}_{\mathbf{x},g}(\mathbf{x}, \widetilde{g})} - n\widehat{\lambda}_g = 0,$$

where $\widehat{F}_g^\dagger\{\mathbf{X}_i\}$ denotes the point mass of $\widehat{F}_g^\dagger$ at $\mathbf{X}_i$ and the integral is interpreted as integration over $\mathbf{x}$ and summation over g. Because $\sum_{i=1}^n \widehat{F}_g^\dagger\{\mathbf{X}_i\} = 1$, $\widehat{\lambda}_g$ satisfies the equation

$$n^{-1}\sum_{i=1}^{n}\frac{I(G_i = g)}{\widehat{\lambda}_g + n_1\eta(1, \mathbf{X}_i, g; \widehat{\boldsymbol{\theta}})\widehat{q}_g\{n\int_{\mathbf{x},\widetilde{g}}\eta(1, \mathbf{x}, \widetilde{g}; \widehat{\boldsymbol{\theta}})\,d\widehat{F}_{\mathbf{x},g}(\mathbf{x}, \widetilde{g})\}^{-1}} = 1 \quad \text{(A.13)}$$

and

$$\min_{1\le i\le n}\left\{\widehat{\lambda}_g + \frac{n_1\eta(1, \mathbf{X}_i, g; \widehat{\boldsymbol{\theta}})\widehat{q}_g}{n\int_{\mathbf{x},\widetilde{g}}\eta(1, \mathbf{x}, \widetilde{g}; \widehat{\boldsymbol{\theta}})\,d\widehat{F}_{\mathbf{x},g}(\mathbf{x}, \widetilde{g})}\right\} > 0.$$

Clearly, $\widehat{\lambda}_g$ must be bounded asymptotically. Thus, by choosing a subsequence, we assume that $\widehat{\lambda}_g \to \lambda^*_g$.

By (A.13) and the Lipschitz continuity of $\eta(1, \mathbf{x}, g; \boldsymbol{\theta}^*)$ in the continuous components of $\mathbf{x}$, we can show that there exists a positive constant δ such that

$$\min_{g,\mathbf{x}}\left\{\left|\lambda^*_g + \frac{\varrho\eta(1, \mathbf{x}, g; \boldsymbol{\theta}^*)q^*_g}{\int_{\mathbf{x},\widetilde{g}}\eta(1, \mathbf{x}, \widetilde{g}; \boldsymbol{\theta}^*)\,dF^*_{\mathbf{x},g}(\mathbf{x}, \widetilde{g})}\right|\right\} > \delta.$$

Consequently, when n is sufficiently large,

$$\begin{aligned}\widehat{F}_g^\dagger(\mathbf{x}) &= n^{-1}\sum_{i=1}^{n}I(G_i = g, \mathbf{X}_i \le \mathbf{x}) \\ &\quad \times \Bigg(\max\Bigg[\Bigg|\widehat{\lambda}_g + \eta(1, \mathbf{X}_i, g; \widehat{\boldsymbol{\theta}})\widehat{q}_g n_1 \\ &\quad \times \left\{n\int_{\mathbf{x},\widetilde{g}}\eta(1, \mathbf{x}, \widetilde{g}; \widehat{\boldsymbol{\theta}})\,d\widehat{F}_{\mathbf{x},g}(\mathbf{x}, \widetilde{g})\right\}^{-1}\Bigg|, \delta\Bigg]\Bigg)^{-1}.\end{aligned}$$

We define an empirical function $\widetilde{F}_g^\dagger$ whose jump size at $\mathbf{X}_i$ is proportional to

$$\begin{aligned}n^{-1}I(G_i = g)\Bigg(&P(G = g, Y = 0) + \eta(1, \mathbf{X}_i, g; \boldsymbol{\theta}_0)q_g\varrho \\ &\times \left\{\int_{\mathbf{x},\widetilde{g}}\eta(1, \mathbf{x}, \widetilde{g}; \boldsymbol{\theta}_0)\,dF_{\mathbf{x},g}(\mathbf{x}, \widetilde{g})\right\}^{-1}\Bigg)^{-1}.\end{aligned}$$

Then it can be verified that $\widetilde{F}_g^\dagger$ converges uniformly to $F_g^\dagger$. In addition, $\widehat{F}_g^\dagger$ is absolutely continuous with respect to $\widetilde{F}_g^\dagger$, and the Radon–Nikodym derivative $d\widehat{F}_g^\dagger(\mathbf{x})/d\widetilde{F}_g^\dagger(\mathbf{x})$ is bounded and converges uniformly to $dF^*_g(\mathbf{x})/dF^\dagger_g(\mathbf{x})$. Let $\widetilde{F}_{\mathbf{x},g}(\mathbf{x}, g) = \widetilde{F}_g^\dagger(\mathbf{x})q_g$, and let $l_n(\boldsymbol{\theta}, \{F_g^\dagger\}, \{q_g\})$ be the log-likelihood based on (8). By the definition of the MLE, $n^{-1}l_n(\widehat{\boldsymbol{\theta}}, \{\widehat{F}_g^\dagger\}, \{\widehat{q}_g\}) - n^{-1}l_n(\boldsymbol{\theta}_0, \{\widetilde{F}_g^\dagger\}, \{q_g\}) \ge 0$. The limit of this difference is the negative Kullback–Leibler information of the distribution for $(\boldsymbol{\theta}^*, \{F^*_g\}, \{q^*_g\})$ with respect to $(\boldsymbol{\theta}_0, \{F_g^\dagger\}, \{q_g\})$ under $P(Y = 1) = \varrho$. The identifiability conditions then yield $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$, $F^*_g = F_g^\dagger$ and $q^*_g = q_g$. Thus the consistency of $\widehat{\boldsymbol{\theta}}$ is established. Because $F_{\mathbf{x},g}$ is continuous, $\sup_{\mathbf{x},g}|\widehat{F}_{\mathbf{x},g}(\mathbf{x}, g) - F_{\mathbf{x},g}(\mathbf{x}, g)| \to 0$ almost surely.

The derivation of the asymptotic distribution is similar to the proof of theorem 1.2 of Murphy and van der Vaart (2001). We first obtain a score function by differentiating $l_n(\boldsymbol{\theta}, \{F_g^\dagger\}, \{q_g\})$ with respect to $\widehat{\boldsymbol{\theta}}$ along the direction $\mathbf{v}$ and with respect to $\widehat{F}_{\mathbf{x},g}$ along the path $\widehat{F}_\epsilon = \widehat{F}_{\mathbf{x},g} + \epsilon\int\psi(\mathbf{x}, g)\,d\widehat{F}_{\mathbf{x},g}$, where $\mathbf{v}$ has a unit norm and $\psi(\cdot, g)$ is any function whose total variation is bounded by 1. The linearization of the score function around the true parameter value yields

$$\begin{aligned}&n^{1/2}\Bigg\{(\mathbf{v}^T\boldsymbol{\Omega}_{11} + \boldsymbol{\Omega}_{21}[\psi]^T)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ &\qquad + \int(\mathbf{v}^T\boldsymbol{\Omega}_{12} + \Omega_{22}[\psi])\,d(\widehat{F}_{\mathbf{x},g} - F_{\mathbf{x},g})\Bigg\} \\ &\quad = n^{-1/2}\sum_{i=1}^{n}y_i\Bigg\{\mathbf{v}^T l_{\boldsymbol{\theta}}(1, \mathbf{X}_i, G_i; \boldsymbol{\theta}_0, F_{\mathbf{x},g}) \\ &\qquad\qquad + l_F(1, \mathbf{X}_i, G_i; \boldsymbol{\theta}_0, F_{\mathbf{x},g})\Bigg[\int\psi\,dF_{\mathbf{x},g}\Bigg]\Bigg\} \\ &\qquad + n^{-1/2}\sum_{i=1}^{n}(1 - y_i)\Bigg\{\mathbf{v}^T l_{\boldsymbol{\theta}}(0, \mathbf{X}_i, G_i; \boldsymbol{\theta}_0, F_{\mathbf{x},g}) \\ &\qquad\qquad + l_F(0, \mathbf{X}_i, G_i; \boldsymbol{\theta}_0, F_{\mathbf{x},g})\Bigg[\int\psi\,dF_{\mathbf{x},g}\Bigg]\Bigg\} \\ &\qquad + o_p(1),\end{aligned}$$

where $\boldsymbol{\Omega}_{11}$ is a constant matrix, $\boldsymbol{\Omega}_{12}$ is a vector function of $\mathbf{x}$, $\boldsymbol{\Omega}_{21}[\psi]$ and $\Omega_{22}[\psi]$ are linear operators of ψ, and $l_{\boldsymbol{\theta}}$ and l_F are the scores with respect to $\boldsymbol{\theta}$ and $F_{\mathbf{x},g}$. The right side of the foregoing equation converges weakly to a Gaussian process, which depends on $(y_1, y_2, \dots)$ only through ϱ. We can show that the operator $\mathcal{B}[\mathbf{v}, \psi] \equiv \{\mathbf{v}^T\boldsymbol{\Omega}_{11} + \boldsymbol{\Omega}_{21}[\psi]^T, \mathbf{v}^T\boldsymbol{\Omega}_{12} + \Omega_{22}[\psi]\}^T$ is invertible along the lines of Murphy and van der Vaart (2001). It then follows from theorem 3.3.1 of van der Vaart and Wellner (1996) that $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, \widehat{F}_{\mathbf{x},g} - F_{\mathbf{x},g})$ converges weakly to a Gaussian process.

Because the asymptotic distribution depends on $(y_1, y_2, \dots)$ only via ϱ, we assume that $(y_1, y_2, \dots)$ are independent realizations from a Bernoulli distribution with mean ϱ. By choosing some ψ such that $\mathcal{B}[\mathbf{v}, \psi] = (\mathbf{v}^T, 0)^T$ for all $\mathbf{v}$, we see that $\widehat{\boldsymbol{\theta}}$ is an asymptotically linear estimator for $\boldsymbol{\theta}_0$ with the influence function in the score space. It follows from proposition 3.3.1 of Bickel, Klaassen, Ritov, and Wellner (1993) that the limiting covariance matrix of $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ attains the semiparametric efficiency bound.

A.4.7 Proof of Theorem 3. We call the probability distribution induced by (9) the pseudoprobability law, denoted by $\widetilde{P}_n$. Let $f(y, \mathbf{x}, g; \boldsymbol{\theta}, \{F_g\}, a_n)$ be the density function under the true probability law P_n. Because $a_n = o(n^{-1/2})$,

$$\frac{dP_n}{d\widetilde{P}_n} = \exp\left\{ a_n \sum_{i=1}^{n} \frac{\partial \log f(y_i, \mathbf{X}_i, G_i; \boldsymbol{\theta}, \{F_g\}, a)}{\partial a}\bigg|_{a=0} + o(1) \right\} \to^{\widetilde{P}_n} 1.$$

Thus any weak convergence under $\widetilde{P}_n$ also holds for P_n. In addition, by the arguments in the proof of Theorem 2, we can easily verify the results of Theorem 3 when the data are generated from $\widetilde{P}_n$. Thus Theorem 3 holds when the data are generated from P_n.

A.5 Cohort Studies

A.5.1 Identifiability. We show that if two sets of parameters $(\boldsymbol{\theta}, \Lambda)$ and $(\widetilde{\boldsymbol{\theta}}, \widetilde{\Lambda})$ yield the same joint distribution, then $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$ and $\Lambda = \widetilde{\Lambda}$. First, it follows from Lemma 1 that $\boldsymbol{\gamma} = \widetilde{\boldsymbol{\gamma}}$. Suppose that

$$\begin{aligned}
&\sum_{H\in\mathcal{S}(G)} \{\widetilde{\Lambda}(\widetilde{Y}) e^{\widetilde{\boldsymbol{\beta}}^T \mathcal{Z}(\mathbf{X},H)} \dot{Q}(\widetilde{\Lambda}(\widetilde{Y}) e^{\widetilde{\boldsymbol{\beta}}^T \mathcal{Z}(\mathbf{X},H)})\}^{\Delta} \\
&\qquad \times \{1 - Q(\widetilde{\Lambda}(\widetilde{Y}) e^{\widetilde{\boldsymbol{\beta}}^T \mathcal{Z}(\mathbf{X},H)})\}^{1-\Delta} P_{\boldsymbol{\gamma}}(H) \\
&= \sum_{H\in\mathcal{S}(G)} \{\dot{\Lambda}(\widetilde{Y}) e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X},H)} \dot{Q}(\Lambda(\widetilde{Y}) e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X},H)})\}^{\Delta} \\
&\qquad \times \{1 - Q(\Lambda(\widetilde{Y}) e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X},H)})\}^{1-\Delta} P_{\boldsymbol{\gamma}}(H).
\end{aligned}$$

By choosing $\Delta = 1$ and integrating Y from 0 to τ on both sides, we obtain

$$\begin{aligned}
&\sum_{H\in\mathcal{S}(G)} Q(\widetilde{\Lambda}(\tau) e^{\widetilde{\boldsymbol{\beta}}^T \mathcal{Z}(\mathbf{X},H)}) P_{\boldsymbol{\gamma}}(H) \\
&\qquad = \sum_{H\in\mathcal{S}(G)} Q(\Lambda(\tau) e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X},H)}) P_{\boldsymbol{\gamma}}(H).
\end{aligned}$$

Because $Q(\cdot)$ is strictly increasing, the foregoing equation implies that $\widetilde{\Lambda}(\widetilde{Y}) e^{\widetilde{\boldsymbol{\beta}}^T \mathcal{Z}(\mathbf{X},H)} = \Lambda(\widetilde{Y}) e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X},H)}$ for $H = (h, h)$ and $H = (h, \widetilde{h})$. It then follows from Condition 8 that $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$ and $\widetilde{\Lambda} = \Lambda$.

A.5.2 Proof of Theorem 4. Our problem is the same as that of Zeng et al. (2005), except replacing the integration over random effects in that article by the sum over $H \in \mathcal{S}(G)$. The asymptotic properties stated in the theorem follow from the identifiability shown in Section A.5.1 and the proofs of Zeng et al. (2005), provided that we can verify the following result: If there exist a vector $\boldsymbol{\mu} = (\boldsymbol{\mu}_{\boldsymbol{\beta}}^T, \boldsymbol{\mu}_{\boldsymbol{\gamma}}^T)^T$ and a function $\psi(t)$ such that

$$\boldsymbol{\mu}^T l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) + l_{\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)\left[\int \psi \, d\Lambda_0\right] = 0, \tag{A.14}$$

where $l_{\boldsymbol{\theta}}$ is the score function for $\boldsymbol{\theta}$ and $l_{\Lambda}[\int \psi \, d\Lambda_0]$ is the score function for Λ along the submodel $\Lambda_0 + \epsilon \int \psi \, d\Lambda_0$, then $\boldsymbol{\mu} = \mathbf{0}$ and $\psi = 0$.

To prove the desired result, we write out (A.14). We then let $\Delta = 1$ and integrate Y from 0 to τ to obtain

$$\begin{aligned}
&\sum_{H\in\mathcal{S}(G)} \{Q(\Lambda_0(\tau) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)})\} P_{\boldsymbol{\gamma}}(H) \\
&\quad \times \left\{ \frac{\dot{Q}(\Lambda_0(\tau) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)}) \Lambda_0(\tau) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)} \boldsymbol{\mu}_{\boldsymbol{\beta}}^T \mathcal{Z}(\mathbf{X}, H)}{Q(\Lambda_0(\tau) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)})} \right. \\
&\qquad + \frac{\dot{Q}(\Lambda_0(\tau) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)}) \int_0^{\tau} \psi(t) \, d\Lambda_0(t) \, e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)}}{Q(\Lambda_0(\tau) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)})} \\
&\qquad \left. + \boldsymbol{\mu}_{\boldsymbol{\gamma}}^T \nabla_{\boldsymbol{\gamma}} \log P_{\boldsymbol{\gamma}}(H) \right\} = 0.
\end{aligned} \tag{A.15}$$

In contrast, by letting $\Delta = 0$ and $Y = \tau$ in (A.14), we have

$$\begin{aligned}
&\sum_{H\in\mathcal{S}(G)} \{1 - Q(\Lambda_0(\tau) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)})\} P_{\boldsymbol{\gamma}}(H) \\
&\quad \times \left\{ - \frac{\dot{Q}(\Lambda_0(\tau) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)}) \Lambda_0(\tau) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)} \boldsymbol{\mu}_{\boldsymbol{\beta}}^T \mathcal{Z}(\mathbf{X}, H)}{1 - Q(\Lambda_0(\tau) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)})} \right. \\
&\qquad - \frac{\dot{Q}(\Lambda_0(\tau) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)}) \int_0^{\tau} \psi(t) \, d\Lambda_0(t) \, e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)}}{1 - Q(\Lambda_0(\tau) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)})} \\
&\qquad \left. + \boldsymbol{\mu}_{\boldsymbol{\gamma}}^T \nabla_{\boldsymbol{\gamma}} \log P_{\boldsymbol{\gamma}}(H) \right\} = 0.
\end{aligned} \tag{A.16}$$

The summation of (A.15) and (A.16) entails $\boldsymbol{\mu}_{\boldsymbol{\gamma}}^T \nabla_{\boldsymbol{\gamma}} \log P_{\boldsymbol{\gamma}}(H) = 0$. From the proof of Lemma 1, $\boldsymbol{\mu}_{\boldsymbol{\gamma}} = \mathbf{0}$. We choose $G = 2h$ or $h + \widetilde{h}$ and let $\Delta = 1$ and $Y = 0$ in (A.14) to obtain $\boldsymbol{\mu}_{\boldsymbol{\beta}}^T \mathcal{Z}(\mathbf{X}, H) + \psi(0) = 0$ for $H = (h, h)$ and $(h, \widetilde{h})$. Thus, $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \mathbf{0}$ and $\psi(0) = 0$ under Condition 8. Finally, (A.14) with $\Delta = 1$ implies that

$$\psi(\widetilde{Y}) + \frac{\ddot{Q}(\Lambda_0(\widetilde{Y}) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)}) \int_0^{\widetilde{Y}} \psi(t) \, d\Lambda_0(t) \, e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)}}{\dot{Q}(\Lambda_0(\widetilde{Y}) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)})} = 0$$

for $H = (h, h)$. Therefore, $\psi = 0$.

[Received September 2004. Revised February 2005.]

REFERENCES

Akaike, H. (1985), "Prediction and Entropy," in *A Celebration of Statistics*, eds. A. C. Atkinson and S. E. Fienberg, New York: Springer-Verlag, pp. 1–24.

Akey, J., Jin, L., and Xiong, M. (2001), "Haplotypes vs. Single-Marker Linkage Disequilibrium Tests: What Do We Gain?" *European Journal of Human Genetics*, 9, 291–300.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore: Johns Hopkins University Press.

Botstein, D., and Risch, N. (2003), "Discovering Genotypes Underlying Human Phenotypes: Past Successes for Mendelian Disease, Future Approaches for Complex Disease," *Nature Genetics Supplement*, 33, 228–237.

Breslow, N., McNeney, B., and Wellner, J. A. (2003), "Large-Sample Theory for Semiparametric Regression Models With Two-Phase, Outcome-Dependent Sampling," *The Annals of Statistics*, 31, 1110–1139.

Clark, A. G. (1990), "Inference of Haplotypes From PCR-Amplified Samples of Diploid Populations," *Molecular Biology and Evolution*, 7, 111–122.

Cohen, J. (1960), "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, 20, 37–46.

Cox, D. R. (1972), "Regression Models and Life-Tables" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 34, 187–220.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood Estimation From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. (2002), *Analysis of Longitudinal Data* (2nd ed.), Oxford, U.K.: Oxford University Press.

Epstein, M. P., and Satten, G. A. (2003), "Inference on Haplotype Effects in Case-Control Studies Using Unphased Genotype Data," *American Journal of Human Genetics*, 73, 1316–1329.

Excoffier, L., and Slatkin, M. (1995), "Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population," *Molecular Biology and Evolution*, 12, 921–927.

Fallin, D., Cohen, A., Essioux, L., Chumakov, I., Blumenfeld, M., Cohen, D., and Schork, N. (2001), "Genetic Analysis of Case-Control Data Using Estimated Haplotype Frequencies: Application to APOE Locus Variation and Alzheimer's Disease," *Genome Research*, 11, 143–151.

Fisher, R. A. (1918), "The Correlation Between Relatives on the Supposition of Mendelian Inheritance," *Transactions of the Royal Society of Edinburgh*, 52, 399–433.

Hallman, D. M., Groenemeijer, B. E., Jukema, J. W., and Boerwinkle, E. (1999), "Analysis of Lipoprotein Lipase Haplotypes Reveals Associations Not Apparent From Analysis of the Constitute Loci," *Annals of Human Genetics*, 63, 499–510.

International Human Genome Sequencing Consortium (2001), "Initial Sequencing and Analysis of the Human Genome," *Nature*, 409, 860–921.

International SNP Map Working Group (2001), "A Map of Human Genome Sequence Variation Containing 1.42 Million Single Nucleotide Polymorphisms," *Nature*, 409, 928–933.
Lake, S. L., Lyon, H., Tantisira, K., Silverman, E. K., Weiss, S. T., Laird, N. M., and Schaid, D. J. (2003), "Estimation and Tests of Haplotype-Environment Interaction When the Linkage Phase Is Ambiguous," *Human Heredity*, 55, 56–65.
Li, H. (2001), "A Permutation Procedure for the Haplotype Method for Identification of Disease-Predisposing Variants," *Annals of Human Genetics*, 65, 180–196.
Liang, K.-Y., and Qin, J. (2000), "Regression Analysis Under Non-Standard Situations: A Pairwise Pseudolikelihood Approach," *Journal of the Royal Statistical Society*, Ser. B, 62, 773–786.
Lin, D. Y. (2004), "Haplotype-Based Association Analysis in Cohort Studies of Unrelated Individuals," *Genetic Epidemiology*, 26, 255–264.
——— (2005), "An Efficient Monte Carlo Approach to Assessing Statistical Significance in Genomic Studies," *Bioinformatics*, 21, 781–787.
McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), New York: Chapman & Hall.
Morris, R. W., and Kaplan, N. L. (2002), "On the Advantage of Haplotype Analysis in the Presence of Multiple Disease Susceptibility Alleles," *Genetic Epidemiology*, 23, 221–233.
Murphy, S. A., and van der Vaart, A. W. (2000), "On Profile Likelihood," *Journal of the American Statistical Association*, 95, 449–465.
——— (2001), "Semiparametric Mixtures in Case-Control Studies," *Journal of Multivariate Analysis*, 79, 1–32.
Niu, T., Qin, Z. S., Xu, X., and Liu, J. S. (2002), "Bayesian Haplotype Inference for Multiple Linked Single-Nucleotide Polymorphisms," *American Journal of Human Genetics*, 70, 157–169.
Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T. N., Norris, M. C., Sheehan, J. B., Shen, N. P., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P. A., and Cox, D. R. (2001), "Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21," *Science*, 294, 1719–1723.
Pettitt, A. N. (1984), "Proportional Odds Models for Survival Data and Estimates Using Ranks," *Applied Statistics*, 26, 183–214.
Prentice, R. L., and Pyke, R. (1979), "Logistic Disease Incidence Models and Case-Control Studies," *Biometrika*, 66, 403–411.
Qin, Z. S., Niu, T., and Liu, J. S. (2002), "Partition-Ligation-Expectation Maximization Algorithm for Haplotype Inference With Single-Nucleotide Polymorphisms," *American Journal of Human Genetics*, 71, 1242–1247.
Risch, N. (2000), "Searching for Genetic Determinants in the New Millennium," *Nature*, 405, 847–856.
Roeder, K., Carroll, R. J., and Lindsay, B. G. (1996), "A Semiparametric Mixture Approach to Case-Control Studies With Errors in Covariables," *Journal of the American Statistical Association*, 91, 722–732.
Satten, G. A., and Epstein, M. P. (2004), "Comparison of Prospective and Retrospective Methods for Haplotype Inference in Case-Control Studies," *Genetic Epidemiology*, 27, 192–201.
Schaid, D. J. (2004), "Evaluating Associations of Haplotypes With Traits," *Genetic Epidemiology*, 27, 348–364.
Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M., and Poland, G. A. (2002), "Score Tests for Association Between Traits and Haplotypes When the Linkage Phase Is Ambiguous," *American Journal of Human Genetics*, 70, 425–434.
Scott, A. J., and Wild, C. J. (1997), "Fitting Regression Models to Case-Control Data by Maximum Likelihood," *Biometrika*, 84, 57–71.
Seltman, H., Roeder, K., and Devlin, B. (2003), "Evolutionary-Based Association Analysis Using Haplotype Data," *Genetic Epidemiology*, 25, 48–58.
Stephens, M., Smith, N. J., and Donnelly, P. (2001), "A New Statistical Method for Haplotype Reconstruction From Population Data," *American Journal of Human Genetics*, 68, 978–989.
Stram, D. O., Pearce, C. L., Bretsky, P., Freedman, M., Hirschhorn, J. N., Altshuler, D., Kolonel, L. N., Henderson, B. E., and Thomas, D. C. (2003), "Modeling and E–M Estimation of Haplotype-Specific Relative Risks From Genotype Data for a Case-Control Study of Unrelated Individuals," *Human Heredity*, 55, 179–190.
Valle, T., Ehnholm, C., Tuomilehto, J., Blaschak, J., Bergman, R. N., Langefeld, C. D., Ghosh, S., Watanabe, R. M., Hauser, E. R., Magnuson, V., Eriksson, J., Ally, D. S., Nylund, S. J., Hagopian, W. A., Kohtamaki, K., Ross, E., Toivanen, L., Buchanan, T. A., Vidgren, G., Collins, F., Tuomilehto-Wolf, E., and Boehnke, M. (1998), "Mapping Genes for NIDDM," *Diabetes Care*, 21, 949–958.
van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer-Verlag.
——— (2001), "Consistency of Semiparametric Maximum Likelihood Estimators for Two-Phase Sampling," *Canadian Journal of Statistics*, 29, 269–288.
Venter, J., Adams, M., Myers, E., Li, P., Mural, R., Sutton, G., Smith, H. et al. (2001), "The Sequence of the Human Genome," *Science*, 291, 1304–1351.
Wang, S., Kidd, K., and Zhao, H. (2003), "On the Use of DNA Pooling to Estimate Haplotype Frequencies," *Genetic Epidemiology*, 24, 74–82.
Weir, B. S. (1996), *Genetic Data Analysis II*, Sunderland, MA: Sinauer Associates.
Zaykin, D. V., Westfall, P. H., Young, S. S., Karnoub, M. A., Wagner, M. J., and Ehm, M. G. (2002), "Testing Association of Statistically Inferred Haplotypes With Discrete and Continuous Traits in Samples of Unrelated Individuals," *Human Heredity*, 53, 79–91.
Zeng, D., and Lin, D. Y. (2005), "Estimating Haplotype-Disease Associations With Pooled Genotype Data," *Genetic Epidemiology*, 28, 70–82.
Zeng, D., Lin, D. Y., and Lin, X. (2005), "Semiparametric Transformation Models With Random Effects for Clustered Failure Time Data," technical report, University of North Carolina, Chapel Hill, Dept. of Biostatistics.
Zhang, S., Pakstis, A. J., Kidd, K. K., and Zhao, H. (2001), "Comparisons of Two Methods for Haplotype Reconstruction and Haplotype Frequency Estimation From Population Data," *American Journal of Human Genetics*, 69, 906–912.
Zhao, L. P., Li, S. S., and Khalid, N. (2003), "A Method for the Assessment of Disease Associations With Single-Nucleotide Polymorphism Haplotypes and Environmental Variables in Case-Control Studies," *American Journal of Human Genetics*, 72, 1231–1250.

Comment

Chiara SABATTI

The detailed and careful article by Lin and Zeng deals with the estimation of haplotype effects. It is perhaps useful to give a little more genetical background on the problem at hand. Through epidemiological studies (where, e.g., one compares risk of siblings or twins of affected individuals with population prevalence) we can identify that some diseases have a clear genetic component. That is, there are modifications in the DNA sequence that predispose carriers to develop the disease. These modifications have varied nature; there may be mutations, insertions, or deletions in the gene sequence that lead to the synthesis of a different protein, or these variations may take place in noncoding portions of DNA, affecting slicing patterns or expression levels. Understanding the nature of these mutations and their functional effects is of considerable importance; it leads to

Chiara Sabatti is Assistant Professor, Departments of Human Genetics and Statistics, University of California—Los Angeles, Los Angeles, CA 90095 (E-mail: *csabatti@mednet.ucla.edu*).

Journal of the American Statistical Association
March 2006, Vol. 101, No. 473, Theory and Methods
DOI 10.1198/016214505000000817

a clear understanding of the disease origins and suggests drug targets. This goal is achieved in a multistep process. Initially, with genome-wide investigations, one tries to identify regions that harbor susceptibility genes. These broad regions are then analyzed in more detail, to lead to finer mapping and eventually to identification of the disease variants. The study of haplotypes (the collection of allele values at measured polymorphic sites on a chromosome) plays a role in both these phases. Note that typically one does not assume that any of the variants recorded in a haplotype is the disease-causing variant; generally, one simply assumes that there may be an association between the disease-causing variant and the alleles characterizing one haplotype.

To understand the origins of such an association, recall that when a disease-causing mutation arises, it occurs on a specific genetic background (a specific selection of alleles at the surrounding polymorphic loci). As the disease mutation is passed down across generations, the portion of the genome closer to it is also transmitted, with boundaries determined by recombination events. Affected descendants are then carriers not only of the disease-specific mutation, but also of the ancestral haplotype. Under the hypothesis of one disease-causing mutation, after a number of generations have intervened since the founding event, affected individuals that may be practically considered as unrelated will be sharing an identical-by-descent genomic segment surrounding the mutation. This is reflected in the association between the disease status and haplotypes in the region of interest. This scenario implies that the effect of the mutation on fitness must be limited through reduced penetrance, late onset, mild disease symptoms, or recessive mode of inheritance, or other factors, for the mutation to be passed down across generations, originating an association effect. If current cases are in the vast majority due to new mutations, then one would not be able to detect association between disease status and haplotype. In present of multiple mutations with founder effect, there may be multiple "disease" haplotypes, and if the consequences of the mutations are slightly different, then different haplotypes may be associated with different risks.

The study of association of haplotypes with disease status helps in the initial localization of the region harboring the disease gene. One can scan through the genome with a window of fixed genetic length and test for association between disease status and the haplotypes formed by the markers in the window. The locations where association is detected can be considered suspicious of harboring the disease gene. The study of haplotypes effects further contributes to the identification of functional variants and understanding their relevance in determining disease risks.

At the beginning of the article, the authors motivate their study with reference to association mapping, and in the conclusion they refer again to application of their methods to genome scans. In my opinion, however, the specific problem that they tackle is not so much related to mapping as to refining the haplotype effects estimates once a location of interest has been identified. This is an important and worthy goal, because (1) ranking the haplotypes in terms of their association with the disease is necessary to identify those that are most likely to include the disease susceptibility locus, and (2) estimating haplotypes effects helps quantify the impact of the mutation in an epidemiological framework. It may seem that (2) should be more easily and effectively done when such a mutation is found, but going from a locus to the identification of a mutation can still require years of work, especially in complex diseases where gene–gene and gene–environment interactions are expected, so that a preliminary quantification of the effect size is important to appropriately allocate resources. The methodology described in the article shows how quantification of these effects can be obtained, paying particular attention to the identifiability issues and asymptotic efficiency.

Genome screens for association mapping of disease genes may not be the most direct application of the methodology described here. For disease mapping, case-control is definitely the preferred design, and so one needs to consider in particular the methods developed here for this purpose, coupled with the idea of studying haplotype effects for sliding marker windows. Computationally, the efficiency of the algorithm may be unsatisfactory in a case-control study that is based on hundreds of thousands of SNPs, as one can expect for genome-wide investigations. It must be emphasized that the methodology presented here leads to efficient estimates of haplotype effects. For the purpose of mapping, it is not really necessary to get a "very good" estimate of the haplotype effect, only to assess whether there is any such effect. For this reason, one may be able to take advantage of less sophisticated methods that are computationally preferable. For example, consider the association tests implemented in Mendel (Lange et al. 2001), which are quite fast and can appropriately handle missing phase information (more later). Aside from this computational issue, and perhaps more important, a methodology that estimates the effects of well-defined haplotypes will encounter difficulties due to the sparsity of the data, as the author themselves note. Disease haplotypes are eroded by the effects of recombination (which can occur in many different locations in the sampled haplotypes) and mutation. Models that allow one to combine information across haplotypes, grouping "similar" ones, are expected to be more powerful. This idea is at the basis of methodologies described by McPeek and Strahs (1999), Service, Temple Lang, Freimer, and Sandkuijl (1999), Lam, Roeder, and Devlin (2000), Morris, Whittaker, and Balding (2000), Liu, Sabatti, Teng, Keats, and Risch (2001), and Molitor, Marjoram, and Thomas (2003), to cite just a few. Finally, when contemplating the idea of a genome-wide series of association tests, one must put in place some control for multiple comparisons. Although there have been some contributions in this direction, the problem is not yet solved (Sabatti, Service, and Freimer 2003; Lin 2005), and the methodology presented here does not appear to be easily amenable to permutation procedures.

Haplotypes are not directly observable. Polymorphism scoring technologies lead to multiloci genotypes, as the authors explain in their introduction. Information on which chromosome is carrying which observed alleles constituting a genotype (phase) can be inferred using family information, when available, or assuming linkage disequilibrium and using EM (Excoffier and Slatkin 1995), resorting to a Bayesian approach (as in Niu et al. 2002; Stephens et al. 2001), or exploiting the existence of block structure (as in Halperin and Eskin 2004). The authors note the importance of not considering such inferred haplotypes as true ones when making inference on their effects, suggesting simultaneously imputing phase and estimate

haplotype effects. This is an important point. Unfortunately, in much of the genetics literature, inferred haplotypes are considered "true," with varying impacts on the results of analysis. For example, it is common practice to reconstruct haplotypes with EM and then measure the amount of LD on the reconstructed haplotypes. Now, given that EM algorithms operate under the assumption of LD, it may well be that the resulting measures are biased upward. But Lin and Zeng are not the first to note the arbitrariness of this practice, or the first to propose mapping methodologies that deal directly with nonphased data. For example, the mentioned association tests coded in Mendel can be run on multilocus genotypes (Lange et al. 2001; Lazzeroni and Lange 1997), and methods for fine mapping via reconstruction of ancestral haplotypes (as in Liu et al. 2001) take as input unphased data and reconstruct haplotypes in the process. Lu, Niu, and Liu (2003) have compared the results obtained by sequentially reconstructing haplotypes and applying association mapping procedures with results derived by joint inference on the data. In the present contribution, Lin and Zeng suggest using an EM algorithm to deal with missing phase information in their likelihood. It is well known that the performance of EM is unsatisfactory with a large number of markers. However, given that the considered analysis is most interesting when applied to rather short haplotypes, this should not represent a serious limitation.

One interesting aspect of the article is that through their regression model, the authors deal simultaneously with quantitative and qualitative traits. In my comments, I have referred mainly to disease traits, which are typically considered qualitative. However, quantitative traits are also the object of much attention, and indeed the distinction between the two types is often arbitrary (consider, e.g., the threshold models). In practice, the mapping of qualitative and quantitative traits is interwoven; faced with the difficulties in identifying susceptibility loci for complex traits, geneticists are increasingly turning their attention to quantitative endophenotypes, measured on a variety of scales (e.g., brain morphology, gene expression values, enzyme concentrations, personality questionnaires) (see Freimer and Sabatti 2003 for a discussion). The possibility of using the same framework to study quantitative and qualitative traits is definitely an advantage.

Another aspect that makes this article very relevant is that the authors consider a variety of designs. The gene mapping community has been very much focused on family-based or case-control designs. But when a locus is identified and one needs to determine the its effect in a population in its interaction with environmental variables, then other designs, such as cross-sectional and cohort studies, may be more relevant. Also, recent years have seen increased interest in cohort-based genetics studies, exploiting cohorts that have been collected and analyzed with respect to a variety of disease/health states/questionnaires. Genotypes of individuals in these cohorts could be used to assess possible associations between loci and a number of phenotypes of interest, translating the costs incurred by genotyping such large and unspecific samples in savings.

The results obtained in the article concern the identifiability of association parameters and the asymptotic behavior of their estimates in the presence of covariate information. Indeed, it is the consideration of covariates that substantially complicates the statistical analysis. In the context of estimating the population effects of haplotypes associated with increased disease risk, or with any phenotype of interest, it is particularly important to consider covariate information. Gene–environment interactions are known to be important in complex diseases, and the authors should be congratulated for their thorough treatment of the subject.

Although haplotype effects are an important step toward identifying the genetic variation underlying the traits under study, ultimately one would like to isolate the polymorphism that is directly responsible for the phenotype. As mentioned earlier, this may or not be scored in the haplotype considered. When the haplotypes are obtained by genotyping any variation in an identified genomic region, one can reasonably assume that the causative polymorphism is scored. One approach of interest, then, is to try to single out this polymorphism among all of the genotyped ones. For this purpose, regression models similar to the one analyzed here have been proposed (see, e.g., Cordell and Clayton 2002). Often the effects of SNPs are considered one at a time, so that phase information is not always as relevant. However, one would expect that if the causal SNP is not included in the typed set, or if causality is achieved by more than one mutation, then shorter haplotypes may be important explanatory variables. It would be interesting to explore the implications of the results presented here for this problem.

ADDITIONAL REFERENCES

Cordell, H., and Clayton, D. (2002), "A Unified Stepwise Regression Procedure for Evaluating the Relative Effects of Polymorphisms Within a Gene Using Case/Control or Family Data: Application to HLA in Type 1 Diabetes," *American Journal of Human Genetics*, 70, 124–141.

Freimer, N., and Sabatti, C. (2003), "The Human Phenome Project," *Nature Genetics*, 34, 15–21.

Halperin, E., and Eskin, E. (2004), "Haplotype Reconstruction From Genotype Data Using Imperfect Phylogen," *Bioinformatics*, 20, 1842–1849.

Lam, J., Roeder, K., and Devlin, B. (2000), "Haplotype Fine Mapping by Evolutionary Trees," *American Journal of Human Genetics*, 66, 659–673.

Lange, K., Cantor, R., Horvath, S., Perola, M., Sabatti, C., Sinsheimer, J., and Sobel, E. (2001), "Mendel Version 4.0: A Complete Package for the Exact Genetic Analysis of Discrete Traits in Pedigree and Population Data Sets," *American Journal of Human Genetics*, 69(suppl.), A1886.

Lazzeroni, L., and Lange, K. (1997), "Markov Chains for Monte Carlo Tests of Genetic Equilibrium in Multidimensional Contingency Tables," *The Annals of Statistics*, 25, 138–168.

Liu, J., Sabatti, C., Teng, J., Keats, B., and Risch, N. (2001), "Bayesian Analysis of Haplotypes for Linkage Disequilibrium Mapping," *Genome Research*, 11, 1716–1724.

Lu, X., Niu, T., and Liu, J. (2003), "Haplotype Information and Linkage Disequilibrium Mapping for Single Nucleotide Polymorphisms," *Genome Research*, 13, 2112–2117.

McPeek, M., and Strahs, A. (1999), "Assessment of Linkage Disequilibrium by the Decay of Haplotype Sharing, With Application to Fine-Scale Genetic Mapping," *American Journal of Human Genetics*, 65, 858–875.

Molitor, J., Marjoram, P., and Thomas, D. (2003), "Fine-Scale Mapping of Disease Genes With Multiple Mutations via Spatial Clustering Techniques," *American Journal of Human Genetics*, 73, 1368–1384.

Morris, A. P., Whittaker, J. C., and Balding, D. J. (2000), "Bayesian Fine-Scale Mapping of Disease Loci, by Hidden Markov Models," *American Journal of Human Genetics*, 67, 155–169.

Sabatti, C., Service, S., and Freimer, N. (2003), "False Discovery Rates in Linkage and Association Linkage Genome Screens for Complex Disorders," *Genetics*, 164, 829–833.

Service, S., Temple Lang, D., Freimer, N., and Sandkuijl, L. (1999), "Linkage-Disequilibrium Mapping of Disease Genes by Reconstruction of Ancestral Haplotypes in Founder Populations," *American Journal of Human Genetics*, 64, 1728–1738.

Comment

Glen A. SATTEN, Andrew S. ALLEN, and Michael P. EPSTEIN

We congratulate Lin and Zeng for their ambitious and thorough treatment of statistical procedures for haplotype analysis of complex genetic traits. As the authors note, the development of haplotype models is complicated by many factors, including missing data arising from haplotype ambiguity in unphased genotype data and (potentially) high-dimensional nuisance parameters arising from the modeling of covariates. These complications illustrate the many challenging statistical problems in genetic epidemiology that warrant the attention of the wider statistical community.

Our interest has been primarily in case-control studies, and this interest colors our subsequent comments. Several methods have been proposed for modeling haplotype–disease association (e.g., Zhao et al. 2003; Stram et al. 2003; Epstein and Satten 2003; Lake et al. 2003). Of these, only the prospective approaches of Zhao et al. and Lake et al. explicitly incorporate covariates. In the absence of covariates, Satten and Epstein (2004) showed that there could be a remarkable difference in efficiency between prospective and retrospective approaches. This difference seems remarkable in light of the classic result of Prentice and Pyke (1979) on the equivalence of retrospective and prospective analyses of case-control data. In fact, Prentice and Pyke showed that the retrospective likelihood for case-control data was proportional to the prospective likelihood times a factor related to the distribution of exposures, so that if a saturated (nonparametric) model for the exposure distribution were used, then inference based on prospective and retrospective likelihoods would be equivalent. In the haplotype problem, a saturated model for the distribution of haplotypes would not be identifiable given only genotype data; for this reason, the result of Prentice and Pyke does not apply. Carroll, Wang, and Wang (1995) gave some guidance, indicating that a retrospective analysis will always be at least as efficient as a prospective analysis and situations in which the retrospective analysis will be more efficient correspond to cases where the distribution of exposures is restricted. Haplotype models such as those in Lin and Zeng's (3) and (4) correspond to such restrictions.

The situation is considerably more complicated when covariates are included in a model. Here it would appear that the retrospective likelihood is inconvenient, because it requires us to specify the distribution of covariates given disease status, a potentially infinite-dimensional nuisance parameter. Fortunately, several authors, including Lin and Zeng, have shown how to avoid this inconvenience. We feel that these are exciting and important results. The difficulty of modeling the effects of covariate and haplotype–covariate interactions using the retrospective likelihood depends substantially on the presumed relationship between covariates and haplotypes in the population. The simplest model is to assume that haplotypes and covariates are independent at the population level. This model is biologically plausible if population stratification (confounding) is absent and if certain haplotypes or genotypes do not cause the exposure to occur. Although it is not hard to imagine behavioral covariates having genetic influence, the assumption of haplotype–covariate independence is reasonable in many cases.

If haplotypes and covariates are associated at the population level, then the situation is much more complicated. If for each value of covariates X we can assume Hardy–Weinberg equilibrium, then Lin and Zeng's model 3 applies marginally. However, conditionally on covariates X, we would expect a different set of haplotype frequencies for each covariate value. This is a modeling nightmare. Lin and Zeng have avoided this, however, but at a different cost; they make the assumption that haplotypes and covariates are conditionally independent, given genotypes. This assumption is used when defining $m_g(y, x; \theta)$, where the integration is over the distribution of marginal haplotypes corresponding to genotypes g, rather than the distribution of haplotypes conditional on X. With this assumption, specifying the haplotype frequencies at each covariate is unnecessary. However, this assumption is unlikely to hold when X and G are associated, unless genotypes specify haplotypes completely (in which case there are no missing data). Thus we come to a matter of style: Is it better to make a mathematically convenient assumption that leads to more flexible models that may not themselves be biologically plausible, or to stick with a simpler model that may not describe the data as well but that does have a chance of being correct? In fact, the assumption made by Lin and Zeng includes the simpler model as a special case, so that the sole cost of the assumption is an increase in complexity. How does this play out in the case-control setting? If we make the rare disease assumption (corresponding to the only situation where a case-control study makes sense) and assume haplotype–covariate independence at the population level, then it is possible to show that the retrospective likelihood factors into one part involving only the parameters of interest and a second part involving only the nuisance distribution of exposures in the population. This factorization is analogous to the classic result of Prentice and Pyke (1979) that enables prospective analysis of a case-control study even though data are gathered retrospectively. As a final comment on this issue, we note that Lin and Zeng's simulations assume that X and H are independent at the population level, not just conditional on G.

Lin and Zeng give several results on the joint identifiability of parameters governing relative risk and parameters governing the distribution of haplotypes. We applaud these results, even as we note some limitations. First, these results appear limited

Glen A. Satten is Mathematical Statistician, Division of Laboratory Science, Centers for Disease Control and Prevention, Atlanta, GA 30341 (E-mail: *GSatten@cdc.gov*). Andrew S. Allen is Assistant Professor, Department of Biostatistics and Bioinformatics and Duke Clinical Research Institute, Duke University, Durham, NC 27710. Michael P. Epstein is Assistant Professor, Department of Human Genetics, Emory University, Atlanta, GA 30322.

In the Public Domain
Journal of the American Statistical Association
March 2006, Vol. 101, No. 473, Theory and Methods
DOI 10.1198/016214505000000826

to situations in which the model allows only for the effect of a single nonnull haplotype, effectively comparing this "target" haplotype with all others. Tests based on such a procedure can perform very poorly in cases where more than one haplotype affects disease risk. We feel that the modeling approach is especially important in just this case, because hypothesis tests for single haplotype effects are tests of a composite null hypothesis; such tests require the capacity to *estimate* relative risk parameters for those haplotypes not constrained by the null hypothesis. But estimation requires modeling the effects of multiple haplotypes, which appears to go beyond the identifiability results presented by Lin and Zeng. When interest is limited to models of a single haplotype, *tests* can be constructed without much difficulty that are valid regardless of the distribution of haplotypes (see, e.g., Schaid et al. 2002; Zaykin et al. 2002). Thus the identifiability results of Lin and Zeng are most important in situations where one wishes to *estimate* the effect of a single haplotype (relative to all of the others). Can these results actually be used to analyze real data? The answer to this question is less clear, because the haplotype distribution parameters may be only weakly identified in finite samples, especially when the true parameters are close to the null hypothesis or where the true risk model is close to dominant (a fact that will not always be known a priori). Along these lines, we note that in their analyses of the FUSION and simulated data, Lin and Zeng use the stronger assumption that model 3 is correct.

Finally, some quibbles. Lin and Zeng claim to describe analytical methods for "all commonly used study designs." In fact, genetic epidemiologists often use family-based association studies, such as case-parent trio studies, that are not covered by Lin and Zeng's article. Recently, we have developed methods for fitting haplotype risk models using case-parent trio data that are robust to misspecification of the parental haplotype distribution (Allen, Satten, and Tsiatis 2005). We have extended our approach to include haplotype–covariate interactions, where the robustness to misspecification of the parental haplotype distribution enables a general dependence of haplotype frequencies on covariates. These methods are based on the efficient score function; we are currently studying the application of our approach to case-control studies. In particular, it appears possible to remove any dependence of the distribution of H given G in models with no covariates; given the assumption that Lin and Zeng were forced to make, this will be of particular interest should these methods extend to models that include haplotype–covariate interactions in case-control studies. Another design also not considered by Lin and Zeng corresponds to conditional logistic regression of finely stratified data. Here we note that the retrospective approach of Epstein and Satten (2003) can be used with highly stratified data because the intercept parameter is conditioned out; as a result, we can use this approach when we have a large number of intercept parameters. We have also developed an extension of the Epstein and Satten approach that includes covariate effects in addition to haplotype effects (and their interactions) for matched or highly stratified studies.

In summary, we congratulate Lin and Zeng on an interesting and stimulating article.

ADDITIONAL REFERENCES

Allen, A. S., Satten, G. A., and Tsiatis, A. A. (2005), "Locally-Efficient Robust Estimation of Haplotype–Disease Association in Family-Based Studies," *Biometrika*, 92, 559–571.

Carroll, R. J., Wang, S., and Wang, C. Y. (1995), "Prospective Analysis of Logistic Case-Control Studies," *Journal of the American Statistical Association*, 90, 157–169.

Comment

Nilanjan CHATTERJEE, Christine SPINKA, Jinbo CHEN, and Raymond J. CARROLL

Lin and Zeng are to be congratulated on an article that describes identifiability and estimation of haplotype distributions and risk parameters for very general models, both prospectively and for case-control studies. In particular, the identifiability conditions will give important guidance to researchers as they attempt to use different models for haplotypes besides Hardy–Weinberg equilibrium (HWE).

Our major aim in this comment is to place Lin and Zeng's article in the broader context of various alternative methods for haplotype-based regression analysis. We point out the connections and the differences between these alternative methods, to shed light on their relative merits. In particular, we note that in some important subproblems, other methods are available. These methods are efficient and simple to implement, and they avoid the need to estimate possibly high-dimensional nuisance parameters.

1. CASE–CONTROL STUDIES

Because haplotype-based association studies are becoming increasingly popular, a number of researchers have developed methods for logistic regression analysis of case-control studies in the presence of phase ambiguity. The methods can be broadly classified into two categories: prospective and retrospective. Before going into technical details, it is useful to understand the main principles behind these two classes of methods.

Nilanjan Chatterjee is Senior Investigator, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD 20852 (E-mail: *chattern@mail.nih.gov*). Christine Spinka is Assistant Professor, Department of Statistics, University of Missouri, Columbia, MO 65211-6100 (E-mail: *spinkac@missouri.edu*). Jinbo Chen is Research Fellow in the Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD 20852 (E-mail: *chenjin@mail.nih.gov*). Raymond J. Carroll is Distinguished Professor of Statistics, Nutrition and Toxicology, Department of Statistics, Texas A&M University, College Station, TX 77843 (E-mail: *carroll@stat.tamu.edu*). Carroll's research was supported by a grant from the National Cancer Institute (CA-57030) and by the Texas A&M Center for Environmental and Rural Health through a grant from the National Institute of Environmental Health Sciences (P30-ES09106).

In the Public Domain
Journal of the American Statistical Association
March 2006, Vol. 101, No. 473, Theory and Methods
DOI 10.1198/016214505000000835

With a slightly different notation than that of Lin and Zeng, let D be disease status, H^d be the "diplotypes" (i.e., the two haplotypes an individual carries in his or her pair of homologous chromosomes), G be the observed genotype, and X be the nongenetic/environmental covariates. Let the risk function satisfy

$$\text{logit}\{\text{pr}(D = 1|H^d, X)\} = \beta_0 + m(H^d, X, \beta_1), \qquad (1)$$

where $m(\cdot)$ is known but of completely general form.

Under the foregoing notation, the prospective likelihood of the data is given by $\text{pr}(D|G, X)$, which ignores the fact that under the case-control sampling design, data are observed on (G, X) conditional on D. In contrast, the retrospective likelihood of the data is given by $\Pr(G, X|D)$ and accounts for the underlying case-control sampling design.

When there are no missing data (i.e., $G = H^d$), it follows from the well-known results of Prentice and Pyke (1979) that the prospective approach is actually equivalent to the retrospective maximum likelihood analysis, provided that the distribution of the covariates (G, X) is treated completely nonparametrically. Thus the prospective method is a "robust approach" for analysis of case-control studies that does not rely on any assumption about the covariate distribution. In studies of genetic epidemiology, however, it often may be reasonable to assume certain parametric or semiparametric models for the covariate distribution in the underlying source population. The assumptions of HWE and gene–environment independence are examples of such models. The retrospective likelihood can directly incorporate such assumptions into the analysis and can be much more efficient than the prospective method when the assumptions are valid (Epstein and Satten 2003; Chatterjee and Carroll 2005).

1.1 Retrospective Maximum Likelihood Analysis With Haplotype-Phase Ambiguity

Epstein and Satten (2003) first described the retrospective maximum likelihood method for haplotype-based association analysis of case-control studies. Incorporation of nongenetic covariates X in this method is complicated by the fact that the retrospective likelihood involves potentially high-dimensional nuisance parameters that specify the distribution of X in the underlying population. In the gene–environment interaction context, and as in the simulation study and example of Lin and Zeng, it is often reasonable to assume that H^d and environmental factors X are independent in the population, with a parametric form

$$\text{pr}(H^d = h^d|X) = \text{pr}(H = h^d) = q(h^d, \theta), \qquad (2)$$

where the model $q(h^d; \theta)$ in turn could be specified according to HWE or some of its extensions, as considered by Lin and Zeng. More generally, one can assume a parametric model for the diplotype distribution of the form

$$\text{pr}(H = h^d|X = x) = q(h^d, x, \theta). \qquad (3)$$

Model (3), for example, can incorporate departure from gene–environment independence and HWE that may be caused by "population stratification." In particular, one could assume HWE and gene–environment independence conditional on various demographic factors, such as ethnicity and geographic regions, and specify the haplotype frequencies conditional on these factors according to a parametric model, such as the polytomous logistic regression model (Spinka, Carroll, and Chatterjee 2005). Moreover, (3) potentially can be used to directly model the association between haplotypes and environmental exposure X.

Under models (2) and (3), Spinka et al. (2005) described simple and easily computable methods that avoid estimating the nonparametric marginal distribution of X and exploit the information available in (2) or (3) to increase efficiency. Chatterjee, Kalaylioglu, and Carroll (2005) described similarly simple methods applicable for family-based or other types of individually matched case-control studies. Let there be n_1 cases and n_0 controls, and let $\pi = \text{pr}(D = 1)$ be the marginal probability of the disease in the population. Assume the definitions

$$\kappa = \beta_0 + \log(n_1/n_0) - \log\{\pi/(1 - \pi)\}$$

and

$$S(d, h, x, \Omega) = q(h, x, \theta)\frac{\exp[d\{\kappa + m(h, x, \beta_1)\}]}{1 + \exp\{\beta_0 + m(h, x, \beta_1)\}},$$

where $\Omega = (\beta_0, \kappa, \theta^{\text{T}}, \beta_1^{\text{T}})^{\text{T}}$. Let $\mathcal{H}_G$ be the set of diplotypes consistent with the observed genotype G. Define

$$L^*(D, G, X, \Omega) = \frac{\sum_{h\in\mathcal{H}_G} S(D, h, X, \Omega)}{\sum_h \sum_d S(d, h, X, \Omega)}.$$

Spinka et al. (2005) first showed that under certain conditions, which are easily verifiable from the data, all of the parameters in Ω, including the intercept parameter β_0, are identifiable from the retrospective likelihood $\prod_i \Pr(G_i, X_i|D_i)$, as long as the underlying models are specified in such a way that Ω would be identifiable from prospective studies. Moreover, the maximum retrospective likelihood estimate of Ω can be obtained as a solution of the score equation corresponding to the pseudolikelihood

$$l^* = \sum_{i=1}^{N} \log\{L^*(D_i, G_i, X_i, \Omega)\}. \qquad (4)$$

Spinka et al. described strategies for estimating the regression parameter β_1 based on l^* for both known and unknown values of the marginal probability of the disease in the underlying population. If one is also willing to make the rare disease assumption *for all* H^d and X, then ℓ^* effectively becomes equivalent to the method that Lin and Zeng derived in their section A.4.5 under the assumption of gene–environment independence. Note, however, that neither the rare disease approximation nor the gene–environment independence assumption is necessary to derive the simple pseudolikelihood ℓ^*.

An alternative representation of l^* is very revealing. Consider a sampling scenario where each subject from the underlying population is selected into the case-control study using a Bernoulli sampling scheme where the selection probability for a subject given his or her disease status $D = d$ is proportional to $\mu_d = N_d/\text{pr}(D = d)$. Let $R = 1$ denote the indicator of whether a subject is selected in the case-control sample under the foregoing Bernoulli sampling scheme. With some algebra, one can

now show that the pseudolikelihood l^* can be expressed in the form

$$l^* = \sum_{i=1}^{N} \log\Bigg\{ \sum_{h^d \in \mathcal{H}_{G_i}} \mathrm{pr}(D_i | H_i^d = h^d, X_i, R_i = 1) \times \mathrm{pr}(H_i^d = h^d | X_i, R_i = 1) \Bigg\}$$
$$= \sum_{i=1}^{N} \log\{\mathrm{pr}(D_i, G_i | X_i, R_i = 1)\}. \quad (5)$$

When no environmental factors are involved, Stram et al. (2003) proposed an analysis of haplotype-based case-control studies using an "ascertainment-corrected joint likelihood" of the form $\prod_i \mathrm{pr}(D_i, G_i | R_i = 1)$. The representation of the l^* given in (5) suggests that under model (2) or (3) with $F(x)$ treated completely nonparametrically, the efficient retrospective maximum likelihood estimate of the haplotype frequency and the regression parameters can be obtained by conditioning on X in the approach of Stram et al. (2003).

In most parts of their article, Lin and Zeng considered retrospective maximum likelihood estimation under the model that assumes H^d and X are independent given G and then allows the distribution of $[X|G]$ to be completely nonparametric. This model has advantages and disadvantages. It is more flexible than the model (2) that assumes H^d and X are independent unconditionally; however, unlike model (3), it cannot allow direct association between haplotypes and environmental/demographic factors. Computationally, retrospective maximum likelihood assuming model (2) or model (3) completely avoids estimation of the distribution of the possibly high-dimensional covariates X. In contrast, under the model considered by Lin and Zeng, one must estimate the nonparametric distribution of X for each different genotype G, possibly stratified by subpopulations—a potentially daunting task. Finally, in situations where the gene–environment independence assumption is likely to be valid, either in the entire population or within subpopulations, based on results of Chatterjee and Carroll (2005) and Spinka et al. (2005), we conjecture that the retrospective maximum likelihood method assuming model (2) or model (3) can be much more efficient than that assuming the model of Lin and Zeng.

1.2 Prospective Methods for Retrospective Data

Lake et al. (2003) described methods for haplotype-based regression analysis based on the prospective likelihood of the data (D, G, X), ignoring the true case-control sampling design. For fixed values of the haplotype-frequency parameter θ, the score equations for the regression parameters $\beta^* = (\kappa, \beta_1)$ under model (2) corresponding to the prospective likelihood of the data is given by

$$0 = \sum_{i=1}^{N} \sum_{h^d \in \mathcal{H}_{G_i}} \frac{\partial}{\partial \beta^*} \log\{\mathrm{pr}_{\beta^*}(D_i | h^d, X_i)\} \times \mathrm{pr}_{\beta^*}(D_i | h^d, X_i) q(h^d; \theta) \times \Bigg(\sum_{h^d \in \mathcal{H}_{G_i}} \mathrm{pr}_{\beta^*}(D_i | h^d, X_i) q(h^d; \theta) \Bigg)^{-1}. \quad (6)$$

Unfortunately, this purely prospective score equation is biased under the case-control sampling design, even if the true haplotype frequencies were known and the underlying HWE and gene–environment independence assumptions were valid. However, a simple modification of the prospective score equation is unbiased,

$$0 = \sum_{i=1}^{N} \sum_{h^d \in \mathcal{H}_{G_i}} \frac{\partial}{\partial \beta^*} \log\{\mathrm{pr}_{\beta^*}(D_i | h^d, X_i)\} \times \mathrm{pr}_{\beta^*}(D_i | h^d, X_i) r_\Omega(h^d, X_i) q(h^d; \theta) \times \Bigg(\sum_{h^d \in \mathcal{H}_{G_i}} \mathrm{pr}_{\beta^*}(D_i | h^d, X_i) r_\Omega(h^d, X_i) q(h^d; \theta) \Bigg)^{-1}, \quad (7)$$

where

$$r_\Omega(h^d, X) = \frac{1 + \exp\{\kappa + m(h^d, x, \beta_1)\}}{1 + \exp\{\beta_0 + m(h^d, x, \beta_1)\}}.$$

Spinka et al. showed that with an appropriate rare disease approximation, the modified prospective estimating equation (7) is equivalent to the approximate estimating equation approach proposed by Zhao et al. (2003). Spinka et al. described strategies for estimating β_1 and κ based on the modified prospective estimating equation (7), where the nuisance parameters θ, and possibly β_0, are estimated based on score equation derived from the pseudolikelihood l^*. Simulation studies show that such a prospective approach generally tends to be much more robust to violation of both HWE and the gene–environment independence assumption compared with the retrospective maximum likelihood method (see also Satten and Epstein 2004).

2. COHORT–BASED STUDIES AND THE COX PROPORTIONAL HAZARDS MODEL

Lin and Zeng admirably describe fully efficient nonparametric maximum likelihood estimation for fitting a general haplotype-based semiparametric linear transformation model to cohort studies with unphased genotype data. An alternative estimator considered by Chen, Peters, Foster, and Chatterjee (2004) and Chen and Chatterjee (2005) for the popular Cox proportional hazard (CPH) model deserves attention. Consider the CPH model for specifying the hazard function for a subject given his or her diplotype status (H^d) and environmental covariates (X) as

$$\lambda[t | H^d, X] = \lambda_0(t) R(H^d, X; \beta_1), \quad (8)$$

where $\lambda_0(t)$ is the unspecified baseline hazard function, $R(H^d, X; \beta_1)$ is a parametric function describing the relative risk associated with the exposure (H^d, X), and β_1 is the vector of associated regression parameters of interest. As before, assume that $\Pr(H^d) = q(H^d; \theta)$ is specified according to HWE and that θ denotes the associated haplotype frequency parameters. The model (8) cannot be used directly, because H^d is not observable. Following Prentice (1982), one can derive the hazard function for disease conditional on the observable genotype data G and

covariates X in the form

$$\lambda[t|G, X] = \lambda_0(t)R^*\{G, X; t, \beta_1, \theta, \Lambda_0(\cdot)\}, \quad (9)$$

where

$$R^*\{G, X; t, \beta_1, \theta, \Lambda_0(\cdot)\} = E\{R(H^d, X; \beta_1)|G, X, T \geq t\} = \frac{\sum_{H^d \in \mathcal{H}_G} R(H^d, X; \beta_1)\,\mathrm{pr}[T > t|H^d, X]q(H^d; \theta)}{\sum_{H^d \in \mathcal{H}_G} \mathrm{pr}[T > t|H^d, X]q(H^d; \theta)}.$$

In general, standard partial likelihood inference cannot be performed based on (9), because the relative risk function $R^*\{G, X; t, \beta_1, \theta, \Lambda_0(\cdot)\}$ itself depends on the baseline hazard function $\lambda_0(t)$. However, Chen et al. (2004) showed that an omnibus score test for genetic association can be performed using outputs from standard statistical software for partial likelihood analysis. Based on (9), Chen and Chatterjee (2005) also described alternative strategies for estimation of the risk parameters β_1. In particular, the authors observed that for rare disease, one could assume that $\mathrm{pr}[T > t|H^d, X] \approx 1$. The corresponding induced relative risk function,

$$R^*(G, X; \beta_1; \theta) = \frac{\sum_{H^d \in \mathcal{H}_G} R(H^d, X; \beta_1)q(H^d; \theta)}{\sum_{H^d \in \mathcal{H}_G} q(H^d; \theta)},$$

is free of the baseline hazard function $\lambda_0(t)$. Thus, under the rare disease approximation, one could estimate β by maximizing the partial likelihood associated with the relative risk function $R^*(G, X; \beta_1; \widehat{\theta})$, where $\widehat{\theta}$ is a consistent estimate of the haplotype-frequency parameters θ. Chen and Chatterjee described alternative strategies for obtaining consistent estimate of θ for cohort and nested case-control studies. A simple asymptotic variance estimator was also provided. Simulation studies for the full cohort design show that the loss of efficiency in this pseudolikelihood method was quite small compared with the fully efficient nonparametric maximum likelihood estimator (NPMLE) estimator proposed by Lin (2004).

An advantage of pseudolikelihood approach of Chen and Chatterjee (2005) is its wide applicability to alternative cohort-based study designs. In particular, for studies of rare diseases such as cancer, it is common to conduct case-control or case-cohort sampling within a cohort to select a subset of people for whom genotype and expensive environmental exposure information will be collected. Various alternative types of partial likelihoods that are currently available for analysis of nested case-control and case-cohort studies can be applied to estimate β_1 based on the induced relative risk function $R^*(G, X; \beta_1; \widehat{\theta})$. Future research is merited to study whether and how one can obtain the NPMLE for these alternative designs, especially when *both* genotype and environmental exposure data are available only for the selected subsample of the subjects. We look forward to Lin and Zeng's further innovations in this area.

ADDITIONAL REFERENCES

Chatterjee, N., and Carroll, R. J. (2005), "Semiparametric Maximum Likelihood Estimation in Case-Control Studies of Gene–Environment Interactions," *Biometrika*, 92, 399–418.

Chatterjee, N., Kalaylioglu, Z., and Carroll, R. J. (2005), "Exploiting Gene–Environment Independence in Family-Based Case-Control Studies: Increased Power for Detecting Associations, Interactions and Joint Effects," *Genetic Epidemiology*, 28, 138–156.

Chen, J., and Chatterjee, N. (2005), "Haplotype-Based Association Analysis in Cohort and Nested Case-Control Studies," *Biometrics*, in press.

Chen, J., Peters, U., Foster, C., and Chatterjee, N. (2004), "A Haplotype-Based Test of Association Using Data From Cohort and Nested Case-Control Epidemiologic Studies," *Human Heredity*, 58, 18–29.

Prentice, R. L. (1982), "Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model," *Biometrika*, 69, 331–342.

Spinka, C., Carroll, R. J., and Chatterjee, N. (2005), "Analysis of Case-Control Studies of Genetic and Environmental Factors With Missing Genetic Information and Haplotype-Phase Ambiguity," *Genetic Epidemiology*, 29, 105–127.

Comment

Jung-Ying TZENG and Kathryn ROEDER

All data analysis relies on a model that is, strictly speaking, not correct. Choices about which features to model and which to ignore distinguish successful models from the rest. Without artful modeling, statisticians would be unable to make inferences based on finite samples. In this wide-ranging article, Lin and Zeng (LZ hereinafter) make novel contributions to the statistical genetics literature by introducing new models and providing a rigorous statistical analysis of these models. Specifically, their article builds on a series of related works modeling the effect of haplotypes on the risk of disease. We congratulate the authors for providing a firm theoretical foundation in this exciting area of research. The authors investigate a family of models that address a broad range of sampling designs commonly used in genetic epidemiology, but for brevity we focus our remarks on those models appropriate to case-control data.

Schaid et al. (2002) published a practical methodological approach for haplotype association analysis using a prospective model to link the risk of disease to observed genetic data. The chosen model ignored two features of the data: the case-control sampling scheme that typically generates the data and poten-

Jung-Ying Tzeng is Assistant Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695 (E-mail: *jytzeng@stat.ncsu.edu*). Kathryn Roeder is Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 (E-mail: *roeder@stat.emu.edu*).

Journal of the American Statistical Association
March 2006, Vol. 101, No. 473, Theory and Methods
DOI 10.1198/016214505000000844

tial violations of "Hardy–Weinberg equilibrium (HWE)" in the pairwise distribution of haplotypes in the population. Zhao et al. (2003) proposed a similar model. By ignoring the retrospective nature of the data, these authors were able to easily incorporate environmental covariates into their models. Both of these articles took a hypothesis testing approach and focused on testing for the presence of haplotype effects, rather than estimating the size of such effects.

Subsequently, Epstein and Satten (2003) introduced an approach based on a retrospective model that accounts for case-control sampling of the data. Building on these results, Satten and Epstein (2004) chose not to assume HWE. They used a population genetics model that permits "inbreeding" (a particular violation of independence of haplotype pairings within individuals). Their retrospective approach does not easily facilitate the inclusion of environmental covariates. LZ continue in this line by developing more general models that handle all three data features described earlier: inbreeding, environmental covariates, and retrospective sampling.

Each step in this progression has led to models of increasing complexity. Clearly, more complex models may more closely reflect reality. The question is whether the extra complexity improves the inferences in realistic situations. This is the question that we pursue in this discussion.

Consider "inbreeding." This term is generally used to describe a situation in which there is an increase in the probability of observing matching genetic information at the pair of haplotypes within an individual, that is, $\pi_{kk} > \pi_k^2$. This excess of matching haplotypes (called homozygotes) tends to follow the model $\pi_{kk} = \rho\pi_k + (1 - \rho)\pi_k^2$, where ρ is the probability that both of a pair of inherited haplotypes trace back to a common ancestor. Excess homozygosity in the controls arises naturally under four scenarios: cultural practices that encourage marriage between close relatives, insular populations for whom the present generation traces back to a small number of ancestors, populations consisting of subpopulations whose members rarely intermarry (i.e., population substructure), and laboratory error.

Cultural norms generally discourage marriage among relatives, and hence inbreeding levels in human populations tend to be very low. For instance, LZ estimate $\rho = .0002$ in the FUSION data described in their article. In populations in which inbreeding is considered acceptable, even encouraged, Donbak (2004) assessed the level of inbreeding to be $\rho = .015$. Contrast this with the level of inbreeding ($\rho = .05$) used by Satten and Epstein (2004) and LZ in their simulation studies. For a large population to attain inbreeding levels of .05 would require a substantial fraction of the marriages to be between first and second cousins (Lange 1997).

The other potential for inbreeding, insular populations, is best illustrated by the Hutterite population of North Dakota, whose ancestry can be traced back to 90 ancestors in the 1,700s/1,800s. Yet even this unique population has only a moderate amount of inbreeding ($\rho = .03$) (Bourgain et al. 2003).

Thus true inbreeding in most human populations is considerably less than .05. Yet we sometimes observe a substantial excess of "homozygotes" in control subjects, leading to strong violations of HWE. Presumably the third and fourth features are the likely causes.

Population substructure is known to lead to excesses of homozygosity in practice (e.g., Devlin, Risch, and Roeder 1990). If the violation in HWE is due to population substructure, then we argue that it is not acceptable to perform further analysis of the data using the type of association analysis developed by LZ. Why? Geneticists are not interested in simply finding associations; rather, they wish to discover causal associations. However, population substructure leads to confounding due to Simpson's paradox. The lurking variable in this context is subpopulation membership. For example, suppose that the disease is more common in one subpopulation than the other. Any genetic marker that differs strongly in distribution across the subpopulations will exhibit a spurious association with the disease (Lander and Schork 1994; Devlin and Roeder 1999). Consequently, great care must be taken to control for population substructure in good quality studies of genetic association. If it is impossible to control for this feature directly by stratifying the data by subpopulation, then a different experimental design similar to matched case-control is often used (Ewens and Spielman 1995). Thus by careful design, we can usually rule out population substructure as the cause of excess homozygosity.

Laboratory error often yields an excess of apparent homozygotes; however, these violations are rarely seen in the final stages of data analysis. Standard laboratory practice involves screening all genetic markers that result in violations of HWE to determine whether they are the consequence of errors. Thus in careful studies, violations of HWE in the control sample are unlikely in practice. If a testing framework were used, it would be reasonable to assume HWE under the null hypothesis.

Next, we consider "testing" versus "estimating" the effect of haplotypes. Traditionally, genetic epidemiologists have tested $\beta = 0$ rather than attempting to estimate the haplotypic effect. To understand why, it is necessary to consider the history of genetic epidemiology in this domain. Although geneticists tracing back to Fisher have been remarkably adept in their use of quantitative tools, the data that they deal with are not always amenable to modeling that is excessively detailed. Moreover, progress in mapping the genetic risk factors for complex disease has been slow, due to the "needle in a haystack" nature of the quest. Millions of polymorphisms potentially could be tested for association. A disease may be caused by multigenic and/or environmental factors, which are likely to interact in complex ways. Effects of individual genes are likely to be small and, even worse, to vary across ethnic groups. For these reasons, the odds are against discovering risk factors, let alone refined estimates of the magnitude of these risks.

In practice, however, a number of steps are taken to improve the likelihood of success. Although some steps could potentially introduce estimation bias of β, they are amenable to simple testing for the presence of a genetic effect. For example, to enhance the chance that participants in a study have the disease due to a genetic cause (rather than environmental ones), diseased subjects often are included only if they have close relatives with the disease. This type of sampling leads to an ascertainment bias in β. This is just one of the substantial biases that might be present in a genetic study.

LZ's model requires specification of a particular haplotype as the associated one. This requirement is made rather cavalierly, considering that there typically is no prior knowledge

about the relationship between haplotypes and a disease phenotype. Indeed, for the most part haplotypes do not cause disease; rather, one or more genetic variants do. Haplotypes are used as proxies for these variants in association studies because the spatial correlation within the haplotypes often leads to a correlation between a causal variant and one or more distinct haplotypes. Certainly there is no guarantee that there will be a one-to-one mapping between a haplotype and a causal variant. Consequently, there is no reason to believe that specific haplotypic effects are interpretable or reproducible across ethnic groups or studies.

What is the alternative to the choice of preselecting the associated haplotype? Even in the testing framework this is a challenging problem. Chapman, Cooper, Todd, and Clayton (2003) performed some intriguing simulations and reached the conclusion that haplotypes have very low power in a search for genetic risk factors. Their work suggests that greater power can be obtained by analyzing a set of single nucleotide polymorphisms (SNPs) with a simple application of Hotelling's T^2 (Fan and Knapp 2003). Roeder, Bacanu, Sonpar, Zhang, and Devlin (2005) took this view one step further and showed that the maximum of single SNP tests is yet more powerful than Hotelling's T^2 in some instances. The reason that naive haplotype analysis performs poorly is because it requires too many degrees of freedom when the causal haplotype is not known a priori. To fully capitalize on their potential, haplotype-based methods that do not dilute their power to detect interesting relationships between haplotypes and phenotypes in the sheer volume of distinct haplotypes are needed. Tzeng (2005) used haplotype similarity to cluster related types and reduce the dimension of the problem. This approach can improve the power of the test. Seltman, Roeder, and Devlin (2001, 2003) proposed a method of haplotype analysis that exploits the evolutionary history of the haplotypes to limit the tests performed. This approach increases the interpretability and the power of generic haplotype analysis. Alternatively, Tzeng, Byerley, Devlin, Roeder, and Wasserman (2003) described an approach that tests for any unusual sharing of haplotypes among the cases that requires only a single degree of freedom. (For a summary of the challenges in this area, see Clayton, Chapman, and Cooper 2004.)

Thus far we have discussed three scenarios that may introduce bias in estimators of β: (A) inbreeding, (B) ascertainment bias, and (C) a causal SNP not uniquely associated with a haplotype. To examine the impact of these violations of assumptions, we generated $n = 500$ cases and controls using (12) from LZ with $\alpha = -4$ and $\beta_2 = \beta_3 = 0$. To expedite the simulations, we made a simplification: We simulated the data from a list of three haplotypes ($h_1 = 11$, $h_2 = 01$, and $h_3 = 00$) with probabilities ($\pi_1 = .2$, $\pi_2 = .3$, and $\pi_3 = .5$). To estimate β_1 we used LZ's rare disease algorithm as described in section A.4.5, but assuming HWE. Under scenario (A) we allowed a realistic violation of HWE and simulated data with an inbreeding coefficient of $\rho = .015$. In scenario (B), we drew the cases from families with an affected sibpair (one child sampled per family). In scenarios (A) and (B), the causal SNP is in position 1, which leads to a unique mapping between h_1 and SNP_1. In scenario (C), the causal SNP is in position 2. Consequently, for this scenario both h_1 and h_2 are associated with the phenotype. For all three scenarios, only the effect of h_1 was investigated.

Table 1. Bias and Mean Squared Error (MSE) of $\hat{\beta}_1$

β_1	Scenario	Bias	MSE
.5	Baseline	−.002	.010
	[A] $\rho = .015$	.023	.012
	[B] Ascertainment bias	.273	.083
	[C] Multiple causal haplotypes	−.217	.058
0	Baseline	−.002	.013
	[A] $\rho = .015$	.001	.012
	[B] Ascertainment bias	.002	.008
	[C] Multiple causal haplotypes	.002	.012

NOTE: The results are based on 200 simulations with 500 cases and 500 controls. Scenario "Baseline" refers to $\rho = 0$, no ascertainment bias, and one causal haplotype h_1.

The bias of $\hat{\beta}_1$ is indicated in Table 1. Note that the observed bias is negligible when inbreeding is ignored. This small level of bias is likely the reason why virtually every published method for estimating haplotype distribution is based on a model that assumes HWE. Alternatively, ascertainment bias leads to a substantial bias in the estimated effect of the haplotype. Clearly, if the size of β_1 is of interest, then care must be taken to model the ascertainment process. Finally, the bias due to nonunique association between SNPs and haplotypes is also considerable. This supports our view that it may be difficult to interpret β_1 in practice.

Thus, although LZ provide models that are closer to truth, we believe the nature of the data is not yet in sync with the refinements to the model that LZ have introduced. Nevertheless, technology continues to improve the quality and amount of data available. The likelihood of success in the quest to discover the genetic risk factors for complex disease rises accordingly. So although we think LZ are ahead of their time for most genetic analyses at the moment, as the data improve and become more focused, it will be advantageous for statisticians to have refined models with good properties at their fingertips.

ADDITIONAL REFERENCES

Bourgain, C., Hoffjan, S., Nicolae, R., Newman, D., Steiner, L., Walker, K., Reynolds, R., Ober, C., and McPeek, M. S. (2003), "Novel Case-Control Test in a Founder Population Identifies P-Selection as an Atopy-Susceptibility Locus," *American Journal of Human Genetics*, 73, 612–626.

Clayton, D., Chapman, J., and Cooper, J. (2004), "Use of Unphased Multilocus Genotype Data in Indirect Association Studies," *Genetic Epidemiology*, 27, 415–428.

Chapman, J. M., Cooper, J. D., Todd, J. A., and Clayton, D. G. (2003), "Detecting Disease Associations Due to Linkage Disequilibrium Using Haplotype Tags: A Class of Tests and the Determinants of Statistical Power," *Human Heredity*, 56, 18–31.

Devlin, B., Risch, N., and Roeder, K. (1990), "No Excess of Homozygosity at Loci Used for DNA Fingerprinting," *Science*, 240, 1416–1420.

Devlin, B., and Roeder, K. (1999), "Genomic Control for Association Studies," *Biometrics*, 55, 997–1004.

Donbak, L. (2004), "Consanguinity in Kahramanmaras City, Turkey, and Its Medical Impact," *Saudi Medical Journal*, 25, 1991–1994.

Ewens, W. J., and Spielman, R. S. (1995), "The Transmission/Disequilibrium Test: History, Subdivision, and Admixture," *American Journal of Human Genetics*, 57, 455–464.

Fan, R., and Knapp, M. (2003), "Genome Association Studies of Complex Diseases by Case-Control Designs," *American Journal of Human Genetics*, 72, 850–868.

Lander, E. S., and Schork, N. J. (1994), "Genetic Dissection of Complex Traits," *Science*, 265, 2037–2048.

Lange, K. (1997), *Mathematical and Statistical Methods for Genetic Analysis*, New York: Springer-Verlag.

Roeder, K., Bacanu, S. A., Sonpar, V., Zhang, X., and Devlin, B. (2005), "Analysis of Single-Locus Tests to Detect Gene/Disease Associations," *Genetic Epidemiology*, 28, 207–219.

Seltman, H., Roeder, K., and Devlin, B. (2001), "TDT Meets MHA: Family-Based Association Analysis Guided by Evolution of Haplotypes," *American Journal of Human Genetics*, 68, 1250–1263.

Tzeng, J. Y. (2005), "Evolutionary-Based Grouping of Haplotypes in Association Analysis," *Genetic Epidemiology*, 28, 220–231.

Tzeng, J. Y., Byerley, W., Devlin, B., Roeder, K., and Wasserman, L. (2003), "Outlier Detection and False Discovery Rates for Whole-Genome DNA Matching," *Journal of the American Statistical Association*, 98, 236–247.

Comment

Hongzhe Li

1. INTRODUCTION

I congratulate Professors Lin and Zeng (LZ for short) on their fine work on inference on haplotype effects in genetic association studies. The completion of the Human Genome Project and near completion of the HapMap project make genomewide genetic association studies of complex traits now a reality. One of the main challenges is performing rigorous and efficient statistical analysis for identifying genetic variants that explain the variation of the traits of interest in the population. Such traits can be binary, such as disease status; quantitative, such as blood pressure; censored survival data, such as age of disease onset; or longitudinal or functional data, such as growth curve. LZ provide a unified framework for haplotype association analysis for different types of traits and several different study designs. Although the likelihood-based methods using the EM algorithm for inferring the missing phases of the haplotypes have been studied and reported by many authors in recent years, most of these articles were published in genetics journals or genetic epidemiological journals, and some of them lack statistical rigor. The major contribution of LZ's article is to provide a rigorous treatment to the problem and the methods developed previously, especially in terms of parameter identifiability and the assumptions for validity of the likelihood-based statistical inferences.

Instead of commenting on the theoretical content of the article, I provide some comments on the following aspects: the design of genetic association studies, alternative parameterization of genetic variants, and incorporation of additional biological information into genetic association analysis.

2. STUDY DESIGNS

LZ consider several commonly used study designs in traditional epidemiological studies, including cross-sectional studies, case-control studies with known or unknown population totals, and cohort designs. Due to the cost of genotyping and collection of environmental covariates, the most practical and the most commonly used design among these listed for large-scale genetic association studies is the case-control design with unknown population totals. Inferences for such a design were investigated by LZ in their simulations and application to the FUSION data. The traditional cohort design is probably not practical or necessary for large-scale genetic association studies. If age of onset and haplotype-specific risk ratios are important, then case-cohort or nested case-control design may provide a feasible alternative, especially for relatively rare diseases. Chen, Peters, Foster, and Chatterjee (2004) and Chen and Chatterjee (2005) proposed likelihood-based score tests for haplotype association and likelihood-based procedures for estimating the haplotype risk ratio parameters in the framework of the Cox proportional hazards model, where they proposed first estimating the haplotype frequencies and then estimating the risk ratio parameters. It should be easy to develop a nonparametric maximum likelihood estimator (NPMLE) procedure in the framework of missing data. Under the case-cohort or nested case-control design, there are two types of missing data. For those who were cases or were selected as controls or a subcohort, we may miss the phases of the haplotypes; for those who were disease-free but were not selected as controls or a subcohort, we miss their genotypes and of course also their haplotypes. The EM algorithm can be developed for the NPMLE to estimate the haplotype frequencies, the baseline hazard function, and the haplotype-specific risk ratios simultaneously. It is also interesting to develop rigorous statistical inference procedure for the class of semiparametric linear transformation models considered by LZ for case-cohort or nested case-control designs. Such procedures should contribute greatly to the future of the haplotype-based genetic association analysis.

3. PARAMETERIZATION OF GENETIC VARIANTS

Haplotype-based genetic association analysis as considered by LZ provides one way of identifying genetic variants that are related to complex traits. However, there are some unsolved practical issues when a large set of SNPs are typed and investigated; for example, how many SNPs one should consider in haplotype analysis, and how one can efficiently deal with many rare haplotypes. Recent idea is to use evolutionary-based grouping of rare haplotypes in association analysis to reduce the degrees of freedom (Tzeng 2005). Such grouping depends, of course, on the population models assumed, which sometimes can be difficult to justify. In addition, the models parameterized in terms of haplotypes may not be appropriate for modeling complex higher-order interactions between

Hongzhe Li is a Professor of Biostatistics, Department of Biostatistics and Epidemiology University of Pennsylvania School of Medicine, Philadelphia, PA 19104 (E-mail: *hli@cceb.upenn.edu*). This work was done when he was on the faculty at the University of California—Davis. His research is supported by National Institutes of Health grant R01 ES009911.

Journal of the American Statistical Association
March 2006, Vol. 101, No. 473, Theory and Methods
DOI 10.1198/016214505000000853

the SNPs. Recently there have been some new developments in modeling more complex interactions among the SNPs than what the haplotypes can model. These include the logic regression models (Ruczinski, Kooperberg, and LeBlanc 2003; Kooperberg and Ruczinski 2004), where logic combinations of the binary variables coded for the SNPs are treated as predictors. Huang et al. (2004) developed a tree-structured supervised learning methods, called FlexTree, for studying gene–gene and gene–environment interactions in which they considered both an additive score model involving many genes and a model in which only a precise list of aberrant genotypes is predisposing. These models hold great promise in modeling complex interactions among the gene variants. Alternatively, one can model the genotype effects over many SNPs and use the recently developed methods in the analysis of high-dimensional genomic data in selecting relevant SNPs and their interactions (Efron, Hastie, Johnstone, and Tibshirani 2004; Gui and Li 2005a,b; Li and Luan 2005). Among these methods, the threshold gradient descent procedure (Friedman and Popescu 2004; Gui and Li 2005a) can be used to implicitly model the linkage disequilibrium between the SNPs and the boosting procedure with regression trees as a base learner (Friedman 2001; Li and Luan 2005) provides an interesting way to model complex interactions among the SNPs. Which of these models is the best for identifying the gene variants related to complex traits remains to be seen.

4. INCORPORATING ADDITIONAL BIOLOGICAL INFORMATION

Although new high-throughput technologies are continuously being developed and elaborated for biomedical research, it is also extremely important to make full use of the data and knowledge accumulated by traditional experiments. This knowledge is often stored in databases as "metadata," usually defined as "data about data." For example, it is now known that many aspects of cancer result from deregulation or disturbance of interacting signaling pathways and regulatory circuits that normally control processes of the cell cycle, apoptosis, cell–cell adhesion, and cytoskeletal organization. Deregulation of these pathways often leads to aberrant signaling and the corresponding cancer hallmarks, such as increased proliferation, decreased apoptosis, genome instability, sustained angiogenesis, tissue invasion, and metastasis (Hanahan and Weinberg 2000). Studies of these deregulated pathways have revealed genomic and biological differences between tumor and normal cells. These observations suggest that genomic data and metadata such as known pathways or networks have great potential in identifying genetic variants and pathways that contribute to the risk of cancers and to the variability in treatment responses among cancer patients. Important metadata that have been widely used in biomedical research include Gene Ontology (GO) (Asburner et al. 2000), the KEGG metabolic pathways database (Kanehisa, Goto, Kawashima, and Nakaya 2002), and BioCarta pathways (*www.biocarta.com*).

One limitation of almost all of the available methods for genetic association analysis of SNP data is that known biological knowledge derived from metadata, such as known biological pathways, is rarely incorporated into the analysis. From a statistical standpoint, doing this may require a whole new set of regression analysis methods, where the levels of activity of several biological networks are treated as predictors. However, such network activities cannot be measured directly, but they may be inferred from complex combinations of genetic variants in genes within the pathways. Such biological knowledge can be used to form more biologically relevant disease-predisposing models, including complex interactions among the genetic variants. In addition, biological information also provides an alternative for mediating the problem of a large number of potential interactions by limiting the analysis to biologically plausible interactions between genes in related pathways. In general, risk interactions are more plausible between genes involved in a physical interaction, found in the same pathways, or involved in the same regulatory network (Carlson, Eberle, Kruglyak, and Nickerson 2004). Incorporating these metadata into current genetic association analysis would appear to hold great promise in large-scale genetic association analysis.

Finally, I agree with LZ that there are many interesting statistical problems related to large-scale genetic association analysis and more efforts from statisticians are needed to develop robust and theoretically sound strategies for identifying genetic contributions to disease and pharmaceutical responses and for identifying gene variants that contribute to good health and resistance to disease.

ADDITIONAL REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000), "Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium," *Nature Genetics*, 25, 25–29.

Carlson, C. S., Eberle, M. A., Kruglyak, L., and Nickerson, D. A. (2004), "Mapping Complex Disease Loci in Whole-Genome Association Studies," *Nature*, 429, 446–452.

Chen, J., and Chatterjee, N. (2005), "Haplotype-Based Associaiton Analysis in Cohort and Nested Case-Control Studies," *Biometrics*, to appear.

Chen, J., Peters, U., Foster, C., and Chatterjee, N. (2004), "A Haplotype-Based Test of Association Using Data From Cohort and Nested Case-Control Epidemiologic Studies," *Human Heredity*, 58, 18–29.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499.

Friedman, J. (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, 29, 1189–1232.

Friedman, J. H., and Popescu, B. E. (2004), "Gradient Directed Regularization," technical report, Stanford University.

Gui, J., and Li, H. (2005a), "Threshold Gradient Descent Method for Censored Data Regression, With Applications in Pharmacogenomics," *Pacific Symposium on Biocomputing*, 10, 272–283.

——— (2005b), "Penalized Cox Regression Analysis in the High-Dimensional and Low-Sample Size Settings, With Applications to Microarray Gene Expression Data," *Bioinformatics*, Advance Access published on April 6, 2005, doi:10.1093/bioinformatics/bti422.

Hanahan, D., and Weinberg, R. A. (2000), "The Hallmarks of Cancer," *Cell*, 100, 57–70.

Huang, J., Lin, A., Narasimhan, B., Quertermous, T., Hsiung, C. A., Ho, L. T., Grove, J. S., Olivier, M., Ranade, K., Risch, N. J., and Olshen, R. A. (2004), "Tree-Structured Supervised Learning and the Genetics of Hypertension," *Proceedings of National Academy of Sciences*, 101, 10529–10534.

Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. (2002), "The KEGG Databases at GenomeNet," *Nucleic Acids Resources*, 30, 42–46.

Kooperberg, C., and Ruczinski, I. (2004), "Identifying Interacting SNPs Using Monte Carlo Logic Regression," *Genetic Epidemiology*, 28, 157–170.

Li, H., and Luan, Y. (2005), "Boosting Proportional Hazards Models Using Smoothing Splines, With Applications to High-Dimensional Microarray Data," *Bioinformatics*, Advance Access published on February 15, 2005, doi:10.1093/bioinformatics/bti324.

Ruczinski, I., Kooperberg, C., and LeBlanc, M. (2003), "Logic Regression," *Journal of Computational and Graphical Statistics*, 12, 475–511.

Tzeng, J. Y. (2005), "Evolutionary-Based Grouping of Haplotypes in Association Analysis," *Genetic Epidemiology*, 28, 220–231.

Rejoinder

D. Y. LIN and D. ZENG

We are grateful to the editor and associate editor for organizing the discussion of our article and to the discussants for their valuable contributions. The discussants are leading researchers in statistical genetics, and we greatly appreciate their expert opinions and insightful comments.

Inference on haplotype effects is a hot topic in genetics. From a statistical standpoint, the problem is very interesting and challenging because of haplotype ambiguity and high-dimensional nuisance parameters. Because the number of haplotypes within a candidate gene can be large and the underlying biology is mostly unknown, modeling the haplotype effects correctly is difficult. The task becomes more daunting when the study involves a large number of SNPs. The comments provided by the discussants underscore the importance of haplotype analysis and reflect the many challenges confronted by data analysts. Indeed, a major motivation for writing our article was to bring these challenges to the attention of the broader statistical community. Here we address the main issues raised by each group of discussants.

1. RESPONSE TO SABATTI

Dr. Sabatti offered an excellent description of the important roles of haplotype-based studies in localizing regions harboring disease susceptibility genes and in identifying the functional variants. She also provided a nice discussion of the potential use of our methods in different types of studies. Although the numerical results in our article pertain to the inference on haplotype effects within a region of interest, the theory developed in the article allows one to perform association mapping as well. As we mention in section 5 of our article, one can scan through the genome with sliding marker windows and perform an overall likelihood-ratio test for association between the disease phenotype and the haplotypes formed by the markers in each window. In genomewide investigations, environmental factors are typically ignored, so the maximization of the likelihood given in (9) of our article is very fast. Rare haplotypes need to be removed or combined in the analysis so as to avoid the sparsity of the data. The effects of multiple comparisons can be adjusted by permuting the data or by applying the Monte Carlo method of Lin (2005). An article on haplotype-based association mapping is currently under preparation.

2. RESPONSE TO SATTEN, ALLEN, AND EPSTEIN

When we started this project 3 years ago, the article by Epstein and Satten (2003) was unpublished. The authors kindly provided us with earlier versions of their article, from which we learned a great deal. In fact, our work on case-control studies with unknown population totals was largely motivated by their work. We have also greatly benefited from personal communications with them.

The first major comment of Drs. Satten, Allen, and Epstein (SAE hereinafter) pertains to the efficiency advantages of the retrospective likelihood analysis over the prospective likelihood analysis for case-control studies. Satten and Epstein (2004) found considerable efficiency differences in estimating main haplotype effects. Our simulation studies revealed that the differences can be even more substantial in estimating haplotype-environment interactions. (The results are not given in the final version of our article.) There is a potential price for the efficiency gains: the retrospective likelihood is less robust against violation of Hardy–Weinberg equilibrium (HWE).

We agree with SAE that the dependence between haplotypes and covariates is a difficult issue and that the independence assumption is reasonable in many cases. For all study designs, the likelihoods involve the conditional distribution of covariates given haplotypes and genotypes, so one must either assume the conditional independence of covariates and haplotypes given genotypes or model the conditional distribution of covariates given haplotypes. Our work accommodates both the situations of conditional independence and unconditional independence. The distinction between the two situations is immaterial to cross-sectional and cohort studies, because the likelihoods for the parameters of interest are the same. For case-control studies, the computations are slightly more demanding under conditional independence than under unconditional independence.

As we mention in section 2.1 of our article, there is much flexibility in specifying the regression effects of haplotypes on the phenotype. With K haplotypes, we can either compare one haplotype with all the others or compare each of the first $(K-1)$ haplotypes with the last haplotype. The numerical results in our article pertain to the first formulation; however, all of the theoretical results, including those on identifiability, and the numerical algorithms apply to the second formulation as well. We agree with SAE that the identifiability results under arbitrary distributions of haplotypes do not guarantee stable estimates in small samples, which is why all of our data analysis relies on model (3).

Our article deals exclusively with studies of unrelated individuals. The methods developed by Allen et al. (2005) for case-parent trio studies are very clever and provide another illustration of the usefulness of semiparametric efficiency theory in statistical genetics. We look forward to learning about the extension of that work to the situations with covariates and to case-control studies, as well as the extension of the work of Epstein and Satten (2003) to matched case-control studies.

D. Y. Lin is Dennis Gillings Distinguished Professor (E-mail: *lin@bios.unc.edu*) and D. Zeng is Assistant Professor (E-mail: *dzeng@bios.unc.edu*), Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599.

Journal of the American Statistical Association
March 2006, Vol. 101, No. 473, Theory and Methods
DOI 10.1198/016214505000000862

3. RESPONSE TO CHATTERJEE, SPINKA, CHEN, AND CARROLL

3.1 Case-Control Studies

This group of discussants has done an impressive amount of work on case-control studies. The original work by Carroll and colleagues (e.g., Roeder et al. 1996) and the recent work by Chatterjee and Carroll (2005) provide fundamental insights into the analysis of case-control data, particularly the relative merits of the retrospective-likelihood versus prospective-likelihood analyses.

As indicated in our response to SAE, it is necessary to assume the conditional independence of haplotypes and covariates given genotypes (or the stronger unconditional independence), unless one is willing and able to model the relationship between haplotypes and covariates. It is straightforward to incorporate a parametric model for the relationship between haplotypes and covariates into our likelihoods. It is also possible to account for latent population substructure with the aid of genomic markers, as we mention in section 5 of our article.

The identifiability results of Spinka, Carroll, and Chatterjee (2005) seem to be related to those presented in sections A.4.1 and A.4.2 of our article. In our view, such identifiability conditions are difficult to verify in practice. Although the intercept term can be identified in some situations, the estimator of this parameter is likely to be unstable in finite samples. Thus, we advocate the methods based on the rare-disease assumption.

We agree with Drs. Chatterjee, Spinka, Chen, and Carroll (CSCC hereinafter) that one can avoid the nonparametric estimation of the covariate distribution whether or not the disease is rare. In fact, the profile likelihood method presented in section A.4.4 of our article does not impose the rare-disease assumption. Expression (5) of CSCC appears to be similar to the second paragraph of our section A.4.5.

Our work covers both the situation of conditional independence between haplotypes and covariates given genotypes and that of unconditional independence under HWE or one-parameter extensions of HWE. We showed that our estimators achieve the semiparametric efficiency bounds. It does not seem possible to obtain more efficient estimators without imposing additional restrictions.

The prospective methods of Spinka et al. (2005) improve on those of Zhao et al. (2003). We agree with CSCC that the prospective methods are more robust than the retrospective methods against violation of HWE and that of the assumption of independence between haplotypes and covariates. In contrast, the prospective methods can be considerably less efficient. The choice between the prospective and retrospective methods would depend on the plausibility of the assumptions.

3.2 Cohort Studies

The work by Chen and colleagues on cohort studies is a novel application of the results of Prentice (1982) for the proportional hazards model with covariate measurement error. Because the relative-risk function shown in (9) of CSCC involves the baseline hazard function, it is very challenging, both computationally and theoretically, to make inference about the relative-risk parameters on the basis of (9). For rare diseases, the relative risk function is free of the baseline hazard function, so that the partial likelihood principle can be applied to both cohort studies and nested case-control studies given consistent estimators of the haplotype frequencies. The bias and efficiency of the resultant relative risk estimators require further investigation.

The NPMLE is very easy to calculate for the proportional hazards model. The EM algorithm guarantees an increase in the likelihood function at each step of the iteration and facilitates calculation of the information matrix. In the M-step of the EM algorithm, estimation of the relative risk parameters and the cumulative baseline hazard function is very similar to calculation of the maximum partial likelihood estimator and the Breslow estimator. We have not encountered any convergence problem of the EM algorithm in our extensive simulation studies.

A major advantage of the NPMLE approach is that it can be applied to the entire class of transformation models, not just the proportional hazards model. In addition, this approach is applicable to case-cohort and nested case-control studies whether environmental factors are measured for the selected sample or the entire cohort and the estimators continue to be asymptotically efficient. A manuscript on the NPMLEs for case-cohort and nested case-control designs is currently under review.

4. RESPONSE TO TZENG AND ROEDER

HWE requires stringent conditions (i.e., random mating, no viability and/or fertility differential of alleles, no immigration or emigration, no mutation, and infinite population size), and is unlikely to hold in human populations. We agree with Drs. Tzeng and Roeder (TR hereinafter) that population substructure is the primary source of departure from HWE (assuming that laboratory errors have been corrected). Population substructure will cause spurious associations if and only if the risks of disease vary among subpopulations and the substructure is not accounted for in the analysis. Our work contains HWE as a special case and enables one to test this assumption. We wish to accommodate Hardy–Weinberg disequilibrium because the analysis of case-control data is highly sensitive to violation of HWE, and using model (3) greatly reduces the bias even when the disequilibrium does not conform to (3) (see Satten and Epstein 2004). Allowing Hardy–Weinberg disequilibrium adds very little complexity to our algorithms.

Although genomewide association studies are on the horizon, most genetic epidemiology studies are concerned with candidate genes. In those studies, epidemiologists are often interested in both testing and estimation. Because our methods are likelihood-based, parameter estimation and hypothesis testing are encompassed within a common framework. We share TR's view that testing for the presence of a genetic effect is more realistic than refined estimates of the effect size. Although our numerical illustrations focus on models that specify a disease-susceptibility haplotype, our theory and numerical algorithms allow other types of models. Extension to genomewide association mapping is also possible, as we mentioned earlier in our response to Dr. Sabatti.

The relative efficiency of haplotype analysis versus single marker analysis depends on a number of factors, including the nature of the SNP–disease association, number and positions of disease-causing SNPs, extent and strength of linkage disequilibrium, and selection of markers. Haplotype analysis is likely to be more powerful than single marker analysis if the causal

SNPs are not typed or if there are strong interactions of multiple SNPs on the same chromosome. When there are a large number of haplotypes, it is sensible to group them by the methods of Tzeng (2005) and Seltman et al. (2001, 2003). The haplotype-sharing statistic developed by Tzeng et al. (2003) is a useful approach for testing the presence of a genetic effect.

As with any type of observational study, it is highly challenging to come up with the correct or even an approximately correct model. The small simulation study conducted by TR demonstrated the potential bias of parameter estimators under model misspecification and ascertainment bias. Because there is no bias when the true parameter value is 0, hypothesis testing still may be appropriate. We certainly agree with TR that refined models will become more useful as the data improve and become more focused.

5. RESPONSE TO LI

As mentioned in our response to CSCC, we have already developed the NPMLEs for the class of semiparametric transformation models under the case-cohort and nested case-control designs. The EM algorithm indeed can be used to estimate all of the parameters simultaneously. The estimators again achieve the semiparametric efficiency bounds. Our simulation studies showed that the case-cohort and nested case-control designs are highly cost-effective.

Dr. Li described a number of methods for parameterizing genetic variants. He also suggested incorporating known biological information into the modeling and analysis. Exploring these interesting ideas will entail considerable statistical innovation and yield useful new methods for genetic association analysis.