Analysis of Case-Control Studies of Genetic and Environmental Factors With Missing Genetic Information and Haplotype-Phase Ambiguity

Christine Spinka,1 Raymond J. Carroll2 and Nilanjan Chatterjee3*

1Department of Statistics, University of Missouri, Columbia, Missouri
2Department of Statistics, Texas A&M University, College Station, Texas
3Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland

Case-control studies of unrelated subjects are now widely used to study the role of genetic susceptibility and gene-environment interactions in the etiology of complex diseases. Exploiting an assumption of gene-environment independence, and treating the distribution of environmental exposures as completely nonparametric, Chatterjee and Carroll [2005] (Biometrika 92:399–418) recently developed an efficient retrospective maximum-likelihood method for analysis of case-control studies. In this article, we develop an extension of the retrospective maximum-likelihood approach to studies where genetic information may be missing on some study subjects. In particular, special emphasis is given to haplotype-based studies where missing data arise due to linkage-phase ambiguity of genotype data. We use a profile likelihood technique and an appropriate expectation-maximization (EM) algorithm to derive a relatively simple procedure for parameter estimation, with or without a rare disease assumption, and possibly incorporating information on the marginal probability of the disease for the underlying population. We also describe two alternative robust approaches that are less sensitive to the underlying gene-environment independence and Hardy-Weinberg-equlibrium assumptions. The performance of the proposed methods is studied using simulation studies in the context of haplotype-based studies of gene-environment interactions. An application of the proposed method is illustrated using a case-control study of ovarian cancer designed to investigate the interaction between BRCA1/2 mutations and reproductive risk factors in the etiology of ovarian cancer.


Key words: case-control studies; gene-environment interactions; EM algorithm; haplotype; semiparametric methods

INTRODUCTION

Risks of complex diseases such as cancers are determined by both genetic and environmental factors. Advances in human genome research have thus led to epidemiologic investigations not only of the effects of genes alone, but also of their effects in combination with environmental exposures. The case-control study design, which has been widely used in classical questionnaire-based epidemiologic studies, is being increasingly used to study the role of genes and gene-environment interactions in the etiology of complex diseases.

The traditional approach for analysis of case-control studies is prospective logistic regression. Here the basis of inference is formed by the likelihood of the disease (D) outcome data conditional on covariate information (X), ignoring the fact that under the case-control sampling design, data are observed on X conditional on D. Andersen [1970] and Prentice and Pyke [1979] showed that such a prospective approach is actually equivalent to the retrospective maximum likelihood analysis that properly accounts for the case-control sampling design, provided the distribution of covariates is treated completely nonparametrically. Roeder et al. [1996] generalized
these results to show that even in the presence of covariate missing data and/or measurement error, the prospective and retrospective maximum-likelihood methods for analyzing case-control studies are equivalent, as long as the underlying model for the covariate distribution is nonparametric.

In studies of genetic epidemiology, it often may be reasonable to assume certain parametric or semiparametric models for the covariate distribution in the underlying source population. For example, if $G$ represents one of the three possible genotypes a subject can have at a particular biallelic locus, the population frequencies of the three genotypes could be specified in terms of the frequency of one of the alleles under the Hardy-Weinberg equilibrium (HWE) assumption. Another assumption that is commonly invoked in practice is that genetic susceptibility and environmental exposures are independently distributed in the population. The prospective logistic regression analysis, being the semiparametric maximum likelihood solution for the problem that allows an arbitrary covariate distribution, clearly remains a valid option for analyzing case-control studies in such a setting. However, retrospective methods that can exploit these various covariate distributional assumptions can be more efficient [Epstein and Satten, 2003; Satten and Epstein, 2004; Chatterjee and Carroll, 2005].

Chatterjee and Carroll [2005] developed a retrospective maximum-likelihood approach for analysis of case-control studies exploiting the gene-environment independence and possibility the HWE assumption. In this article, we extend this approach for dealing with missing data on genetic risk factors ($G$). Missing data on genetic factors could arise due to incomplete genotyping information. Moreover, in haplotype-based studies, where the effect of a gene is studied in terms of “haplotypes” (the combination of alleles at multiple loci along individual chromosomes), missing data arise due to the intrinsic “phase ambiguity” of locus-specific genotype data. For example, if $A/a$ and $B/b$ denote the major/minor alleles in two biallelic loci, then subjects with genotypes (Aa) and (Bb) at the first and the second locus, respectively, are considered “phase ambiguous;” their genotypes could arise from either haplotype-pair $(A-B, a-b)$ or haplotype-pair $(A-b, a-B)$.

As 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. Zhao et al. [2003] described an estimating-equation approach where the logistic regression parameters are estimated based on score equations derived from a prospective likelihood of the disease outcome data, given covariates. The estimates of haplotype frequencies, which are required for evaluation of the prospective score equations, were proposed to be estimated using an expectation-maximization (EM) algorithm [Excoffier and Slatkin, 1995] applied to the genotype data of the controls. Lake et al. [2003] described a similar prospective approach, except that they proposed estimating haplotype frequencies jointly with regression parameters from the prospective likelihood itself. Incorporation of environmental factors is straightforward in these approaches under the assumption of gene-environment independence.

Epstein and Satten [2003] described an alternative approach for haplotype-based analysis of case-control studies that jointly estimates the regression parameters and haplotype frequencies by maximizing the proper retrospective likelihood of data under the case-control sampling design. The authors observed that the retrospective likelihood approach yielded more efficient estimates of regression parameters than the previously proposed prospective methods, a consequence of the fact that the retrospective approach fully exploited the HWE assumption for the underlying population. Incorporation of environmental factors, however, is complicated in this approach, because the retrospective likelihood involves potentially high-dimensional nuisance parameters that specify the distribution of the environmental factors in the underlying population. Stram et al. [2003] described yet another approach based on the joint likelihood of disease and genotype data, after accounting for the ascertainment scheme that cases and controls are selected with differential probabilities from the underlying population. We will show later that an extension of this ascertainment-corrected joint-likelihood method, which can incorporate environmental covariates, is equivalent to the retrospective-maximum likelihood method we propose in this article.

In this article, we extend the profile likelihood approach of Chatterjee and Carroll [2005] to develop a relatively simple algorithm for obtaining the efficient retrospective maximum-likelihood estimator for case-control studies that can incorporate both genetic and environmental factors and can account for the presence of missing data in the genetic factors. We first describe the key results for derivation of the profile likelihood...
and related asymptotic theory in a general missing-data setting. We then describe a representation of the profile likelihood that links the retrospective maximum-likelihood procedure to the ascertainment-corrected joint-likelihood approach of Stram et al. [2003]. Then we describe a computational algorithm for implementation of the profile likelihood method in the context of haplotype-based gene-environment interaction studies. Further simplification of the proposed methodologies under the rare disease assumption is also described.

Afterwards, we describe an extension of the methods to account for possible correlation between genetic and environmental factors that may arise due to their dependence on other common factors, such as ethnicity. We next describe a modified prospective estimating equation approach that is fairly robust to violation of gene-environment independence and HWE assumptions. We discuss how this latter approach contrasts with some of the recently proposed “prospective” methods that could be inconsistent under the case-control design, even if the true haplotype frequencies were known and the model assumptions were valid. We then study the performance of the proposed estimators, using simulated data in the context of gene-environment interaction studies involving haplotypes. Finally, we illustrate the application of the proposed method based on a case-control study of ovarian cancer aiming to ascertain the interaction of reproductive risk factors and BRCA1/2 mutation.

**METHODS: THE GENERAL SETTING**

**NOTATIONS AND MODEL ASSUMPTIONS**

Let $D$ be the binary indicator of the presence, $D = 1$, or the absence, $D = 0$, of a disease. Suppose the prospective risk model for the disease given a subject’s genetic covariate of interest, $H$, and environmental risk factors, $X$, is given by the logistic regression model

$$
\text{pr}(D = 1|H, X) = \mathcal{L} \left( \beta_0 + m(H, X; \beta_1) \right),
$$

where $\mathcal{L}(u) = \frac{1}{\left(1 + \exp(-u)\right)}$ is the logistic distribution function, and $m(\cdot)$ is a known but arbitrary function. We assume $H$ and $X$ are independently distributed in the underlying population, and their joint distribution is given by the product form $V(H, X) = Q(H)F(X)$, where $Q$ and $F$ are the marginal distribution functions of $H$ and $X$, respectively. We assume $H$ is discrete with

$$
\text{pr}(H = h_j) = q(h_j; \theta),
$$

where $q(\cdot)$ is a known function and $\theta$ is a vector of parameters. The environmental covariates $X$ can be of arbitrary type, possibly including both continuous and discrete components. The corresponding distribution $F(x)$ is left completely unspecified.

Suppose that the true genetic covariate of interest, $H$, may not be always directly observed. Let $G$ denote all the genetic information for a subject that is directly observed. We assume that $G$ is independent of $(D, X)$ given $H$, i.e., $G$ does not contain any additional information on $D$ and $X$ given $H$. Let $\Delta$ be a variable whose values indicate what sort of genetic information is measured in $G$. For example, in a haplotype-based study, we could have

$$
\Delta = \begin{cases} 
1 & \text{if no genetic information is measured;} \\
2 & \text{if unphased genotypes are measured;} \\
3 & \text{if phased haplotypes are measured.}
\end{cases}
$$

Suppose that $N_0$ controls and $N_1$ cases are sampled from the conditional distributions $\text{pr}(\Delta, G, X|D=1)$ and $\text{pr}(\Delta, G, X|D=0)$, respectively, and let $(G_i, X_i)_{i=1}^{N_0+N_1}$ denote the corresponding covariate data of the $N_0+N_1$ study subjects. We assume that $\text{pr}(\Delta|D, X, H) = \text{pr}(\Delta|D, X, G)$, i.e., the type of genetic information measured does not depend on the individual’s true genetic covariate $(H)$, given disease status $(D)$, environmental covariates $(X)$, and the measured genetic information $(G)$.

Define $H$ to be the set of all possible values of $H$, and $H_G = \{h_j : h_j \text{ is consistent with } G\}$ to be the set of all possible values of $H$ that are consistent with the observable genetic information $G$. Then,

$$
\text{pr}(D|X, G) = \sum_{h \in H_G} \text{pr}(D|X, H = h, G)\text{pr}(H = h|X, G)
\begin{align*}
&= \sum_{h \in H_G} \text{pr}(D|X, H = h)\text{pr}(H = h|G) \\
&= \sum_{h \in H_G} \frac{\text{pr}(D|X, H = h)q(h; \theta)}{\sum_{h' \in H_G} q(h'; \theta)}.
\end{align*}
$$

The log-likelihood of the data under the case-control sampling scheme assuming the above
model is given by

\[
L = \sum_{i=1}^{N_0+N_1} \log\{\text{pr}(G_i, X_i | D_i)\} \\
= \sum_{i=1}^{N_0+N_1} \log\{\text{pr}(D_i | G_i, X_i) \text{pr}(G_i) \text{pr}(X_i) / \text{pr}(D_i)\}
\]

where

\[
\text{pr}(D_i) = \int \sum_{x \in H} \text{pr}(D_i | X = x, H = h) \\
\times \text{pr}(H = h) dF(x).
\]

**IDENTIFIABILITY**

In a nonparametric setting, where no assumption is made about the form of the covariate distribution \(V(h, x)\), it is well-known that neither \(V(\cdot)\) nor the intercept parameter \(\beta_0\) is identifiable from case-control data [Prentice and Pyke, 1979]. In contrast, under the assumption of gene-environment independence, Chatterjee and Carroll [2005] noted that except for some boundary situations, the intercept parameter \(\beta_0\) and the covariate distribution \(V(\cdot)\) are identifiable from the retrospective case-control likelihood. In general, the identifiability of \(\beta_0\) is intrinsically related to the class of \(V(\cdot)\) that is under consideration.

In the presence of missing data on \(H\), the identifiability of the parameter estimates also depends on the nature of missing data and the form of the functions \(m(H, X, \beta_1)\) and \(q(H, \theta)\). In haplotype-based studies, for example, where \(H\) reflects the pair of haplotypes (diplotypes) a subject carries in two homologous chromosomes, certain diplotypes may never be directly observed from the unphased genotype data. In such a situation, identifiability of parameter estimates requires specifying the distribution \(q(H, \theta)\), using the HWE assumption (see Haplotype-Based Gene-Environment Studies, below) and restricting the model \(m(H, E, \beta_1)\) so that it does not involve interactions between pairs of haplotypes which are never directly observed together. For subsequent calculations, we will assume that, depending on the missing data structure of \(H\), the models \(q(H; \theta)\) and \(m(H, X, \beta_1)\) are chosen in such a way that all of the parameters \(\beta_0\), \(\beta_1\), and \(\theta\) and the nonparametric distribution function \(F(x)\) are identifiable from prospective studies. In what follows, we state easily verifiable conditions for identifiability of parameters of a prospective model from retrospective studies.

We will assume \(X\) to be discrete, with \(K\) possible values. Although the results we state below can be expected to hold for a continuous \(X\), a rigorous proof would require a more sophisticated argument. Let \(q(g)\) and \(f(x)\) denote the marginal probability mass functions for \(G\) and \(X\) in the underlying population. Further define

\[
\phi(G, X) = \log \frac{\text{pr}(D = 1 | G, X) \text{pr}(D = 0 | G_0, X_0)}{\text{pr}(D = 0 | G, X) \text{pr}(D = 1 | G_0, X_0)}
\]

to be the log odds-ratio of the disease associated with the joint exposure \((G, X)\) in reference to a chosen baseline value \((G_0, X_0)\), and let \(\alpha = \logit\{\text{pr}(D = 1 | G_0, X_0)\}\), so that \(\alpha\) is the corresponding baseline odds of the disease. With a slight abuse of notation, let \(f, q, \phi\) denote the vectors that contain the values of \(f(X), q(G), \phi(G, X)\), respectively, for distinct values of \(X\) and \(G\). We note that the parameter vector \(\theta = (\alpha, \phi^T, q^T, f^T)^T\) completely characterizes the joint distribution \(\text{pr}(D, G, X)\). It is clear that \(\phi\) is identifiable from retrospective studies, because prospective and retrospective odds-ratios are equivalent. In the following Lemma, we state conditions under which the other components of \(\theta\) are identifiable from retrospective studies.

**Lemma 1.** Define \(\alpha^* = \alpha + \log\{\text{pr}_g(D = 0) / \text{pr}_g(D = 1)\}\). Let \(B_0 \subset B\) be the subspace for the parameter vector \(\theta\) that satisfies the constraint

\[
V^*(G, X) \equiv \left[ 1 + \exp\{\alpha^* + \phi(G, X)\} \right] \left[ 1 + \exp\{\alpha + \phi(G, X)\} \right]^{-1} q(G)f(X) = q^*(G)f^*(x)
\]

for some probability mass functions \(q^*(G)\) and \(f^*(X)\). Then, for all \(\theta \notin B_0\),

\[
\text{pr}_g(X = x, G = g | D = d) = \text{pr}_g(E = e, G = g | D = d)
\]

if and only if \(\theta = \theta^*\). Moreover, if the models \(q(H; \theta)\) and \(m(H, X, \beta_1)\) are chosen in such a way that \(\gamma = (\beta_0, \beta_1, \theta)\) is uniquely identifiable from the prospective likelihood \(\text{pr}(D, G, X)\), then for all
Lemma 1 first ensures the conditions under which the joint distribution \( \Pr(D, G, X) \) of the observable variables \( (D, G, X) \) can be nonparametrically identified from retrospective studies. Further, it states the condition under which the parameters \( \beta_0, \beta_1, \) and \( \theta \) that characterize the joint distribution \( \Pr(D, H, X) \) involving the potentially unobservable variable \( H \), can be identified from retrospective studies. The proof of our Lemma 1 follows from the Lemma 1 of Roeder et al. [1996], which states that \( V^*(G, X) \) is the only distribution of \( (G, X) \) that can yield the same value of retrospective likelihood as the true distribution \( V(G, X) = q(G)f(X) \). Now, for \( \delta \notin B_0 \) \( V^*(G, X) \) lies outside the model space under the consideration that assumes \( G \) and \( X \) are independent. Thus, for \( \delta \notin B_0 \) the retrospective-likelihood uniquely identifies the joint distribution \( V(G, X) \), which together with the odds-ratio parameters \( \phi(G, X) \) further identifies the intercept parameter \( \alpha \).

Consider the hypothetical population \( \mathcal{P}^* \) that could be obtained by sampling each subject from the original population \( \mathcal{P} \) according to a Bernoulli sampling, with the selection probability for cases and controls being proportional to \( N_1/\Pr(D = 1) \) and \( N_0/\Pr(D = 1) \). A case-control sample from population \( \mathcal{P} \) can be viewed as a random sample from population \( \mathcal{P}^* \). Moreover, with some algebra, it can be seen that \( V^*(G, X) \) represents the distribution of \( (G, X) \) for the selected population \( \mathcal{P}^* \). Thus the constraint (2) can be checked in the data by testing for the independence of \( G \) and \( X \) in the combined case-control sample. The boundary condition (2) implies that if \( G \) and \( X \) are assumed to be independently distributed in the underlying population, then the departure of the distribution of \( (G, X) \) in the case-control sample from independence is informative for the estimation of \( V(G, X) \) and \( \alpha \). Similarly, if certain parametric models, such as HWE, are assumed to hold for \( q(G) \) in the underlying population, then the departure of the distribution of \( G \) in the case-control sample from the assumed parametric models is informative for the estimation of \( q(G) \) and \( \alpha \).

### Estimation

Now we consider maximization of \( L \) with respect to the underlying parameters of the model, \( \beta_0, \beta_1, \) and \( \theta \), and the nonparametric distribution function \( F(x) \). We consider the restricted nonparametric maximum likelihood estimator of \( F \) that allows positive masses only within the set \( X = \{x_1, \ldots, x_K\} \) that represents the unique values of \( X \) that are observed in the case-control sample of \( N = N_0 + N_1 \) study subjects. Thus, for obtaining the maximum likelihood estimator, it is sufficient to consider the class of discrete \( F \) that has support points within the set \( X \). Any \( F \) in this class can be parameterized with respect to the probability masses \( \{\delta_1, \ldots, \delta_K\} \) that assigns to the points \( \{x_1, \ldots, x_K\} \).

Since the dimension of \( \delta \) could easily becomes very large when \( X \) consists of multiple covariates, possibly including continuous ones, direct maximization of the log-likelihood with respect to \( (\beta_0, \beta_1, \theta, \delta) \) may be complex or even infeasible. Following Chatterjee and Carroll [2005] we consider deriving the profile likelihood for the lower-dimensional parameters \( \gamma = (\beta_0, \beta_1, \theta) \) by maximizing the likelihood with respect to \( \delta \) for fixed values of \( \gamma \). The result in the following Lemma shows that the profile likelihood \( L(\gamma, \delta(\gamma)) \) can be obtained in a closed form up to only one additional parameter \( \kappa \), and thus numeric maximization of the likelihood \( L(\gamma, \delta) \) with respect to the potentially high-dimensional nuisance parameter \( \delta \) can be avoided.

**Lemma 2.** Let \( \kappa = \beta_0 + \log\{N_1/\Pr(D = 0)\}/\{N_0/\Pr(D = 1)\} \), \( \Omega = (\gamma, \kappa) \), and \( S(D, X, H, \Omega) = q(H, \theta)\exp[D(\kappa + m(X, H, \beta_1))]/\{1 + \exp[\beta_0 + m(X, m(X, H, \beta_1))]\} \). The profile log likelihood \( L(\gamma, \delta(\gamma)) \) can be computed as \( L^*(\gamma, \delta(\gamma)) \) where

\[
L^*(\gamma, \kappa) = \sum_{i=1}^{N} \left[ \log \left( \sum_{h_j \in H_{i,j}} S(D_i, X_i, h_j, \Omega) \right) - \log \left( \sum_{d=0}^{1} \sum_{h_j \in H} S(d, X_i, h_j, \Omega) \right) \right]
\]

and \( \delta(\gamma) \) is defined by the solution of the equation \( \partial L^*(\kappa, \gamma)/\partial \kappa = 0 \) for fixed \( \gamma \).

The proof of the Lemma is given in the Appendix.

In the above approach, for rare diseases, the estimate of the parameter \( \beta_0 \) itself can be expected to be imprecise because of the intrinsic nonformativeness of the retrospective likelihood.
Much more precise estimation of $\beta_0$ is possible when the marginal probability of the disease, $\Pr(D = 1) = p_1$, for the underlying population is known, which is often the case for case-control studies conducted within a well-defined population or an established cohort. In this case, we observe that $\kappa$ and $\beta_0$ are uniquely determined from each other, based on the formula

$$\kappa = \beta_0 + \log \frac{N_1}{N_0} - \log \frac{p_1}{1 - p_1}. \quad (4)$$

Thus the profile-likelihood can be defined in terms of the reduced set of parameters $\Omega = (\beta_0, \beta_1, \theta)$. Hereafter, we will use the generic notation $\Omega$ so that our results are valid for both the cases of $\Pr(D = 1)$ being known and $\Pr(D = 1)$ being unknown, with the convention that $\Omega = (\beta_0, \beta_1, \theta)$ in the former case and $\Omega = (\beta_0, \kappa, \beta_1, \theta)$ in the latter case.

The score function is given by $\partial L^\ast(\Omega)/\partial \Omega = \sum_{i=1}^{N} \Psi(D_i, X_i, G_i, \Omega)$ where

$$\Psi(D_i, X_i, G_i, \Omega) = \frac{\sum_{h \in H_{1}} S_{\Omega}(D_i, X_i, h, \Omega)}{\sum_{h \in H_{1}} S(D_i, X_i, h, \Omega)} - \frac{\sum_{d=0,1} \sum_{h \in H} S_{\Omega}(d, X_i, h_i, \Omega)}{\sum_{d=0,1} \sum_{h \in H} S(d, X_i, h_j, \Omega)}$$

and $S_{\Omega}(D_i, X_i, h, \Omega) = \delta S_{\Omega}(D_i, X_i, h, \Omega)/\partial \Omega$. Further define $I = -N^{-1}E\{\partial^2 L^\ast(\Omega)/\partial \Omega^2\}$, with the expectation being taken under the case-control sampling design. Let

$$\Lambda = \sum_{d=0}^{1} \{N_d/N\} E\{\Psi(\Delta, D, X, G, \Omega)|D = d\} \times \{E\{\Psi(\Delta, D, X, G, \Omega)|D = d\}\}^T.$$

In the following Lemma, we state the main asymptotic result, which in turn is used to obtain estimates of the asymptotic variance-covariance matrix of the parameter estimates.

**Theorem 1.** Under suitable regularity conditions, the following results hold:

(i) The estimating equations $\partial L^\ast(\Omega)/\partial \Omega = \sum_{i=1}^{N} \Psi(\Delta, D_i, X_i, G_i, \Omega) = 0$ have a unique, consistent sequence of solutions, $\{\Omega_1^N\}_{N \geq 1}$;

(ii) Moreover, $N^{1/2} \left( \Omega_N^N - \Omega_0 \right) \rightarrow \text{Normal}(0, \Sigma)$ in distribution, with $\Sigma = I^{-1} - I^{-1} \Lambda I^{-1}$.

**HAPLOTYPE-BASED GENE ENVIRONMENT STUDIES**

**BACKGROUND, NOTATION, AND MODEL**

For haplotype-based studies, the underlying genetic factor ($H$) of interest for a subject is defined by “diplotypes,” i.e., the two haplotypes the individual carries in his/her pair of homologous chromosomes, where each “haplotype” is the combination of alleles at the loci of interest along an individual chromosome within the genomic region of interest. We denote the diplotype data for a subject by $H^d = (H_1, H_2)$, where $H_1$ and $H_2$ denote the constituent haplotypes. The diplotype data, however, are not directly observable using standard polymerase chain reaction (PCR) methods. Instead, for each subject, the multilocus genotype data $G$ are observed, which contain information on the pair of alleles the subject carries on the pair of homologous chromosomes at each locus, but does not provide the “phase information,” i.e., which combination of alleles appears along each of the individual chromosomes. Thus, the same genotype data $G$ could be consistent with multiple diplotypes. We will denote $H^d_G$ to be the set of all possible diplotypes that are consistent with the genotype data $G$. We observe that for subjects who carry two copies of the same allele (homozygous genotype) at all loci or all but one locus, the diplotype information is uniquely identifiable. It is for subjects who are heterozygous at two or more loci that the phase remains ambiguous.

Given the diplotype data $H^d$ and environmental covariate $X$, we assume that the risk of the disease is given by the logistic regression model

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

Often, one imposes structural assumptions on the risk associated with $H^d$ by modeling its effect through the constituent haplotypes according to a dominant, additive, or recessive model [Wallenstein et al., 1998]. Such modeling may be necessary due to identifiability considerations [Epstein and Satten, 2003]. Such modeling may also be desirable when the effects of the haplotypes themselves are of direct scientific interest. For example, a logistic regression model which assumes an additive effect for each copy of a haplotype (additive model) corresponds to

$$m(H^d = (h_1, h_2), X; \beta_1) = \beta_1 X + \beta_{h_1} + \beta_{h_2} + \beta_{h_1}X + \beta_{h_2}X.$$
where $\beta_X$ is the main effect of $X$, $\beta_{h_i}$ is the main effects of haplotypes $h_i$, $i = 1, 2$ and $\beta_{h_i,X}$ is the interaction effect of $X$ with haplotype $h_i$, $i = 1, 2$.

We assume that $H^d$ is independent of $X$ in the population. Moreover, we assume that the distribution of diplotypes is specified by the HWE

$$\Pr_0\{H^d = (H_i, H_j)\} = \theta_i^2 \quad \text{if} \quad H_i = H_j$$
$$= 2\theta_i \theta_j \quad \text{if} \quad H_i \neq H_j$$

where $\theta_i$ denotes the frequency for haplotype $H_i$.

In the following, we present an alternative representation of $L^*$ that links the retrospective-maximum-likelihood approach to an extension of the approach of Stram et al. [2003], to account for environmental covariates. For algebraic convenience, we now introduce some further notations. Define

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

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/her disease status $D = d$ is proportional to $\mu_d = N_d/\Pr(D = d)$. Let $R = 1$ denote the indicator of whether a subject is selected in the case-control sample under the above Bernoulli sampling scheme. We observe the following probability equalities

$$S(D, H^d, X, \Omega) = \Pr(D|H^d, X, R = 1) \times q(H^d, \theta)r_\Omega(H^d, X),$$

$$\Pr(D = 1|H^d, X, R = 1) = \left[1 + \exp\{-\kappa - m(H^d, X, \beta_1)\}\right]^{-1},$$

$$\Pr(H^d|D, G, X, R = 1) = \frac{\Pr(D|H^d, X, R = 1)\Pr(H^d|X, R = 1)}{\sum_{H^d \in H_0} \Pr(D|H^d = h^d, X, R = 1)\Pr(H^d = h^d|X, R = 1)},$$

and

$$\Pr(H^d|X, R = 1) = \frac{q(H^d, \theta)r_\Omega(H^d, X)}{\sum_{H^d} q(h^d, \theta)r_\Omega(h^d, X)}.$$  

With some algebra, one can now show that the log-profile-likelihood given in Lemma 1 can be expressed in the form

$$L^* = \sum_{i=1}^{N} \log\left( \sum_{h^d \in H_{c_i}} \Pr(D_i|H^d_i = h^d, X_i, R_i = 1) \times \Pr(H^d_i = h^d|X_i, R_i = 1) \right)$$

$$= \sum_{i=1}^{N} \log\left( \sum_{h^d \in H_{c_i}} \Pr(D_i, H^d_i = h^d|X_i, R_i = 1) \right)$$

$$= \sum_{i=1}^{N} \log\{ \Pr(D_i, G_i|X_i, R_i = 1) \}.$$  

When no environmental factors are involved, Stram et al. [2003] proposed analysis of haplotype-based case-control studies using an “ascertainment-corrected joint-likelihood” of the form $\prod_i \Pr(D_i, G_i|R_i = 1)$. The representation of the profile likelihood $L^*$ given in (10) suggests that when $F(x)$ is 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].

Next, we develop an algorithm for estimating $\Omega = (\kappa, \beta_1, \theta)$ using $L^*$, assuming $\Pr(D = 1)$ is known. Then we describe a modification of the methods required when $\Pr(D = 1)$ is unknown.

**ESTIMATION OF HAPLOTYPE FREQUENCIES**

Here, we describe an estimation method for the haplotype-frequency parameters ($\theta$) for fixed $\lambda = (\kappa, \beta)$. Let $N_k(H^d)$ be the number of copies of haplotype $H_k$ contained in the diplotype $H^d$. Note that $N_k(H^d)$ could be 0, 1, or 2. The value of $\theta$ that maximizes $L^*$ with the constraints $\sum_{k=1}^{K} \theta_k = 1$ will satisfy the equation

$$\frac{\partial}{\partial \theta_k} \left\{ L^* + \lambda \sum_{k=1}^{K} \theta_k \right\} = 0.$$

The resulting estimating equation can be shown to be

$$0 = \sum_{i=1}^{N} E_\Omega \left[ \frac{\partial \log\{q(H^d_i, \theta)\}}{\partial \theta_k} \right]_{D_i, G_i, X_i, R_i = 1}$$

$$- \sum_{i=1}^{N} E_\Omega \left[ \frac{\partial \log\{q(H^d_i, \theta)\}}{\partial \theta_k} \right]_{X_i, R_i = 1} + \lambda$$

where $\partial \log\{\Pr(H^d_i|D_i)\}/\partial \theta_k = N_k(H^d_i)/\theta_k$, and the expectations in the first and second terms are taken with respect to the distribution $\Pr(H^d_i|D_i)$,


\( \theta \) under the current parameter estimates. We have shown that the estimating function for the form given by

\[
\sum_{i=1}^{N} E_{\Omega}(N_k(H^d)|D_i, G_i, X_i, R = 1) - \sum_{i=1}^{N} E_{\Omega}(N_k(H^d)|X_i, R = 1).
\]

Now we notice that

\[
\sum_{i=1}^{N} E_{\Omega}(N_k(H^d)|X_i, R = 1) = \theta_0 \sum_{i=1}^{N} \sum_{h^d} 20 h^d_r \{H^d = (h_k, h^d_k), X_i\}
\]

This representation suggests the following iterative approach for solving (12) in terms of \( \theta \):

\[
\theta^{(s+1)}_k = N_k^{(s)} \left\{ \frac{\sum_{i=1}^{N} \sum_{h^d} 20 h^d_r \{H^d = (h_k, h^d_k), X_i\} \theta^{(s)}_0}{\sum_{i=1}^{N} \sum_{h^d} \theta^{(s)}_0 \{H^d = h^d\} r_{\Omega}(h^d, X_i)} \right\}^{-1}
\]

where \( N_k^{(s)} = \sum_{i=1}^{N} E_{\Omega=\{\kappa, \beta^{(s)}_1\}}(N_k(H^d)|D_i, G_i, X_i, R = 1) \) is the expected count for the \( k \)th haplotype under the current parameter estimates. We observe that by definition, \( \theta^{(0)}_0 > 0 \). Further, in each iteration, we will normalize \( \theta^{(s+1)}_k = \theta^{(s+1)}_k / \sum_{k=1}^{K} \theta^{(s+1)}_k \). Thus, we note that the estimate of haplotype-frequencies using formula (13) is given by the expected haplotype-count as a ratio of an “effective sample-size” formula.

**ESTIMATION OF \( \beta_1 \) AND \( \kappa \)**

Define \( \beta^* = (\kappa, \beta_1) \). The estimating equation corresponding to \( \partial L^*/\partial \beta^* = 0 \) can be written in the form \((B1) + (B2) + (B3) = 0\), where

\[
(B1) = \sum_{i=1}^{N} E_{\Omega}(\frac{\partial}{\partial \beta} \log\{p_{\beta}(D_i|H^d, X_i, R = 1)\} | D_i, G_i, X_i, R_i = 1);
\]

\[
(B2) = \sum_{i=1}^{N} E_{\Omega}(\frac{\partial}{\partial \beta} \log\{r_{\Omega}(H^d, X_i)\} | D_i, G_i, X_i, R = 1);
\]

\[
(B3) = - \sum_{i=1}^{N} E_{\Omega}(\frac{\partial}{\partial \beta} \log\{r_{\Omega}(H^d, X_i)\} | X_i, R_i = 1).
\]

Let \( V_{\beta^*}(\Omega) = (B2) + (B3) \). We propose to estimate \( \beta^* \) by iteratively solving

\[
\sum_{i=1}^{N} E_{\Omega^{(t-1)}}(\frac{\partial}{\partial \beta} \log\{pr_{\beta}(D_i|H, X_i, R = 1)\} | D_i, G_i, X_i, R = 1) = -V_{\beta^*}(\Omega^{(t-1)}).
\]

We observe that the estimating equations given in (14) are similar to the corresponding estimating equations for \( \beta_0 \) and \( \beta_1 \) in a logistic regression model in the presence of missing data, except that we are equating them to a nonzero term. Because of the similarity with the parameter estimation in the standard logistic regression model, we can get a fairly stable algorithm for solving these equations.

**Unknown \( pr(D = 1) \)**

We observe that in the calculations given above, the value of \( pr(D = 1) \) is only needed to get an estimate \( \beta_0 \) from the estimate of \( \kappa \). Moreover, the parameter \( \beta_0 \) enters into computations only through the function \( r_{\Omega}(H, X) \). If we assume a rare disease, then we have

\[
r_{\Omega}(H, X) \approx 1 + \exp\{\kappa + m(H, X; \beta_1)\}.
\]

Thus, if one assumes a rare disease, \( \beta_0 \) need not be estimated, and hence \( pr(D = 1) \) need not be known. Under this rare disease approximation, the proposed retrospective maximum-likelihood method reduces to that of Epstein and Satten [2003] in the absence of environmental covariates. If one is not willing to make the rare disease assumption, we propose to estimate \( \theta, \kappa, \) and \( \beta_1 \) by maximizing \( L^* \) for fixed values of \( \beta_0 \) and then do a one-dimensional grid-search to find the estimate of \( \beta_0 \) that maximizes the profile likelihood \( L^*(\beta_0, \beta, \theta(\beta_0)) \). We found that the grid-search method performs very well for unbiased estimation of the odds-ratio parameters \( (\beta_1) \) of interest. The estimates of the intercept parameter \( \beta_0 \), however, are typically imprecise. Gains in precision are possible if one places reasonable bounds on \( pr(D = 1) \).

**POPULATION STRATIFICATION**

Although in many situations genetic susceptibility and environmental exposures are unlikely to be causally related at an individual level, these factors may be correlated at a population
level due to their dependence on other factors. A classic example is "population stratification" due to ethnicity. Allele frequencies for many genes vary widely across different races. Moreover, environmental covariates such as lifestyle or dietary factors also often have different distributions for people of different races. Thus, although genetic and environmental factors may be independently distributed within an ethnic group, there could be a spurious correlation between these factors when ethnicity is ignored.

Here, we will briefly describe how to generalize our methods to handle such "population stratification."

We assume there is a set of cofactors $W$ so that gene-environment independence and HWE hold, conditional on $W$. We consider a polytomous logistic regression model for specifying the haplotype-frequencies, given $W$, as

$$\log \{ p_r(H = h_k | W) / p_r(H = h_0 | W) \} = \log \{ \theta_k(W)/\theta_0(W) \} = \gamma_k + \gamma_k^T W$$

for $k = 1, \ldots, K$, where $h_0$ is a reference haplotype, typically chosen to be the most common haplotype. We further assume HWE conditional on $W$, i.e.,

$$p_r\{H^d = (H_i, H_j) | W\} = \{ \theta_i(W) \}^2 \text{ if } H_i = H_j$$

$$= 2\theta_i(W)\theta_j(W) \text{ if } H_i \neq H_j.$$ 

We also allow $W$ to be potential risk factors for the disease by simply extending the disease-risk model to be

$$\log \{ p_r(D = 1 | H^d, X, W) \} = \beta_0 + m(H^d, X, W; \beta_1).$$

Define

$$r_\alpha(H^d, X, W) = \frac{1 + \exp \{ \alpha + m(H^d, X, W; \beta_1) \}}{1 + \exp \{ \beta_0 + m(H^d, X, W; \beta_1) \}}.$$ 

Following previous arguments, the estimating equation for $\gamma_k$ that corresponds to maximization of the profile likelihood $L^* = \sum_{i} \sum_{H^d \in H_i} \log \{ p_r(D_i, H^d | X_i, W_i, R = 1) \}$ can be shown to be

$$0 = \sum_{i=1}^{N} E_\alpha \left[ \frac{\partial \log \{ p_r(H^d | W) \}}{\partial \gamma_k} | D_i, G_i, X_i, W_i, R = 1 \right]$$

$$- \sum_{i=1}^{N} E_\alpha \left[ \frac{\partial \log \{ p_r(H^d) \}}{\partial \gamma_k} | X_i, W_i, R = 1 \right]$$

(16)

where

$$\frac{\partial \log \{ p_r(H | W) \}}{\partial \gamma_k} = W\{ N_k(H) - 20_k(W) \}. \quad (17)$$

One can get a fairly stable Newton-Raphson or related algorithm for solving (16) by exploiting the generalized linear model (GLM) form of (17). Finally, the updating procedures for $\alpha$ and $\beta_1$ remain the same as before, except that throughout, we condition on $W$.

**ALTERNATIVE ROBUST ESTIMATION OF $\beta$**

Although exploitation of the gene-environment independence and the HWE assumptions can lead to major efficiency gains for analysis of case-control studies, we recommend cautious use of these assumptions, because violation of them can lead to major bias in parameter estimation [Albert et al., 2001; Satten and Epstein, 2004; Chatterjee and Carroll, 2005]. The gene-environment independence assumption, for example, is likely to be satisfied in a wide range of studies involving external environmental agents, exposure to which is not directly controlled by an individual’s own behavior. When an exposure depends on a subject’s individual behavior, on the other hand, the independence assumption could be violated due to direct or indirect association. Family history of a disease, for example, which is associated with genetic risk factors, may influence a subject to change his/her behavior regarding established environmental risk factors, such as smoking for lung cancer. In estimation of $\beta_1$ and $\kappa$, we proposed a possible remedy for accounting for such indirect association between $G$ and $E$ due to their dependence on other common factors $S$. There could also be direct associations. Genetic polymorphisms in the smoking metabolism pathway, for example, may not only modify a subject’s risk from smoking, but may also influence a subject’s degree of addiction to smoking.

When violation of HWE and/or the gene-environment independence assumption seems plausible, it is important to consider alternative methods for analysis of case-control studies that are less sensitive to these assumptions. In the absence of missing data, it is well-known that the standard prospective logistic regression analysis is such an option, because it does not rely on any assumption on covariate distribution. In the presence of missing data, a prospective likelihood-based method that treats the distribution of cofactors as completely nonparametric will also be such a robust option [e.g., Roeder et al., 1996].
For haplotype-based studies, however, a complete nonparametric treatment of the covariate distribution may not be possible because of a lack of parameter identifiability. Nevertheless, when no environmental factors are involved, Satten and Epstein [2004] showed that methods that estimate the regression parameters from the prospective likelihood of data are less sensitive to the violation of the HWE assumption than those based on the true retrospective likelihood. Below, we point out a problem for the use of the prospective estimating equation for analysis of case-control studies, and propose an appropriate remedy.

For fixed values of the haplotype-frequency parameter \( \theta \), the score equations for the regression parameters \( \beta^* = (\kappa, \beta_1) \), corresponding to the prospective likelihood of the data, are given by

\[
0 = \sum_{i=1}^{N} \sum_{h^d \in H_0^d} \frac{\partial}{\partial \beta^0} \log \{ \text{pr}_\beta(D_i|h^d, X_i) \} \frac{\text{pr}_\beta(D_i|h^d, X_i) q(h^d; \theta)}{\sum_{h^d \in H_0^d} \text{pr}_\beta(D_i|h^d, X_i) q(h^d; \theta)}. \tag{18}
\]

We argue in the Appendix that this “purely prospective” score equation (18) is biased under the case-control sampling design due to the underlying covariate distributional assumptions. In other words, even if the true haplotype-frequencies were known and the underlying HWE and gene-environment independence assumptions were valid, the estimator of the regression parameter \( \beta_1 \) based on solving the score equation (18) is not consistent. However, we show that the following simple modification of the prospective score equation is unbiased:

\[
0 = \sum_{i=1}^{N} \sum_{h^d \in H_0^d} \frac{\partial}{\partial \beta^0} \log \{ \text{pr}_\beta(D_i|h^d, X_i) \} \frac{\text{pr}_\beta(D_i|h^d, X_i) r_G(h^d, X_i) q(h^d; \theta)}{\sum_{h^d \in H_0^d} \text{pr}_\beta(D_i|h^d, X_i) r_G(h^d, X_i) q(h^d; \theta)}. \tag{19}
\]

The only structural difference between the two sets of score equations is that (19) is obtained from (18) by replacing \( q(h^d, \theta) \) with \( r_G(h^d, X_i) q(h^d; \theta) \). The unbiasedness of the modified prospective-score equations under the case-control sampling design is shown in the Appendix. We also show that with an appropriate rare disease approximation, the proposed method is equivalent to the estimating equation approach proposed by Zhao et al. [2003]. However, we note that the proof of the asymptotic unbiasedness of the estimating equation approach that is given in Zhao et al. [2003] assumes random sampling of subjects, and does not properly account for the case-control sampling design. Thus our derivation justifies the validity of the procedure of Zhao et al. [2003] under the case-control sampling design. Moreover, it shows how one can avoid the rare disease approximation by using the exact score equation (19) itself.

We observe that evaluation of the score function (19) requires knowing \( \theta \) and \( \beta_0 \). Similar to Satten and Epstein [2004], we propose estimating \( \theta \) for a fixed value of \( \beta^* \) and \( \beta_0 \) by maximization of the retrospective likelihood, the algorithm for which we described in Estimation of Haplotype Frequencies, above. As before, we observe that if \( \text{pr}(D = 1) \) is known, then \( \beta_0 \) could be evaluated as a function of \( \kappa \) using the relationship (4). If \( \text{pr}(D = 1) \) is unknown, one can use the rare disease approximation given in equation (15), so that evaluation of (19) does not require knowing \( \beta_0 \). Alternatively, one can estimate \( \theta \) and \( \beta \) for fixed values of \( \beta_0 \) following the above procedures, and then do a one-dimensional grid-search to estimate \( \beta_0 \) as the maximizer of the profile likelihood \( L^*\{\beta_0, \beta^*(\beta_0), \theta(\beta_0)\} \). Finally, we observe that the functional form of the right-hand side of the score equation (19) is equivalent to that of \( B_1 \), the first of the three terms of the score equations corresponding to the retrospective likelihood that are given in Estimation of \( B_1^* \) and \( \kappa \), above. Thus, the proposed prospective estimation method can be implemented with minimal modification of the algorithm for the retrospective method, and vice versa. A sandwich variance estimator, which properly accounts for the case-control design, can also be easily obtained based on the estimating equation theory. A general formula for the variance estimator is given in the Appendix.

**SIMULATION STUDIES**

\( H \text{ AND } X \text{ ARE INDEPENDENT} \)

In the first set of simulation studies, we examined the performance of the proposed retro-
We analyzed each data set using the retrospective maximum-likelihood method, under the assumption that \(H^d\) and \(X\) are independent in the population, with \(\text{pr}(D = 1)\) being known and unknown, the algorithms for which are described above. In the case of \(\text{pr}(D = 1)\) being unknown, we used the grid-search method for estimation of \(\beta_0\) that we described earlier. Although we know that there are eight true haplotypes in the underlying population, for analysis of each datum, we allowed all possible 32 haplotypes to arise, and let the algorithm estimate the frequencies of each haplotype separately. For estimation of regression parameters \((\beta)\), we pooled three rare haplotypes \(h_6, h_7,\) and \(h_8\) and all of the artificial haplotypes which may appear in the given data to have nonzero, but small, frequencies. The performance of the proposed method for estimation of the haplotype frequency (\(\theta\)) and regression parameters \((\beta)\) is shown in Table II. For convenience of presentation, frequency estimates are shown for the non-null haplotype \(h_3\), which is known to be associated with the disease, for one null “common” \((f = 15\%)\) haplotype, and for one null “rare” \((f = 5\%)\) haplotype. The estimates of regression parameters are shown for the non-null haplotype \((h_3)\), for one “common” haplotype \((h_3)\), and for the pooled category of rare haplotypes.

Using the results shown in Table II, we observe that the proposed method performed very well in estimating both the regression parameters \((\beta)\) and haplotype frequencies \((\theta)\). The proposed standard error estimator also performed very well, and the corresponding 95% confidence intervals had coverage that was very close to their nominal values. Estimates of the interaction parameter \(\beta_{H^dX}\) for the non-null haplotype \(h_3\) were more precise when \(\text{pr}(D = 1)\) was known than when \(\text{pr}(D = 1)\) was unknown.

\(H\) ARE \(X\) ARE INDEPENDENT, GIVEN S

In the second simulation study, we examined the robustness of alternative methods in a scenario where the assumptions of gene-environment independence and HWE hold only within subpopulations. We consider a population composed of two strata, with frequencies \(0.40\) \((S=1)\) and \(0.60\) \((S=2)\), which differed in their distribution of both haplotypes and environmental factors. We assumed a simple scenario involving four haplotypes constructed from two binary SNPs with the haplotypes \{\(0, 0\), \(0, 1\), \(1, 0\), \(1, 1\)\} having frequencies \((0.35, 0.30, 0.15, 0.20)\) and \((0.35, 0.20,\)
TABLE II. Results from 1,000 simulated case-control studies from a population under HWE, with independent distributions for haplotypes (H) and environmental covariates (X)*

<table>
<thead>
<tr>
<th>Pr(D=1)</th>
<th>Parameter</th>
<th>Value</th>
<th>Bias</th>
<th>Observed standard error</th>
<th>Estimated standard error</th>
<th>Coverage probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known</td>
<td>$\beta_X$</td>
<td>0.10</td>
<td>-0.009</td>
<td>0.053</td>
<td>0.054</td>
<td>0.961</td>
</tr>
<tr>
<td></td>
<td>$\beta_H$</td>
<td>0.15</td>
<td>-0.013</td>
<td>0.119</td>
<td>0.122</td>
<td>0.945</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.06</td>
<td>-0.007</td>
<td>0.171</td>
<td>0.172</td>
<td>0.951</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>-0.007</td>
<td>0.147</td>
<td>0.145</td>
<td>0.948</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta_{HX}$</td>
<td>0.20</td>
<td>0.009</td>
<td>0.036</td>
<td>0.037</td>
<td>0.939</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>-0.001</td>
<td>0.049</td>
<td>0.050</td>
<td>0.953</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.001</td>
<td>0.041</td>
<td>0.042</td>
<td>0.964</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.25</td>
<td>0.001</td>
<td>0.009</td>
<td>0.009</td>
<td>0.954</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td>&lt;0.001</td>
<td>0.009</td>
<td>0.009</td>
<td>0.954</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>&lt;0.001</td>
<td>0.004</td>
<td>0.004</td>
<td>0.938</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>$\beta_X$</td>
<td>0.10</td>
<td>-0.006</td>
<td>0.054</td>
<td>0.056</td>
<td>0.961</td>
</tr>
<tr>
<td></td>
<td>$\beta_H$</td>
<td>0.15</td>
<td>-0.008</td>
<td>0.123</td>
<td>0.129</td>
<td>0.961</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.006</td>
<td>0.171</td>
<td>0.172</td>
<td>0.951</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>-0.007</td>
<td>0.147</td>
<td>0.145</td>
<td>0.949</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta_{HX}$</td>
<td>0.20</td>
<td>0.0007</td>
<td>0.040</td>
<td>0.043</td>
<td>0.964</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>-0.001</td>
<td>0.049</td>
<td>0.049</td>
<td>0.953</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.001</td>
<td>0.041</td>
<td>0.042</td>
<td>0.963</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.25</td>
<td>&lt;0.001</td>
<td>0.010</td>
<td>0.011</td>
<td>0.967</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td>&lt;0.001</td>
<td>0.009</td>
<td>0.009</td>
<td>0.956</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>&lt;0.001</td>
<td>0.004</td>
<td>0.004</td>
<td>0.939</td>
<td></td>
</tr>
</tbody>
</table>

*Each replicate contains 1,000 cases and 1,000 controls, and is analyzed using proposed retrospective maximum-likelihood method, assuming HWE and $H \rightarrow X$ independence. Estimates are shown (1) using known probability of disease in population, and (2) estimating probability from data, using grid-search method.

0.30, 0.15) in strata 1 and 2, respectively. We chose the frequencies for the larger stratum (stratum 2) to correspond to the haplotypes defined by the third and fourth SNPs listed in Table 1. The values of $R^2_H$, a popular measure for haplotype-phase uncertainty [Stram et al., 2003], for the haplotypes \{(0, 0), (0, 1), (1, 0), (1, 1)\} were (0.88, 0.87, 0.79, 0.83) for stratum 1 and (0.88, 0.83, 0.87, 0.78) for stratum 2. Thus, in this setting, the degree of phase-uncertainty was modest, but not negligible.

We generated the environmental covariate from a log normal distribution, with the mean and variance for the underlying normal distribution at 0.67 and 1 for stratum 1, and 0 and 1 for stratum 2. Again, we truncated the environmental exposure above at 10 for both strata. Additionally, we assumed that the stratification variable, $S$, is a risk factor for disease. In particular, the disease status for each subject was generated according to the model

$$\logit\{\text{Pr}(D|H^d, X, S)\} = \beta_0 + \beta_X X + \beta_H N_2(H^d) + \beta_{HX} N_2(H^d)X + \beta_S S + \beta_{HS} N_2(H^d)S$$

where $N_2(H^d)$ denotes the number of copies of $h_2 = (0, 1)$ contained in $H^d$, and where the parameters ($\beta_0$, $\beta_X$, $\beta_H$, $\beta_{HX}$, $\beta_S$, $\beta_{HS}$) were chosen to be $(-3.5, 0.1, 0.15, 0.20, 0.69, 1.10)$. For each replicate of our simulation, we first generated data for a large random sample of subjects, which was then treated as the underlying study base for selection of 1,000 cases and 1,000 controls.

During analysis of each data set, as before, we assumed only that the locus-specific genotype data were available, but the phase information was unknown. Each data set was analyzed using a) the retrospective maximum-likelihood method under the assumption that $H^d$ and $(X, S)$ are independently distributed in the population; b) the retrospective maximum-likelihood method under the assumption that $H^d$ and $X$ are independent, conditional on $S$ (see population stratification, above); and c) the modified prospective estimating equation method (see Alternative Robust Estimation of $\beta$, above). We assumed Pr($D=1$) to be known for this set of simulations.

The results, shown in Table III, suggest the following important observations. First, when the true model assumed that $H^d$ and $X$ were independent conditional on $S$, but we analyzed the data as though $H^d$ and $(X, S)$ were independent in the entire population, we induced substantial bias in estimating the parameters $\beta_H$, $\beta_S$, and $\beta_{HS}$. Neither the prospective method nor the
method which accounts for population stratification suffered from such bias. Secondly, the prospective method had the largest variance of the three methods, while the maximum-likelihood method under the unconditional independence assumption had the smallest. The retrospective method which took population stratification into account provided both small bias and relatively small variance. These observations suggest that when gene-environment dependence is suspected, the use of the retrospective maximum-likelihood method under the conditional gene-environment independence model could be optimal, assuming that factors which may induce such dependence are observable. If such factors are not observable, or if direct association between genetic and environmental factors may exist, then the use of the modified prospective method should be considered.

BIAS OF ALTERNATIVE PROSPECTIVE METHODS

Stram et al. [2003] observed that although a naive prospective method which ignores the case-control sampling design may not be strictly correct, the bias in such a method is typically small, unless the predictability of haplotypes given the genotypes, as measured by the $R^2_S$ statistics, is low and the magnitudes of the true risk-parameters are high. We evaluated the bias of alternative prospective methods in a situation where the bias of a naive prospective method was expected to be high. We implemented three procedures: a) the naive prospective method based on cohort likelihood [Lake et al., 2003]; b) the estimating equation approach of Zhao et al. [2003], assuming rare disease; and c) the proposed modified prospective-score-equation approach, assuming $Pr(D=1)$ to be known. We considered a simulation scenario involving three SNPs. To generate a maximal amount of phase ambiguity, we assumed all of the $2^3=8$ haplotypes to be equally likely. We generated diplotypes for each subject under the assumption of Hardy-Weinberg equilibrium. We generated a continuous environmental covariate for each subject, independent of the subject’s diplotype status, using a log-normal distribution, with the mode of the effect being dominant. The true values of parameters for the underlying logistic regression model were $(\beta_0, \beta_X, \beta_H, \beta_{HX}) = (-3.5, 0.1, 0.69, 1.60)$, which corresponded to an overall disease rate of 10.7%. In each replication, we generated data for 1,000 cases and 1,000 controls.

We implemented all three methods to estimate the regression parameters associated with the known “risk haplotype.” From the results shown in Table IV, we observe that while the proposed

| TABLE III. Results from 1,000 simulated case-control studies from population where HWE and independence between haplotypes $(H)$ and environmental covariate $(X)$ holds within strata defined by $S$ |
|---|---|---|---|---|---|
| $\beta_X$ | $\beta_H$ | $\beta_{HX}$ | $\beta_S$ | $\beta_{HS}$ |
| a) Unconditional RML |
| Bias | 0.009 | −0.547 | −0.033 | −0.102 | 0.755 |
| Empirical SE | 0.059 | 0.290 | 0.064 | 0.305 | 0.264 |
| Estimated SE | 0.062 | 0.303 | 0.063 | 0.318 | 0.281 |
| Coverage probability | 0.956 | 0.551 | 0.897 | 0.941 | 0.223 |
| b) Conditional RML |
| Bias | −0.004 | −0.012 | 0.010 | 0.008 | 0.002 |
| Empirical SE | 0.061 | 0.313 | 0.065 | 0.331 | 0.307 |
| Estimated SE | 0.063 | 0.327 | 0.067 | 0.339 | 0.317 |
| Coverage probability | 0.955 | 0.963 | 0.954 | 0.957 | 0.965 |
| c) Modified PSE |
| Bias | 0.001 | −0.022 | −0.003 | 0.010 | 0.029 |
| Empirical SE | 0.070 | 0.326 | 0.075 | 0.337 | 0.302 |
| Estimated SE | 0.069 | 0.350 | 0.076 | 0.349 | 0.330 |
| Coverage probability | 0.942 | 0.966 | 0.947 | 0.960 | 0.972 |

*Each replicate contains 1,000 cases and 1,000 controls, and is analyzed using a) proposed unconditional retrospective maximum-likelihood (RML) method assuming that HWE and $H-X$ independence hold in entire population; b) proposed conditional RML method, assuming that HWE and $H-X$ independence hold conditional on $S$; and c) proposed modified prospective score-equation (PSE) method.

| TABLE IV. Bias and standard errors for regression parameters estimated using three alternative prospective methods |
|---|---|---|
| $\beta_X$ | $\beta_H$ | $\beta_{HX}$ |
| a) Naive prospective |
| Bias | 0.005 | −0.198 | 0.181 |
| Empirical SE | 0.040 | 0.252 | 0.216 |
| b) Zhao et al. [2003] |
| Bias | −0.022 | 0.692 | −0.705 |
| Empirical SE | 0.042 | 0.216 | 0.119 |
| c) Modified PSE |
| Bias | −0.002 | 0.004 | 0.017 |
| Empirical SE | 0.044 | 0.250 | 0.200 |

*Naive prospective method based on cohort likelihood [Lake et al., 2003], b) estimating equation approach of Zhao et al. [2003], assuming rare disease; c) proposed modified prospective-score-equation approach, assuming $Pr(D=1)$ to be known.
modified prospective method was unbiased in estimating all three parameters, both the naive prospective method and the estimating equation approach of Zhao et al. [2003] produced substantial bias for estimation of the parameters $\beta_H$ and $\beta_{HX}$. The large bias in the approach of Zhao et al. [2003] was likely caused by the violation of the underlying rare-disease assumption. In the current simulation setting, although the overall disease rate for the population is low (10.7%), the risk of the disease could become very high for those subjects who carried the risk haplotype and

who also had a high value of environmental exposure. However, it is important to note that the example reflects a fairly extreme scenario involving large amounts of phase ambiguity and strong genetic effects on the risk of disease. In many other examples that involved less extreme parameter settings, the bias for both the naive prospective method and the estimating equation approach of Zhao et al. [2003] was found to be very small or negligible.

**DATA ANALYSIS: ISRAELI OVARIAN CANCER STUDY**

Chatterjee and Carroll [2005] described an application of their proposed methodology on a case-control study of ovarian cancer in Israeli women that was performed to investigate the interaction between the BRCA1/BRCA2 mutations and oral contraceptive use and parity [Modan et al., 2001]. Briefly, this study consisted of all ovarian cancer cases identified in Israel between March 1, 1994–June 30, 1999. For each case, two controls were selected. The selected cases and controls provided blood samples for testing mutations in the BRCA1 and BRCA2 genes. In addition, data were collected on reproductive/gynecological history such as parity, number of years of oral contraceptive use, and gynecological surgery.

Chatterjee and Carroll [2005] restricted their analysis to 832 cases and 747 controls who were genotyped for BRCA1/2 mutations, leaving out 50 cases and 763 controls for whom BRCA1/2 status was missing, but data on all other risk factors were available. We reanalyzed the data using the proposed retrospective maximum-likelihood method, including the subjects with missing genotype information. Similar to Chatterjee and Carroll [2005], we considered the following logistic regression model for risk of ovarian cancer:

\[
\text{logit}\{\Pr(D = 1)\} = \beta_0 + \beta_{BRCA1/2} I(\text{BRCA1/2}) + \beta_{OC} \text{OC} + \beta_{\text{parity}} \text{Parity} \\
+ \beta_{BRCA1/2 \times \text{OC}} I(\text{BRCA1/2}) \times \text{OC} \\
+ \beta_{BRCA1/2 \times \text{parity}} I(\text{BRCA1/2}) \times \text{Parity} + Z^T \gamma
\]

where $I(\text{BRCA1/2})$ denotes the 0–1 indicator of carrying at least one BRCA1/2 mutation, $\text{OC}$ denotes years of oral contraceptive use, $\text{Parity}$ denotes number of children, and $Z$ denotes the set of all cofactors that Modan et al. [2001] used to adjust their regression analysis. Moreover, similar to Chatterjee and Carroll [2005], we assumed the independence between presence of mutation and reproductive risk factors conditional on age, ethnicity, personal history of breast cancer (PHB), and family history of breast and ovarian cancer (FHBO). The genotype frequencies were modeled as a function of these four factors, using the parametric model

\[
\text{logit}\{\Pr(G = 1|S)\} = \theta_0 + \theta_{\text{Age}} I(\text{Age} \geq 50) + \theta_{\text{ASH}} I(\text{Non-Ashkenazi}) \\
+ \theta_{\text{PH}} I(\text{PHB} = 1) + \theta_{1FH} I(\text{FHBO} = 1) + \theta_{2FH} I(\text{FHBO} = 2).
\]

The results of our analysis for the main covariates of interest, i.e., parity, oral contraceptive use, BRCA1/2 mutation, and interactions between the mutations and each of the two reproductive risk factors, are presented in Table V. Compared to the analysis of Chatterjee and Carroll [2005] that included only those individuals with complete genotype information, we observe that there was a important reduction in standard errors for the main effects of the two environmental factors, OC use and parity. This result is intuitive, given that the additional subjects who were incorporated in the new analysis provided data on these two risk factors. In addition, the new analysis confirmed
the original finding of Modan et al. [2001], which suggested an interaction between BRCA1/2 mutation and OC use. In particular, the results suggest that, unlike the situation for noncarriers, the risk of ovarian cancer for carriers did not decrease with increasing oral contraceptive use.

**DISCUSSION**

We developed a method for retrospective maximum-likelihood analysis of case-control studies of genetic and environmental factors that can account for missing genetic information. Particular emphasis was given to haplotype-based studies where missing data arise due to phase ambiguity of available genotype data. By utilizing a profile likelihood of the data under the assumption of gene-environment independence and HWE, we were able to develop a relatively simple computational algorithm for obtaining the estimator. We also showed how this profile likelihood approach established a connection between two seemingly different methods for haplotype-based association analysis of case-control studies: the ascertainment-corrected joint-likelihood approach of Stram et al. [2003], and the retrospective maximum-likelihood approach. Further simplifications of the methodology under a rare disease assumption were also described.

Simulation studies in this article as well as those reported in Epstein and Satten [2003], Satten and Epstein [2004], and Chatterjee and Carroll [2005] show that retrospective methods that can exploit various covariate distributional assumptions, such as HWE and gene-environment independence, can lead to major efficiency gains for analysis of case-control studies. However, caution is needed in the practical use of these methods, because these simulation studies also demonstrate the possibility of major bias in the retrospective methods when the underlying covariate distributional assumptions are violated in truth. In this article, we proposed two alternative methods for relaxing the covariate distributional assumptions.

In the first we proposed explicitly accounting for those factors, such as ethnicity, which could be related to both allele frequencies and environmental factors, possibly inducing an association between these factors in the population. In the second, we proposed a variation of the prospective-estimating equation which we showed to be asymptotically consistent under the retrospective case-control design, assuming that underlying covariate distributional assumptions are valid.

Moreover, in simulation studies, we showed that the method produced very little bias in parameter estimates, even when the covariate distributional assumptions were violated.

A novel finding of our simulation studies, as well as those reported in Chatterjee and Carroll [2005], is that when the gene-environment (G-E) independence assumption holds, incorporation of external information on the marginal probability of disease in the population can lead to further efficiency in the estimation of regression parameters of interest. In traditional logistic regression analysis, knowing the marginal probability of disease allows one to estimate the intercept term of the regression model, but otherwise does not have any effect on the estimation of the other regression parameters of interest. The marginal probability of the disease, possibly stratified by basic demographic factors such as age, sex, and race, is often available or can be estimated precisely in population based case-control studies as well as in case-control studies that are nested within a larger cohort study. The proposed methodology allows incorporation of such additional information into the analysis, and hence can lead to a further precision gain in the estimation of regression parameters under the G-E independence model.

When a study involves a large number of haplotypes, estimation of their frequencies as well as the associated regression parameters could become unstable due to the presence of rare haplotypes. Schaid [2004] gave an excellent review of various currently available techniques for tackling this problem. In principle, in our setting, the parametric model $p_r(H^d)$ can incorporate genetic models based on evolutionary history, thus specifying the haplotype frequencies in terms of a reduced set of genetic parameters. Similarly,
hierarchical modeling techniques can be used to specify regression parameters $\beta$ in terms of a set of lower-dimensional parameters. These and other extensions of the proposed methodology will be pursued in the future.

ACKNOWLEDGMENTS

C.S. was supported by a training grant from the National Cancer Institute (CA-90301). R.C. was supported by a grant from the National Cancer Institute (CA-90301). R.C. was supported by a training grant from the National Institute of Environmental Health Sciences (P30-ES09106).

SOFTWARE

Software implementing the methodology is available upon request from chatterjee@mail.nih.gov.

REFERENCES


PROOF OF LEMMA 2

Recall that $\zeta_m$ is the probability mass function for $X = x_m$, $m = 1, \ldots, K$. For fixed $\gamma = (\beta_0, \beta_1, \theta)$, and except for constants, the log-likelihood function for $\zeta$ has the form

$$
\ell(\zeta|\gamma) = \sum_{i=1}^{N} \log \left( \sum_{m=1}^{K} \zeta_m I(X_i = x_m) \right) - \sum_{i=1}^{N} \log \left\{ \sum_{k} \sum_{h_j \in H} \Pr(D = d_i|X = x_k, H = h_j) q(h_j, \theta) \zeta_k \right\}.
$$

Taking derivatives with respect to each $\zeta_m$ and solving, we find that

$$
\zeta_m = \sum_i I(X_i = x_m) / \sum_{i=1}^{N} \sum_{h_j \in H} \Pr(D = d_i|X = x_k, H = h_j) q(h_j, \theta) \zeta_k.
$$

However, note that $\Pr(D = d) = \sum_k \sum_{h_j \in H} \Pr(D = d|X = x_k, H = h_j) q(h_j, \theta) \zeta_k$, and define $\mu(d) = N_d / \{N \Pr(D = d)\}$. This implies that $\Pr(D = d) = N_d / \{N \mu(d)\}$ and

$$
\zeta_m = \frac{1}{N} \sum_i I(X_i = x_m) / \sum_{h_j \in H} \Pr(D = d_i|X = x_m, H = h_j) \mu(d) q(h_j, \theta).
$$

APPENDIX

Haplotypes and Case-Control Studies 123


It is easily shown that $\sum_{m} \zeta_m = 1$. Substituting, and except for constants, the profile log-likelihood function has the form

$$L\{\gamma, \zeta(\gamma)\} = \sum_{i=1}^{N} \log \{ \sum_{h \in H_{G_i}} \text{pr}(D_i|X_i, h)q(h, \theta) \} + \log \{ \mu(D_i) \}$$

$$- \log \{ \sum_{m} \text{pr}(D = d|X_i, h(j))\mu(d)q(h(j), \theta) \}.$$ 

Now, define $\kappa$ so that $\log \{ \mu(1) \} = \log \{ \mu(0) \} + \kappa - \beta_0$. Then,

$$\ell\{\gamma, \zeta(\gamma)\} = \sum_{i=1}^{N} \log \left[ \sum_{h \in H_{G_i}} \text{pr}(D_i|X_i, h)q(h, \theta) \exp\{D_i(\kappa - \beta_0)\} \right]$$

$$- \sum_{i=1}^{N} \log \left[ \sum_{d} \sum_{h \in H} \text{pr}(D = d|X_i, h)q(h, \theta) \exp\{d(\kappa - \beta_0)\} \right].$$

Defining $\Omega = (\gamma, \kappa)$, and recalling the definition of $S(d, x, h, \Omega)$, simple algebra completes the proof.

**PROOF OF THEOREM 1**

Let subscripted $\Omega$ denote partial derivatives, e.g., $S_{\Omega}(\bullet)$ and $S_{\Omega\Omega}(\bullet)$ are the vector and matrix of the first and second partial derivatives of $S(\bullet)$ with respect to $\Omega$, respectively. Obviously, the semiparametric likelihood score is

$$L_{\Omega}(\Omega) = \sum_{i=1}^{N} \sum_{h \in H_{G_i}} \frac{S\Omega(D_i, X_i, h, \Omega)}{S(D_i, X_i, h, \Omega)} - \sum_{i=1}^{N} \sum_{d} \sum_{h \in H} \frac{S\Omega(d, X_i, h, \Omega)}{S(d, X_i, h, \Omega)}.$$ 

That $L_{\Omega}(\Omega)$ is an unbiased estimating equation is a simple consequence of the following easily proved result. Let $f_{\Omega}(\bullet)$ be the probability density function of $X$. Let the distinct values of $G$ be $(g_1, \ldots, g_M)$, and let $H_{g_j}$ be the values of $h$ consistent with $g_j$. Recall the definition of $\mu(0) = N_0/\{N \text{pr}(D = 0)\}$.

**Lemma A.1.** For any function $R(D, X, G)$, and any function $R_s(D, X, H)$,

$$\mathbb{E}\{N^{-1} \sum_{i=1}^{N} R(D_i, X_i, G_i) \} = \mu(0) \int f_{\Omega}(x) \sum_{j=1}^{M} \sum_{h \in H_{g_j}} R(d, x, g_j) \sum_{h \in H_{g_j}} S(d, x, h, \Omega) dx;$$

$$\mathbb{E}\{N^{-1} \sum_{i=1}^{N} R_s(D_i, X_i, H_i) \} = \mu(0) \int f_{\Omega}(x) \sum_{h \in H} R_s(d, x, h) \sum_{h \in H} S(d, x, h, \Omega) dx.$$ 

In addition, assuming that $N_0/\{N \text{pr}(D = 0)\} = \mu(0)$ converges to a finite, positive constant, the obvious law of large numbers applies to the sums in the expectations.

Lemma A.1 can be used to compute the expectations of the matrix of second partial derivatives (the so-called “bread of the sandwich”) and the variance of the score.
MATRICES OF SECOND PARTIAL DERIVATIVES

Note that

\[ N^{-1}L_{\Omega^T}(\Omega) = N^{-1} \sum_{i=1}^{N} \left[ \sum_{h_1 \in H_{G_1}} S_{\Omega}^1(D_i, X_i, h, \Omega) - \sum_{d} \sum_{h_2 \in H} S_{\Omega^T}(d, X_i, h_2, \Omega) \right] \]

\[ - \sum_{h_3 \in G_3} S_{\Omega}(D_i, X_i, h, \Omega) \left( \sum_{h_1 \in H_{G_1}} S_{\Omega}(D_i, X_i, h, \Omega) \right)^T \]

\[ - \sum_{d} \sum_{h_2 \in H} S_{\Omega}(d, X_i, h_2, \Omega) \left( \sum_{d} \sum_{h_1 \in H} S_{\Omega}(d, X_i, h_1, \Omega) \right)^T \]

\[ = S_{N1} - S_{N2} - S_{N3} + S_{N4}. \]

It is easy to show, using Lemma A.1, that

\[ E(S_{N1}) = E(S_{N2}) = \mu(0) \int f_X(x) \sum_d \sum_{h_2 \in H} S_{\Omega^T}(d, x, h_2, \Omega)dx \]

and that \( S_{N1} - S_{N2} = o_p(1) \). A further application of Lemma A.1 shows that the expectations, and hence the probability limits of \( S_{N3} \) and \( S_{N4} \), are given by

\[ E(S_{N3}) = \eta_3 = \mu(0) \int f_X(x) \sum_d \sum_{h_2 \in H} S_{\Omega}(d, x, h_2, \Omega) \left( \sum_{h_1 \in H_{G_1}} S_{\Omega}(d, x, h_1, \Omega) \right)^T f_X(x)dx \]

\[ E(S_{N4}) = \eta_4 = \mu(0) \int f_X(x) \sum_d \sum_{h_2 \in H} S_{\Omega}(d, x, h_2, \Omega) \left( \sum_{d} \sum_{h_1 \in H} S_{\Omega}(d, x, h_1, \Omega) \right)^T f_X(x)dx. \]

Hence, matrix \( I \) defined in Theorem 1 is \( \eta_3 - \eta_4 \).

VARIANCE OF THE SCORE

Recall that

\[ L_{\Omega}(\Omega) = \sum_{i=1}^{N} \left\{ \sum_{h_1 \in H_{G_1}} S_{\Omega}(D_i, X_i, h, \Omega) - \sum_{d} \sum_{h_2 \in H} S_{\Omega^T}(d, X_i, h_2, \Omega) \right\} \]

\[ = \sum_{i=1}^{N} \{ A_1(D_i, X_i, G_i, \Omega) - A_2(X_i, \Omega) \}. \]

Define \( A_3(d, \Omega) = E(A_1(D, X, G, \Omega) - A_2(X, \Omega)|D = d) \). Then, \( \sum_{i=1}^{N} A_3(D_i, \Omega) = 0 \), because the score is unbiased. Thus we can write

\[ L_{\Omega}(\Omega) = \sum_{i=1}^{N} \{ A_1(D_i, X_i, G_i, \Omega) - A_2(X_i, \Omega) - A_3(D_i, \Omega) \}. \]

Note that each of the terms in this sum is independent with the zero mean. Then,

\[ N^{-1}E\{L_{\Omega}(\Omega)L_{\Omega}^T(\Omega)\} = N^{-1} \sum_{i=1}^{N} E\left[ \{ A_1(D_i, X_i, G_i, \Omega) - A_2(X_i, \Omega) \} \{ \bullet \}^T \right] \]

\[ - N^{-1} \sum_{i=1}^{N} A_3(D_i, \Omega)A_3(D_i, \Omega)^T \]
where the expression \{\bullet\} means a repetition of the previous argument. The first term can be written as
\[ D_1 = -N^{-1}E[L_{\Omega}^T(\Omega)] \]
and
\[ D_3 = -N^{-1}E[L_{\Omega}L_{\Omega}^T(\Omega)] \]
for random sampling.

Since \( D_1 - D_3 = N^{-1}E[L_{\Omega}^T(\Omega)] \), we have shown that
\[ N^{-1}E[L_{\Omega}(\Omega) L_{\Omega}^T(\Omega)] = I - \Lambda. \]

Application of the central limit theorem yields Theorem 1.

**CONSISTENCY ISSUE FOR PROSPECTIVE ESTIMATING EQUATIONS**

We first prove that the modified prospective score equation (19) is unbiased for estimation of \( \kappa \) and \( \beta_1 \), assuming that \( \theta \) and \( \beta_0 \) are fixed at their true values. We note that the joint distribution of \( D, H^d, \) and \( X \) in the underlying population is characterized by the parameters \( \beta_0, \beta_1, \) and the distribution function \( V(h^d, x) = g_0(h^d) \times F(x) \). Using Lemma 1 of Roeder et al. [1996], we observe that for any given value of the parameters \( \beta_0 \) and \( \beta_1 \) and any given function \( V(\cdot) \), one can choose \( \beta_0 \) and a distribution function \( V^*(\cdot) \) such that
\[ \text{pr}_{\beta_0, \beta_1, V}(H^d, X|D) = \text{pr}_{\beta_0, \beta_1, V^*}(H^d, X|D) \]
and
\[ \text{pr}_{\beta_0, \beta_1, V^*}(D = 1) = N_1/N. \]

In particular, by construction, the authors showed that \( \beta_0^* = \kappa \) and
\[ V^*(h^d, x) \propto r_\Omega(h^d, x) V(h^d, x). \]

Let \( P \) and \( P^* \) denote the probability law for \( (D, H^d, X) \) under \( (\beta_0, \beta_1, V) \) and \( (\kappa, \beta_1, V^*) \), respectively. Let \( E \) and \( E^* \) denote expectations under the probability law \( P \) and \( P^* \). Now, the right-hand side of equation (19), when evaluated at true values of \( \kappa, \beta_1, \theta, \) and \( \beta_0^* \), can be expressed as
\[
\sum_{i=1}^{N_1} \sum_{j=1}^{N_0} E^* \left[ \frac{\partial}{\partial \beta} \log \{ \text{pr}_{\kappa, \beta_1}(D_i|H^d, X_i) \} | D_i = 1, G_i, X_i \right]
+ \sum_{i=1}^{N_1} E^* \left[ \frac{\partial}{\partial \beta} \log \{ \text{pr}_{\kappa, \beta_1}(D_i|H^d, X_i) \} | D_i = 0, G_i, X_i \right].
\]

Let
\[ q_{\kappa, \beta_1}(D, G, X) = E^* \left[ \frac{\partial}{\partial \beta} \log \{ \text{pr}_{\kappa, \beta_1}(D|H^d, X) \} | D, G, X \right]. \]

Thus, under the case-control sampling design, the asymptotic limit of (19) divided by the total sample size \( N = N_0 + N_1 \) can be written as
\[
\frac{N_1}{N} E_{G,X} \left[ q_{\kappa, \beta_1}(D, G, X) | D = 1 \right] + \frac{N_0}{N} E_{G,X} \left[ q_{\kappa, \beta_1}(D, G, X) | D = 0 \right].
\]

Since \( P^*(D = 1) = N_1/N \) and \( P^*(G, X|D) = P(G, X|D) \), we can write (A.2) as
\[
E^* D E_{G,X} \left[ q_{\kappa, \beta_1}(D, G, X) | D \right] = E_{D,H,X} \left[ \frac{\partial}{\partial \beta} \log \{ \text{pr}_{\kappa, \beta_1}(D|H^d, X) \} \right]
\]
which in turn can be shown to be zero by following the standard theory of unbiasedness of the prospective-score equations under random sampling.
To see why the proof of consistency fails for the ordinary prospective-estimating equation, we note that each individual term of equation (18), when evaluated at true values of $\kappa, \beta_1$ and $\theta$ can be written as

$$\sum_{h' \in H_0^d} \sum_{i \in I_0} \frac{h'}{h} \log \left[ \text{pr}_{\kappa, \beta_1} (D_i | h', X_i) \right] \frac{\text{pr}_{\kappa, \beta_1} (D_i | h', X_i) \cdot q(h'; \theta)}{\sum_{h' \in H_0^d} \text{pr}_{\kappa, \beta_1} (D_i | h', X_i) \cdot q(h'; \theta)}.$$

The above, however, cannot be written in the usual expectation form, because while $\text{pr}_{\kappa, \beta_1} (D_i | h', X_i)$ corresponds to the probability law of $[D | H, X]/C_{138}$ under $P/\text{C}^3$, $q(h'; \theta)$ corresponds to the probability law of $[H | X]/C_{138}= [H]/C_{138}$ under $P$. Thus, the ordinary prospective-score equation, when evaluated at $\kappa, \beta_1$, and $\theta$, does not have a conditional expectation form, which was key to the proof given in Zhao et al. [2003]. Nevertheless, we observe that

$$\text{pr}_{\kappa, \beta_1} (D_i | h', X_i) r_{\Omega} (H_i, X) = \frac{\exp \left[ D \{ \kappa + m(\beta_1, H_i, X) \} \right]}{1 + \exp \left[ \beta_0 + m(\beta_1, H_i, X) \right]}.$$

Assuming that the disease is rare in the population for all combinations of $H_i$ and $X$, one can make the approximation $[1 + \exp \{ \beta_0 + m(\beta_1, H_i, X) \}]^{-1} \approx 1$, which, when substituted in equation (19), yields the approximate estimating function of Zhao et al. [2003].

**SANDWICH VARIANCE ESTIMATOR UNDER CASE-CONTROL DESIGN**

Let $\hat{\Omega} = (\hat{\beta}^*, \hat{\theta})$ be the estimate of $\Omega = (\beta^*, \theta)$ that solves the estimating equation $\sum_{i=1}^N \Psi_{\Omega} (D_i, G_i, X_i) = 0$ for a vector-valued kernel function $\Psi_{\Omega} (D_i, G_i, X_i)$ that has the same dimension as $\Omega$. Using the standard estimating equation theory, it follows that under suitable regularity conditions,

$$N^{1/2} \left( \hat{\Omega} - \Omega_0 \right) \Rightarrow \text{Normal} \{ 0, \Sigma = \Psi_{\Omega}^{-1} (A - B)(\Psi_{\Omega}^{-1})^T \}$$

where

$$\Psi_{\Omega} = \lim_{N \to \infty} N^{-1} \sum_{i=1}^{N_0+N_1} \partial \Psi_{\Omega} (D_i, G_i, X_i) / \partial \Omega;$$

$$A = \lim_{N \to \infty} N^{-1} \sum_{i=1}^{N_0+N_1} \Psi_{\Omega} (D_i, G_i, X_i) \Psi_{\Omega}^T (D_i, G_i, X_i);$$

$$B = \lim_{N \to \infty} \sum_{d=0}^1 (N_d/N) \{ \Psi(D, X, G, \Omega) | D = d \} \{ \Psi(D, X, G, \Omega) | D = d \}^T.$$