The International Journal of Biostatistics

Volume 4, *Issue* 1 2008 *Article* 6

Inference of the Haplotype Effect in a Matched Case-Control Study Using Unphased Genotype Data

Samiran Sinha[*] Stephen B. Gruber[†]

Bhramar Mukherjee[‡] Gad Rennert[**]

[*] Texas A&M University, sinha@stat.tamu.edu

[†] University of Michigan, sgruber@med.umich.edu

[‡] University of Michigan, bhramar@umich.edu

[**] Carmel Medical Center; Technion-Israel Institute of Technology; CHS National Cancer Control Center, rennert@clalit.org.il

Inference of the Haplotype Effect in a Matched Case-Control Study Using Unphased Genotype Data[*]

Samiran Sinha, Stephen B. Gruber, Bhramar Mukherjee, and Gad Rennert

Abstract

Typically locus specific genotype data do not contain information regarding the gametic phase of haplotypes, especially when an individual is heterozygous at more than one locus among a large number of linked polymorphic loci. Thus, studying disease-haplotype association using unphased genotype data is essentially a problem of handling a missing covariate in a case-control design. There are several methods for estimating a disease-haplotype association parameter in a matched case-control study. Here we propose a conditional likelihood approach for inference regarding the disease-haplotype association using unphased genotype data arising from a matched case-control study design. The proposed method relies on a logistic disease risk model and a Hardy-Weinberg equilibrium (HWE) among the control population only. We develop an expectation and conditional maximization (ECM) algorithm for jointly estimating the haplotype frequency and the disease-haplotype association parameter(s). We apply the proposed method to analyze the data from the Alpha-Tocopherol, Beta-Carotene Cancer prevention study, and a matched case-control study of breast cancer patients conducted in Israel. The performance of the proposed method is evaluated via simulation studies.

KEYWORDS: conditional logistic regression, ECM algorithm, haplotype, matched case-control study, unphased genotype data

[*]The authors would like to thank Dr. Nilanjan Chatterjee, Division of Cancer Epidemiology and Genetics, NCI, for many detailed suggestions which greatly strengthened the manuscript. The authors would like to thank Dr. Jinbo Chen, University of Pennsylvania, for kindly providing the ATBC dataset and Professor Raymond J. Carroll, Texas A&M University for his comments. The authors would also like to thank two referees for helpful suggestions. The research of Bhramar Mukherjee was supported by NSF grant DMS 0706935 and NIH grant R03 CA130045-01.

1 Introduction

In a gene-disease association study, attention is paid to certain locations of Deoxyribo Nucleic Acid (DNA) which carry variations in its nucleotide structure. A *single nucleotide polymorphism* (SNP) could be defined as a single base change in a DNA sequence that occurs in a population with a large proportion (more that 1%). Due to rapid growth of molecular techniques, identifying several polymorphic loci or SNP on the same chromosome has become very common, and now is used for mapping complex-disease genes or identifying genetic variants responsible for a disease. However, the simple SNP based association study can be expanded to understand the biologically more relevant contiguous region of DNA containing a risk allele by examining a number of adjacent loci. This type of haplotype based analysis takes advantage of the linkage disequilibrium information from multiple SNPs together in a single analysis, and can be informative regardless of whether these SNPs simply represent markers of risk or interact in some genetically relevant manner.

A haplotype is a set of closely linked SNP's present on one chromosome which tend to be inherited together. Thus two haplotypes one from father and one from mother go together and form a pair of *haplotypes* which is called a *diplotype*. The list of unordered pairs of alleles in a diplotype is called a genotype. That means, a genotype is obtained from a pair of haplotypes without any information regarding the chromosome which is associated with each allele, and this is known as phase information. For example, if we consider three loci, with genotypes Aa, Bb, and cc, then there are two possible pairs of haplotypes ABc/abc and Abc/aBc. Due to ambiguous phase information about multiple SNP-based haplotypes, the standard method for analyzing matched case-control data, such as conditional logistic regression method treating haplotypes as a covariate is inapplicable in its standard form.

Due to high cost of molecular haplotyping, numerous methods have been proposed for assessing association of haplotypes and/or environmental risk factors with disease variable using unphased haplotype data. Clark (1990) introduced a method based on the Hardy Weinberg equilibrium (HWE) assumption to determine the phase information of multiple SNP genotypes. Excoffer and Slatkin (1995) used expectation maximization algorithm to estimate haplotype frequencies from unphased genotype data in a diploid population. Niu et al. (2002) used a Bayesian technique to estimate the unknown haplotype frequencies for large number of linked loci by using *progressive ligation* Gibbs sampler. Greenspan & Geiger (2003) proposed a Bayesian approach incorporating priors which takes into account the aspects of recombination. Kraft et al. (2005) presented a comprehensive review of some of the existing methods for analyzing matched case-control data in presence of unphased genotype data. Recently, Lin and Zeng (2006) presented a likelihood based approach for the analysis of these data collected through cohort or case-control study.

Generally, for a matched case-control study, a random sample of cases (diseased subject) is drawn from a target population, and then each case is matched with a number of controls (nondiseased subject) based on some matching variables. The unphased genotype data on multiple loci (or other covariate) are then collected from each sampled individual. In a naive method, first the haplotype frequencies are estimated by the Expectation-Maximization (EM) based algorithm using case-control sample (Excoffier and Slatkin, 1995; Fallin and Schork, 2000; Qin et al., 2002). Next each subject is assigned the most likely haplotype pair given the observed genotype, and then the standard analysis is carried out as if the haplotypes are exactly observed. The drawback of the naive strategy is that ignoring the uncertainty in the assigned haplotype pair can lead to overly narrow confidence intervals and bias, depending on the degree of misclassification. As noted in Kraft et al. (2005), if the misclassification rates differ between the cases and controls (i.e., if there is differential misclassification), then the test of haplotype-effect parameter estimates can be biased. In order to handle the uncertainty in the haplotype assignment one may employ multiple imputation technique. In multiple imputation, multiple datasets are created by randomly assigning a haplotype pair to each subject which is in accordance with the locus specific genotype data. Then disease-haplotype association parameter is estimated by taking the average of the estimates across the imputed datasets. Kraft et al. proposed to use a weighted average of the conditional likelihood, where the weights are the probabilities corresponding to different haplotype combination for the cases and controls. In a conditional likelihood paradigm Chen and Chatterjee (2006) proposed a semiparametric method for joint estimation of relative-risk parameter and cumulative baseline hazard function using cohort or nested case-control study. They proposed an alternative EM algorithm to estimate haplotype frequency from cohort and nested case-control study under HWE. Recently, Zhang et al. (2006) considered haplotype-based association study under a matched case-control design with the rare disease assumption. In all these methods, the haplotype frequencies are separately estimated from the control sample or from the combined sample of cases and controls under the HWE. Then the estimated haplotype frequencies are plugged into the prescribed likelihood for estimating the disease-haplotype association. Though Epstein and Satten (2003) estimated the relative risk parameters and the haplotype frequencies simultaneously in an unmatched case-control study, it has not been previously done for matched case-control data.

In order to take into account the uncertainty of the haplotype frequency estimates we propose to estimate the haplotype frequencies and the disease-haplotype association parameters simultaneously through a conditional likelihood of the observed data. Considering the fact that the validity of HWE in the population may often be in doubt, following Epstein and Satten (2003), we relax that assumption, and assume that only the control population is under HWE. This assumption along with the disease risk model induces a probability structure for a pair of haplotypes among the cases, which generally do not obey HWE.

The rest of the article is organized as follows. Section 2 contains model and notation while a brief description of existing methods for analyzing matched case-control data is collected in Section 3. Section 4 presents the details of the proposed method. In Section 5, the proposed method is applied to two real matched case-control datasets. The first is based on the data from Alpha-Tocopherol, Beta-Carotene (ATBC) Cancer Prevention Study (Woodson et al., 2003), and the second is a population-based case-control study on incident breast cancer conducted in northern Israel (Pujana et al., 2007). The second study recently associated three haplotype-tagging SNPs (htSNPs) at the HMMR (hyaluronan-mediated motility receptor) locus with increased risk of breast cancer. Since then, haplotype data on additional cases and controls have been added to the study-base and we analyze the same three htSNPs on this extended database including new subjects. Section 6 contains extensive simulation study which assesses the performance of the proposed method in terms of bias, efficiency, and robustness under violation of various model assumptions. Section 7 contains discussion.

Before we conclude this section, we would like to point out the main features of this article. We propose a conditional likelihood approach for inference regarding the disease-haplotype association parameters in matched case-control studies in presence of unphased genotype data. The unknown parameters are all simultaneously estimated by the expectation and conditional maximization (ECM) algorithm. The proposed method depends neither on the rare disease assumption nor on the HWE in the target population. It assumes that only the control population is under HWE. Extensive simulation study is an important asset of this paper, which shows that in terms of bias and efficiency the proposed method works well, and in many situations it outperforms the existing methods.

2 Model and Notation

Suppose we have n matched sets each comprising of 1 case and M (M$\geq$ 1) unrelated controls. Let $\boldsymbol{S}$ be the set of matching variables. Let Y_{ij} denote the binary disease variable for the subject j in the i^{th} matched set, and $G_{ij} = (g_{ij1}, \cdots, g_{ijp})$ be the set of p SNPs within a candidate region. The genotype g_{ijk} at the locus k consists of unordered pair of alleles $g_{ijk} = (g_{ijk}^{(1)}, g_{ijk}^{(2)})$. If $g_{ijk}^{(1)}$ and $g_{ijk}^{(2)}$ are inherited from the father and the mother of the subject, then the corresponding diplotype, the pair of haplotypes, will be $D_{ij} = (h_{ij}^{(1)}, h_{ij}^{(2)})$, where the haplotype $h_{ij}^{(l)} = (g_{ij1}^{(l)}, \cdots, g_{ijp}^{(l)})$, for l=1, 2. The disease-risk model we are interested in is

$$\text{pr}\{Y = 1|D = (h_s, h_t), \boldsymbol{S}_i\} = H\{\beta_0(\boldsymbol{S}_i) + \beta_{(st)}\}, \tag{1}$$

where $H(u) = \{1 + \exp(-u)\}^{-1}$. Here $\beta_0(\boldsymbol{S})$ is the matched set specific intercept parameter, which could be any type of function of the matching variables $\boldsymbol{S}$, and $\beta_{(st)}$ is the log-odds ratio parameter associated with the diplotype D. Depending on the mode of inheritance one can write $\beta_{(st)}$ in different ways. Following the notation of Lin and Zeng (2006), if h^* is the haplotype of our interest, then $\beta_{(st)} = \beta I(h_s = h_t = h^*)$ is termed as the recessive model, $\beta_{(st)} = \beta\{I(h_s = h^*) + I(h_t = h^*) - I(h_s = h_t = h^*)\}$ is termed as the dominant model, and $\beta_{(st)} = \beta\{I(h_s = h^*) + I(h_t = h^*)\}$ is termed as the log-additive model, where β is the effect of the haplotye in the disease risk. Though (1) includes only the haplotype effect, one can easily extend it to incorporate some environmental covariates and its interaction term with the other predictor variables in the model.

3 Previous Work

If the complete information on the haplotype pair were available, one could have estimated the log-odds ratio parameter $\boldsymbol{\beta}$ by using the conditional likelihood of the disease status given the diplotypes, the matching variable, and the conditioning event $T = \sum_{j=1}^{M+1} Y_{ij} = 1$. Let $\boldsymbol{Y}_{i,-k} = (Y_{i1}, \cdots, Y_{ik-1}, Y_{ik+1}, \cdots, Y_{iM+1})$, then the conditional likelihood is

$$L_{CLR}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \text{pr}(Y_{i1} = 1, \boldsymbol{Y}_{i,-1} = \boldsymbol{0}|\boldsymbol{S}_i, \{D_{ij}\}_{j=1}^{M+1}, \sum_{j=1}^{M+1} Y_{ij} = 1). \tag{2}$$

Note that due to conditioning on the number of cases in every matched set, we get rid of the nuisance parameter $\beta_0(\boldsymbol{S}_i)$. Hereafter this method will be referred as *Method-I.* Kraft et al. (2005) proposed the following likelihood for matched case-control study with unphased genotype data.

$$L_K(\boldsymbol{\beta},\boldsymbol{\pi}) = \prod_{i=1}^{n} \sum_{\boldsymbol{D}_i \in \bigotimes_{j=1}^{M+1} \mathcal{D}(G_{ij})} \Big\{ \text{pr}(Y_{i1}=1, \boldsymbol{Y}_{i,-1}=\boldsymbol{0} | \boldsymbol{S}_i, \{D_{ij}\}_{j=1}^{M+1}, \sum_{j=1}^{M+1} Y_{ij}=1) \times \prod_{r=1}^{M+1} \text{pr}(D_{ir}|G_{ir}) \Big\} \quad (3)$$

where $\boldsymbol{D}_i = (D_{i1}, \cdots, D_{iM+1})$ is a set of $(M+1)$ diplotypes, and $\mathcal{D}(G_{ij})$ is the set of all possible diplotypes which are compatible with the unphased genotype G_{ij}. Note that $\text{pr}(D_{ir}|G_{ir})$ is a function of the haplotype frequency $\boldsymbol{\pi}$. This likelihood is the expectation of (2) with respect to missing haplotypes given the unphased genotype. $\widehat{\boldsymbol{\beta}}$ is obtained by solving the score equation $\partial \text{log} L_K(\boldsymbol{\beta}, \widehat{\boldsymbol{\pi}})/\partial\boldsymbol{\beta} = 0$, where $\widehat{\boldsymbol{\pi}}$ is the estimate of the haplotype frequency obtained from the combined sample of cases and controls using the EM algorithm. In future, we will refer this method as *Method-II.*

Although the method proposed in Chen and Chatterjee (2006) was targeted for nested case-control study, it can also be applied for matched case-control data. Suppose t is the true age of onset of the disease, then the disease hazard at age t for a subject with diplotype $D = (h_r, h_s)$ is $\lambda(t) = \lambda_0(t)\exp\{\beta_D\}$, where $\lambda_0(t)$ is the baseline hazard function, and β_D is the log-odds ratio parameter associated with the diplotype D. Then partial likelihood function proposed in Equation (3) of the article involves the cumulative baseline hazard function (CBHF). Since, in general, matched case-control data do not contain age of onset of the disease, the authors proposed the following likelihood under the rare disease assumption to estimate β_D for matched case-control data.

$$L_{\text{CC}} = \prod_{i=1}^{n} \frac{\sum_{D\in\mathcal{D}(G_{i1})} \exp(\beta_D)\text{pr}(D|G_{i1};\widehat{\boldsymbol{\pi}})}{\sum_{j=1}^{M+1}\sum_{D\in\mathcal{D}(G_{ij})} \exp(\beta_D)\text{pr}(D|G_{ij};\widehat{\boldsymbol{\pi}})}, \quad (4)$$

where $\widehat{\boldsymbol{\pi}}$ is the estimated haplotype frequency of the target population obtained by using an EM algorithm under HWE. Under the rare disease assumption, one may obtain $\widehat{\boldsymbol{\pi}}$ by applying an EM algorithm only to the data on the control subjects. In rest of the article this method is referred as *Method-III.*

Zhang et al. (2006) considered a slightly different approach by adopting the following likelihood

$$L_Z(\boldsymbol{\beta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \sum_{D_{i1} \in \mathcal{D}(G_{i1})} \cdots \sum_{D_{iM+1} \in \mathcal{D}(G_{iM+1})} \mathrm{pr}\,(Y_{i1} = 1, \boldsymbol{Y}_{i,-1} = \boldsymbol{0}, \{D_{ij}\}_{j=1}^{M+1} | \boldsymbol{S}_i, \sum_{j=1}^{M+1} Y_{ij} = 1\,)\,, \quad (5)$$

which conditions only on T, but not on the diplotypes, and the contribution of the i^{th} matched set $\mathrm{pr}(Y_{i1} = 1, \boldsymbol{Y}_{i,-1} = \boldsymbol{0}, \{D_{ij}\}_{j=1}^{M+1} | \boldsymbol{S}_i, \sum_{j=1}^{M+1} Y_{ij} = 1)$ is

$$\frac{\mathrm{pr}(Y_{i1} = 1, \boldsymbol{Y}_{i,-1} = \boldsymbol{0} | \{D_{ij}\}_{j=1}^{M+1}, \boldsymbol{S}_i) \prod_{j=1}^{M+1} \mathrm{pr}(D_{ij})}{\sum_{\boldsymbol{D}_i} \sum_{k=1}^{M+1} \mathrm{pr}(Y_{ik} = 1, \boldsymbol{Y}_{i,-k} = \boldsymbol{0} | \{D_j\}_{j=1}^{M+1}, \boldsymbol{S}_i) \prod_{j=1}^{M+1} \mathrm{pr}(D_j)}$$

Note that the above conditional likelihood function involves the matched set specific nuisance parameter $\beta_0(\boldsymbol{S}_i)$. Thus Zhang et al. (2006) assumed that the disease is rare, which makes the likelihood free from the nuisance parameters, and it becomes proportional to the retrospective likelihood of Epstein and Satten (2003), and the estimate of $\boldsymbol{\beta}$ is then obtained by solving the score equation $\partial \mathrm{log} L_Z(\boldsymbol{\beta}, \widehat{\boldsymbol{\pi}})/\partial \boldsymbol{\beta} = 0$, under the rare disease assumption. Here $\widehat{\boldsymbol{\pi}}$ is the estimate of the haplotype frequency among the controls. Hereafter we will refer this method as *Method-IV* whereas the proposed method will be referred as *Method-V*.

4 Proposed Method

4.1 Modeling Haplotype Frequency

Although haplotype frequencies may vary in the target population due to admixture or stratification of the population, for the sake of simplicity we assume that the target population is genetically homogeneous, and the haplotype frequencies do not confound with the matching variables $\boldsymbol{S}$. Later on in the simulation study we consider a departure from these assumptions. Following Epstein and Satten (2003), we assume that the control population is under HWE. Let $\pi_{(st)} = \mathrm{pr}\{D = (h_s, h_t) | Y = 0\}$, then $\pi_{(st)}$ is equal to $2\pi_s\pi_t$ for $s < t$ and π_s^2 otherwise, where $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_m)$ are the haplotype frequencies among the control population, where m is the the total number of haplotypes that are present in the dataset with non-zero frequency. Let the odds of the disease given a specific diplotype $D = (h_s, h_t)$ in the matched set i is

$$\theta_{(st)} = \frac{\mathrm{pr}\{Y = 1 | \boldsymbol{S}_i, D = (h_s, h_t)\}}{\mathrm{pr}\{Y = 0 | \boldsymbol{S}_i, D = (h_s, h_t)\}} = \exp\{\beta_0(\boldsymbol{S}_i) + \beta_{(st)}\},$$

then by using the result of Satten and Kupper (1993) and Satten and Carroll (2000) we obtain the expression for the prevalence of the diplotype $D = (h_s, h_t)$ among the cases, and it is

$$\rho_{(st)} = \frac{\theta_{(st)}\pi_{(st)}}{\sum_{(j,k)} \theta_{(jk)}\pi_{(jk)}} = \frac{\exp\{\beta_{(st)}\}\pi_{(st)}}{\sum_{(j,k)} \exp\{\beta_{(jk)}\}\pi_{(jk)}}.$$

Thus even though the control population is under HWE, in general the case population does not follow HWE unless one assumes log-additive model for the haplotype effect.

4.2 Likelihood of the Proposed Method

We start out with the joint likelihood of the disease variable given the genotype data and condition on the number of cases of every matched set. Thus the likelihood function is

$$\begin{aligned} L_{\text{OBS}}(\boldsymbol{\beta}, \boldsymbol{\pi}) &= \prod_{i=1}^{n} \text{pr}(Y_{i1} = 1, \boldsymbol{Y}_{i,-1} = \boldsymbol{0} | \boldsymbol{S}_i, \{G_{ij}\}_{j=1}^{M+1} \sum_{j=1}^{M+1} Y_{ij} = 1) \\ &= \prod_{i=1}^{n} \frac{\text{pr}(Y_{i1} = 1, \boldsymbol{Y}_{i,-1} = \boldsymbol{0} | \boldsymbol{S}_i, \{G_{ij}\}_{j=1}^{M+1})}{\sum_{k=1}^{M+1} \text{pr}(Y_{ik} = 1, \boldsymbol{Y}_{i,-k} = \boldsymbol{0} | \boldsymbol{S}_i, \{G_{ij}\}_{j=1}^{M+1})} \\ &= \prod_{i=1}^{n} \frac{\text{pr}(Y_{i1} = 1 | G_{i1}, \boldsymbol{S}_i)/\text{pr}(Y_{i1} = 0 | G_{i1}, \boldsymbol{S}_i)}{\sum_{k=1}^{M+1} \text{pr}(Y_{ik} = 1 | G_{ik}, \boldsymbol{S}_i)/\text{pr}(Y_{ik} = 0 | G_{ik}, \boldsymbol{S}_i)}. \qquad (6) \end{aligned}$$

By a straight forward calculation one can show that the odds of the disease given the genotype data is a function of $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$, and it is

$$\frac{\text{pr}(Y_{ij} = 1 | G_{ij}, \boldsymbol{S}_i)}{\text{pr}(Y_{ij} = 0 | G_{ij}, \boldsymbol{S}_i)} = \sum_{(h_s,h_t)\in\mathcal{D}(G_{ij})} \theta_{(st)} \frac{\pi_{(st)}}{\sum_{(h_t,h_t')\in\mathcal{D}(G_{ij})} \pi_{(tt')}}. \qquad (7)$$

Hence using the relation (7) in Equation (6) we obtain

$$\begin{aligned} L_{\text{OBS}} &= \prod_{i=1}^{n} \frac{\sum_{(h_s,h_t)\in\mathcal{D}(G_{i1})} \theta_{(st)}\{\pi_{(st)}/\sum_{(h_t,h_t')\in\mathcal{D}(G_{i1})} \pi_{(tt')}\}}{\sum_{j=1}^{M+1}\sum_{(h_s,h_t)\in\mathcal{D}(G_{ij})} \theta_{(st)}\{\pi_{(st)}/\sum_{(h_t,h_t')\in\mathcal{D}(G_{ij})} \pi_{(tt')}\}} \\ &= \prod_{i=1}^{n} \frac{\sum_{(h_s,h_t)\in\mathcal{D}(G_{i1})} \exp\{\beta_{(st)}\}\{\pi_{(st)}/\sum_{(h_t,h_t')\in\mathcal{D}(G_{i1})} \pi_{(tt')}\}}{\sum_{j=1}^{M+1}\sum_{(h_s,h_t)\in\mathcal{D}(G_{ij})} \exp\{\beta_{(st)}\}\{\pi_{(st)}/\sum_{(h_t,h_t')\in\mathcal{D}(G_{ij})} \pi_{(tt')}\}} \qquad (8) \end{aligned}$$

Note that likelihood (8) is free from the nuisance parameter $\beta_0(\boldsymbol{S}_i)$.
Remark 1. Likelihood (6) is an exact likelihood without the rare disease assumption. It is a function of the haplotype frequency $\boldsymbol{\pi}$ of the control population and the association parameter $\boldsymbol{\beta}$.

Remark 2. One should carefully note that when the phase information is available, likelihood (6) does not reduce to L_{CLR}, which is semiparametrically efficient. Thus in absence of phase ambiguity, the proposed method may not be a better choice for analyzing the data compared to the CLR approach which does not require the estimation of the haplotype frequencies.

Remark 3. Likelihoods (8) and (4) are very similar, except $\boldsymbol{\pi}$ is replaced by $\widehat{\boldsymbol{\pi}}$ in (4). Also, likelihood (4) is obtained under the rare disease assumption.

Remark 4. The difference between different likelihood functions should be noted. Though likelihood (3) is lack of proper probabilistic interpretation, one can think of it as a marginal likelihood. Both the likelihood functions (6) and (4) are conditional on the observed genotype data, whereas likelihood (5) is a direct function of the diplotype. Thus, if the diplotypes were observed, (5) will produce more efficient estimate of the parameter than the other methods when the disease is rare. The other reason for gain in efficiency of (5) is that it extracts information from all matched sets even the ones which are diplotype concordant. In the downside, this likelihood heavily relies on HWE for calculating marginal probability of the diplotypes, compared to likelihoods (6) and (4) which use conditional probability of diplotype given the unphased genotype and the disease status. Nonetheless we only assume HWE among the control population, hence the violation of HWE in the target population will have less impact on the proposed method than the method based on likelihood (5).

Remark 5. As noted in Breslow (1996), the conditional likelihood (6) is both prospective and retrospective, thus it corrects for ascertainment bias. Also (5) implicitly corrects for ascertainment bias as it is equivalent to retrospective likelihood under the rare disease assumption. For discussion related to bias and efficiency for different likelihoods see Kraft and Thomas (2000).

4.3 Method of Estimation

In presence of missing diplotypes, one may think of using EM algorithm to estimate the parameter of interest, where the E-step reduces to finding conditional expectation of the log of complete data likelihood, and the M-step is simply the maximization of the conditional expectation obtained in the previous step. However, if the M-step is not simple, as in this situation, one may consider to break up the M-step into two conditional maximization (CM) steps– which is known as ECM algorithm (Meng and Rubin, 1993). Let $\Theta = (\boldsymbol{\beta}, \boldsymbol{\pi})$. Furthermore we assume that given the genotype and disease status diplotypes and the stratification variables $\boldsymbol{S}$ are independent. If the dipolotyes were observed, then the *complete data likelihood* would be

$$\begin{aligned} L_{\text{compl}}(\Theta) &= \prod_{i=1}^{n} \text{pr}(Y_{i1}=1, \boldsymbol{Y}_{i,-1}=\mathbf{0}, \{D_{ij}\}_{j=1}^{M+1} | \{G_{ij}\}_{j=1}^{M+1}, \boldsymbol{S}_i, \sum_{j=1}^{M=1} Y_{ij}=1) \\ &= \prod_{i=1}^{n} \Bigg\{ \text{pr}(D_{i1}|Y_{i1}=1, G_{i1}) \prod_{j=2}^{M+1} \text{pr}(D_{ij}|G_{ij}, Y_{ij}=0) \\ &\qquad \text{pr}(Y_{i1}=1, \boldsymbol{Y}_{i,-1}=\mathbf{0}|\{G_{ij}\}_{j=1}^{M+1}, \boldsymbol{S}_i, \sum_{j=1}^{M=1} Y_{ij}=1) \Bigg\}. \end{aligned}$$

The ECM algorithm consists of the following three steps. In the E-step of the $(t+1)^{th}$ iteration we take expectation of the log of complete data likelihood with respect to the unobserved D_{ij} with the following probability mass function. Let $c_{(rs),ij}(\Theta^{(t)}) = \text{pr}\{D_{ij}=(h_r,h_s)|G_{ij}, Y_{ij}=1; \Theta^{(t)}\}$ and $\overline{c}_{(rs),ij}(\Theta^{(t)}) = \text{pr}\{D_{ij}=(h_r,h_s)|G_{ij}, Y_{ij}=0; \Theta^{(t)}\}$, then

$$\begin{aligned} \overline{c}_{(rs),ij}(\Theta^{(t)}) &= \frac{\pi^{(t)}_{(st)}}{\sum_{(h_t,h'_t)\in\mathcal{D}(G_{ij})} \pi^{(t)}_{(tt')}} \text{ for } (h_r,h_s)\in\mathcal{D}(G_{ij}) \\ c_{(rs),ij}(\Theta^{(t)}) &= \frac{\pi^{(t)}_{(st)} \exp\{\beta^{(t)}_{st)}\}}{\sum_{(h_t,h'_t)\in\mathcal{D}(G_{ij})} \pi^{(t)}_{(tt')} \exp\{\beta^{(t)}_{(tt')}\}} \text{ for } (h_r,h_s)\in\mathcal{D}(G_{ij}). \end{aligned}$$

Let $Q(\Theta^{(t+1)}|\Theta^{(t)}) = E\Big[\log\{L_{\text{compl}}(\Theta)\}\Big]$, then $Q(\Theta^{(t+1)}|\Theta^{(t)})$ is

$$\begin{aligned} &\sum_{i=1}^{n} \Bigg\{ \sum_{(h_r,h_s)\in\mathcal{D}(G_{i1})} c_{(rs),i1}(\Theta^{(t)}) \text{logpr}\{D_{ij}=(h_r,h_s)|G_{ij}, Y_{ij}=1; \Theta^{(t+1)}\} \\ &+ \sum_{j=2}^{M+1} \sum_{(h_r,h_s)\in\mathcal{D}(G_{ij})} \overline{c}_{(rs),ij}(\Theta^{(t)}) \text{logpr}\{D_{ij}=(h_r,h_s)|G_{ij}, Y_{ij}=0; \Theta^{(t+1)}\} \\ &+ \text{logpr}(Y_{i1}=1, \boldsymbol{Y}_{i,-1}=\mathbf{0}|\{G_{ij}\}_{j=1}^{M+1}, \boldsymbol{S}_i, \sum_{j=1}^{M=1} Y_{ij}=1; \Theta^{(t+1)}) \Bigg\} \qquad (9) \end{aligned}$$

Next we maximize $Q(\Theta^{(t+1)}|\Theta^{(t)})$ with respect to $\Theta^{(t+1)}$.
CM step for $\boldsymbol{\beta}$

We fix $\boldsymbol{\pi}^{(t+1)} = \boldsymbol{\pi}^{(t)}$ in $Q(\Theta^{(t+1)}|\Theta^{(t)})$, and then maximize it with respect to $\boldsymbol{\beta}^{(t+1)}$. The part of (9) which is a function of $\boldsymbol{\beta}^{(t+1)}$ is

$$\sum_{i=1}^{n}\Big\{\sum_{(h_r,h_s)\in\mathcal{D}(G_{i1})} c_{(rs),i1}(\Theta^{(t)})\text{logpr}\{D_{ij}=(h_r,h_s)|G_{ij},Y_{ij}=1;\Theta^{(t+1)}\}$$

$$+\text{logpr}(Y_{i1}=1,\boldsymbol{Y}_{i,-1}=\mathbf{0}|\{G_{ij}\}_{j=1}^{M+1},\boldsymbol{S}_i,\sum_{j=1}^{M=1}Y_{ij}=1;\Theta^{(t+1)})\Big\}.$$

We maximize the above function using Newton-type algorithm, which numerically calculates the gradient and the Hessian matrix of the above function.
CM step for $\boldsymbol{\pi}$

In this step we determine $\boldsymbol{\pi}^{(t+1)}$ by fixing $\beta^{(t+1)}$ to its recently updated value. Note that $\boldsymbol{\pi}$ is present in all three components of $Q(\Theta^{(t+1)}|\Theta^{(t)})$. The score equation corresponding to $\boldsymbol{\pi}^{(t+1)}$ with the constraint $\sum_{l=1}^{m}\pi_l^{(t+1)}=1$ is $\partial Q(\Theta^{(t+1)}|\Theta^{(t)})/\partial\boldsymbol{\pi}^{(t+1)}-\lambda I_m=0$, which is equivalent to

$$\pi_k^{(t+1)}\Big[\sum_{i=1}^{n}\sum_{j=1}^{M+1}\frac{\partial A_{ij}^{(t+1)}/\partial\pi_k^{(t+1)}}{A_{ij}^{(t+1)}}+\sum_{i=1}^{n}\frac{\partial}{\partial\pi^{(t+1)}}\log\{\sum_{j=1}^{M+1}\frac{B_{ij}^{(t+1)}}{A_{ij}^{(t+1)}}\}+\lambda\Big]$$

$$=\sum_{i=1}^{n}\Big[\sum_{(h_r,h_s)\in\mathcal{D}(G_{i1})}c_{(rs),i1}(\Theta^{(t)})\{I(s=k)+I(r=k)\}$$

$$+\sum_{j=2}^{M+1}\sum_{(h_r,h_s)\in\mathcal{D}(G_{ij})}\overline{c}_{(rs),ij}(\Theta^{(t)})\{I(s=k)+I(r=k)\}\Big], \tag{10}$$

for $k=1,\cdots,m$, where λ is the Lagrangian multiplier, and I_m is the identity matrix of order m. Also, $A_{ij}^{(t+1)}=\sum_{(h_r,h_s)\in\mathcal{D}(G_{ij})}\pi_{(rs)}^{(t+1)}$ and $B_{ij}^{(t+1)}=\sum_{(h_r,h_s)\in\mathcal{D}(G_{ij})}\exp\{\beta_{(rs)}^{(t+1)}\}\pi_{(rs)}^{(t+1)}$. Thus λ is determined by setting $\sum_{i=1}^{m}\pi_i$ equal to 1. $\boldsymbol{\pi}^{(t+1)}$ is solved iteratively. We start with an initial value $\boldsymbol{\pi}^{(t+1)}=\boldsymbol{\pi}^{(t+1),0}$, and then calculate $\boldsymbol{\pi}^{(t+1),1}$ using Equation (10), and repeat this step until $\boldsymbol{\pi}^{(t+1)}$ converges. Thus the steps for the ECM algorithm are as follows.

Step 0. Initialize $\boldsymbol{\beta}=\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\pi}=\boldsymbol{\pi}^{(0)}$, and $\Theta^{(0)}=(\boldsymbol{\beta}^{(0)},\boldsymbol{\pi}^{(0)})$.
Step 1. Calculate $Q(\Theta^{(1)}|\Theta^{(0)})$.
Step 2. Determine $\boldsymbol{\beta}^{(1)}$ by fixing $\boldsymbol{\pi}^{(1)}=\boldsymbol{\pi}^{(0)}$.
Step 3. Determine $\boldsymbol{\pi}^{(1)}$. Set $\Theta^{(1)}=(\boldsymbol{\beta}^{(1)},\boldsymbol{\pi}^{(1)})$.
Step 4. Repeat step 1 through 3 until

$$\Big|\Big|\frac{\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)}}{\boldsymbol{\beta}^{(t)}}\Big|\Big|+\Big|\Big|\frac{\boldsymbol{\pi}^{(t+1)}-\boldsymbol{\pi}^{(t)}}{\boldsymbol{\pi}^{(t)}}\Big|\Big|<10^{-5}.$$

In the ECM algorithm the initial values should be carefully chosen. The initial value of $\boldsymbol{\pi}$ was set to $\widehat{\boldsymbol{\pi}}$ which is obtained by applying the EM algorithm to the combined case and control sample, and the initial value of $\boldsymbol{\beta}$ can be set to any reasonable estimate of $\boldsymbol{\beta}$. For our data analysis and simulation we used the estimate of $\boldsymbol{\beta}$ obtained by using Kraft's method as the initial value for $\boldsymbol{\beta}$. For the data analysis and simulation study, the ECM algorithm converged within 5 or 6 iterations. The estimate of the variance of the estimator can be obtained by inverting observed information matrix (Louis, 1982).

Remark 6. The proposed method can easily handle missing genotype data. We assume that data are either missing completely at random (MCAR) or missing at random (MAR), which do not require to model the missingness process in the likelihood-based inference. An unphased data G with missing genotype at one/more locus results in larger set $\mathcal{D}(G)$ than a G without missing genotype. For example if $G = (21100)$, $\mathcal{D}(G) = \{(11100, 10000), (11000, 10100)\}$. On the other hand if $G^* = (2\text{NA}100)$, where the genotype at locus 2 is missing, then $\mathcal{D}(G^*) = \{(11100, 11000), (11100, 10000), (11000, 10100), (10100,\ \ 10000)\}$.
Remark 7. A more flexible model for the diplotype frequency is to assume

$$\pi_{(st)} = 2(1-f)\pi_s\pi_t + f\pi_s I(s=t) \tag{11}$$

among the control population, which involves the fixation parameter f. Following Satten and Epstein (2004) the model parameters are still identifiable as we have already assumed that the haplotype frequencies do not vary across the strata. Furthermore, for the estimation one needs to estimate f using a conditional maximization step. Note that even this flexible model may entail bias if the diplotype frequencies follow some other pattern which is neither captured by HWE nor by (11).

5 Real Data Examples

5.1 Application to Data from the ATBC Study

For illustration purposes we apply all four methods, Method-II through Method-V, on the data from ATBC cancer prevention study conducted in Finland jointly by National Public Health Institute of Finland and National Cancer Institute (NCI) of USA. This data have previously been analyzed by Chen and Chatterjee (2006). It is a large randomized, double blinded, placebo-controlled, primary prevention trial to determine whether daily supplementation of alpha-tocopherol, beta-carotene, or both would reduce the chance of having lung or other cancers among male smokers. The cohort of the ATBC

Method		Haplotype	
		AC/GT	GC
		Association Parameter	
		β_1	β_2
II	EST	-0.1637	0.0922
	SE	0.4339	0.1657
III	EST	-0.1560	0.0909
	SE	0.4225	0.1645
IV	EST	-0.1493	0.0876
	SE	0.4917	0.1749
V	EST	-0.1541	0.0903
	SE	0.4690	0.1754

Table 1: Results of the ATBC data analysis. Here EST and SE stand for the estimate and standard error.

study consisted of 29,133 men between age 50 to 69 years, and who smoked at least 5 cigarettes per day, and the participants were assigned to one of 4 treatment groups. The participants received either alpha-tocopherol, beta-carotene, both supplements, or placebo capsules for 5-8 years. From January 1, 1983 to December 31, 1994 208 prostate cancer patients have been identified in the cohort, and corresponding to each case one control has been chosen based on the matching variables length of follow-up, age at randomization, intervention group, and study clinic. Here we just focus our attention on the effect of interleukin-1 (IL1A) gene cluster on the risk of prostate cancer. Two polymorphisms, IL1A889 (A/G) and IL1A4845 (T/C) were genotyped within the IL1A region. Out of 208 case-control pairs we consider only 179 case-control pairs excluding the matched sets with no information on the genotype of both case and control. There were four possible haplotypes AT, AC, GT, and GC. The estimated frequencies of the haplotypes among the control population are 0.6473, 0.0212, 0.0225, and 0.309. We found that the two polymorphisms are in strong linkage disequilibrium with $D^{'} = 0.8962$ and the test statistic $\chi_1^2 = 144.5953$. Following Chen and Chatterjee (2006) we treat AT as the reference category which has maximum frequency among the possible haplotypes, and assumed additive model for the haplotype effect. As AC and GT have low frequency ($< 3\%$) we decide to clump these two categories and will refer them as HAP-I, and GC will be denoted as HAP-II. Let β_1 and β_2 are the two relative risk parameters corresponding to HAP-I and II respectively, then the disease incidence model can be written as

$$\mathrm{pr}(Y_{ij}=1|D_{ij},\boldsymbol{S}_i)=H\{\beta_0(\boldsymbol{S}_i)+\sum_{k=1}^{2}\beta_k X_k\},$$

where X_k represents the number of copies of haplotype k present in the diplotype D_{ij}, We analyze the data by using methods II, III, IV, and V. The results of the data analysis are given in Table (1). None of the methods shows any significant association between the disease and any of the haplotypes. Under the proposed method the odds ratio estimate of HAP-I and HAP-II are 0.85 (95% CI=(0.07, 1.65), P-value=0.371) and 1.09 (95% CI=(0.72, 1.47), P-value=0.31) respectively. One should also note that there is not much difference among the estimates obtained under different methods. The AIC value corresponding to this additive model was 247.9972. As suggested by a reviewer, we also consider dominant and recessive model alternatives with HAP-I as the main haplotype of interest. Under the dominant model the disease prevalence model is $\mathrm{pr}\{Y_{ij}=1|D_{ij}=(h_r,h_s),\boldsymbol{S}_i\}=H[\beta_0(\boldsymbol{S}_i)+\beta\{I(h_r=\text{HAP-I})+I(h_s=\text{HAP-I})-I(h_r=h_s=\text{HAP-I})\}]$. Using Method V the odds ratio estimate of HAP-I is 1.03 (95% CI=(0.58, 1.48), P-value=0.4463). The AIC value of this dominant model is 246.422. For the recessive model we fit $\mathrm{pr}\{Y_{ij}=1|D_{ij}=(h_r,h_s),\boldsymbol{S}_i\}=H\{\beta_0(\boldsymbol{S}_i)+\beta I(h_r=h_t=\text{HAP-I})\}$. The odds ratio estimate due to Method V is 1.86(95% CI=(0.25, 3.47), P-value=0.079). The AIC value of this model is 242.422, which suggests that the recessive model is preferable than the other models. However, in all the models the effect of HAP-I is statistically insignificant. In contrast to the null results in this example, in the next example we will observe another inferential scenario where a haplotype on the HMMR locus significantly increases the risk of breast cancer in an Israeli population.

5.2 Application to Israeli Breast Cancer Data

This case-control study conducted in northern Israel is a population-based study of incident breast cancer cases identified through rapid case ascertainment between January 2000 and July 2006. Control women were randomly selected from a comprehensive list of insurees which covers approximately 70% of the women of northern Israel who are at risk. For each case, a control woman was identified randomly who was within 1 year age difference with the case subject, who has the same self-reported ancestry as the case (Jewish versus non-Jewish) and who is in the same geographical clinic area. Genomic DNA derived from blood lymphocytes was used for genotyping, and for our analysis, we consider genotype data for three htSNPs namely, rs7712023 (A/T), rs299290 (C/T), and rs10515860 (A/G). Pujana et al. (2007) analyzed the same htSNPs based on 923 1:1 matched case-control pairs and found a statistically significant association between the presence of breast cancer and A-C-A haplotype. They also found that htSNP rs10515860 captured almost all of the variation associated with the risk. Our analysis is based on the combined 1:1 matched case-control dataset with $n = 1445$ matched pairs which included the test data of 923 matched pairs as well as the validation dataset from the Israel study used in the original article. We will note that the exciting finding published in the original scientific article is consistent with the results of our current analysis. Note that based on the three htSNPs there are 8 possible haplotypes A-C-A, T-C-A, A-T-A, T-T-A, A-C-G, T-C-G, A-T-G, and T-T-G. Due to small prevalences we merge the first four haplotypes having allele A at rs10515860, and call them as HAP-I. As haplotype A-C-G has also very small frequency we will merge it with T-T-G. Since this combined category bears maximum frequency we will consider it as the reference category. The remaining haplotypes T-C-G and A-T-G will be referred as HAP-II and HAP-III. Approximately 36% of the genotype data were missing for the location rs7712023 and rs299290. However the difference of the missingness probability does not vary across the case-control status. Therefore we may assume that the data were missing at random and the proposed method can be applied to the dataset. The dataset is analyzed by all four methods with an additive model, and the results are presented in Table 2. Based on all the methods HAP-I is significantly associated with breast cancer incidence. By Method-V the presence of each copy of HAP-I increases a woman's risk of having breast cancer by 29%. The odds ratio estimate is obtained as $\widehat{OR} = 1.29$ with P-value 0.005, and 95% confidence interval $(1.06, 1.57)$. Under the proposed method the AIC value for this additive model is 1991.76. We also fit a dominant model using the proposed method to the data treating HAP-I as the haplotype of interest. The estimate of the corresponding OR is 1.31 (95% CI=(1.08,1.60), P-value=0.007), and the AIC value corresponding to the dominant model is 1994.91. Finally, we fit a recessive model to the data using the proposed method with HAP-I as the main haplotype. The estimate of the odds ratio parameter for the model is 1.44 (95% CI=(0.63,3.3), P-value=0.389), an insignificant association in contrast to the findings of the additive and the dominant model. However, the AIC value corresponding to this recessive model is 2002.30. Thus in terms of AIC values, the additive model is superior than the other models and this result is consistent with the findings of Pujana et al. (2007).

Method		Haplotype		
		HAP-I	HAP-II	HAP-III
		Association Parameter		
		β_1	β_2	β_3
II	EST	0.2515	0.1859	-0.0122
	SE	0.0977	0.1246	0.0770
III	EST	0.2526	0.1893	-0.0089
	SE	0.0973	0.1261	0.0787
IV	EST	0.2788	0.2595	-0.0200
	SE	0.1522	0.1773	0.0995
V	EST	0.2550	0.1893	-0.0083
	SE	0.0995	0.1240	0.0768

Table 2: Results of the Israeli breast cancer data analysis. Here EST and SE stand for the estimate and standard error.

6 Simulation Study

In order to study the performance of the proposed method we conduct a detailed simulation study. For generating realistic data, we consider 17 haplotypes based on 5 tightly linked SNPs on a putative region on chromosome 22 among the control sample of FUSION data as illustrated in Epstein and Satten (2003). Three haplotypes 01110, 10010, and 11110 occurred with frequency less than 10^{-6}, therefore we discarded these haplotypes and the remaining haplotypes with their frequency are given by the second column of Table (3). Following Epstein and Satten (2003) we focused on the haplotype 01100 which may increase the odds of type 2 diabetes. So using the haplotype frequencies and assuming HWE we generate dipoltypes for a cohort of size $N = 20,000$. In addition, we generate a matching variable S from the Gamma($\alpha = 1.25, \beta = 1.65$). We consider only the log-additive effect of the haplotype 01100. Next we generate the disease variable Y from the Bernoulli distribution with the success probability $H\{-3.2 + 0.51S + \beta s(D)\}$, where $s(D)$ represents the number of copies of haplotype 01100 present in the diplotype D. Thus $s(D)$ can only take values 0, 1, or 2. We consider three distinct values of β, a) $\beta = 0$, b) $\beta = 0.3$, and c) $\beta = 0.6$. The coefficient of S is so chosen that the odds of the disease is increased by 2 for changing S from its 10^{th} quantile to the 90^{th} quantile. From this cohort we construct 1:1 matched case-control data as follow. First we randomly sample 150 cases out of the cohort, and for each sampled case we draw a control subject from the cohort so

Haplotype	Frequency		
10000	0.0136	0.0360	0.1360
01000	0.0000	0.1000	0.1000
00100	0.0035	0.1350	0.1350
10100	0.0520	0.0520	0.0520
01100	**0.2514**	**0.1334**	**0.0334**
11100	0.0110	0.0110	0.0110
00010	0.0000	0.1000	0.1000
00110	0.0018	0.0018	0.0018
10110	0.0317	0.0317	0.0317
01101	0.0012	0.0012	0.0012
00011	0.0042	0.0042	0.0042
10011	0.3574	0.1235	0.1235
01011	0.1292	0.1292	0.1292
11011	0.1391	0.1391	0.1391
01111	0.0019	0.0019	0.0019
11111	0.0020	0.0000	0.0000

Table 3: Haplotype frequencies used for simulation studies.

that the values of the matching variable of the case and control subject are at most 0.01 apart. In order to induce more phase ambiguity we randomly select 20% of the subjects and then delete the genotype information of a randomly selected locus.

Each dataset is analyzed by five different methods discussed in Sections 3 and 4. For the simulated dataset we record phase information, and using the phase information we obtain the CLR estimate of the parameter β (*Method-I*). For the other methods we do not use the phase information, rather work only with the unphased genotype data. We simulate $ns = 300$ datasets, and report the estimate (EST) of the parameter which is obtained by taking average of the estimates across the simulated datasets. The performance of the methods is examined through i) simulation variance of the estimate (SVAR) which is $\sum_{i=1}^{ns}(\hat{\beta}_i - \overline{\beta})^2/(ns-1)$, where $\overline{\beta} = \sum_{i=1}^{ns}\hat{\beta}_i/ns$, and ns is the number of simulations, ii) average of the estimated variance (AEV) which is $\sum_{i=1}^{ns}\widehat{var}(\widehat{\beta}_i)/ns$, iii) 95% coverage probability (CP) which is the proportion of times the estimated 95% confidence interval contains the true value of the parameter, iv) power (PWR) which is the proportion of times $|\widehat{\beta}/\widehat{v}ar(\widehat{\beta})| > 1.96$, and v) mean square error (MSE) which is equal to $\sum_{i=1}^{ns}(\widehat{\beta}_i - \beta)^2/ns$. The results are presented in Table (4). As expected, with known phase information, Method I performs well in terms of bias and efficiency. For $\beta = 0.3$ and $\beta = 0.6$ the amount of bias due to Method IV is large. The intuitive reason for this bias may be that Method IV presumes that the disease is rare which is not true when the log odds ratio parameter gets large. Hence, the method produces conservative estimate of the association parameter which is bias towards null. On the other hand, due to smaller value of the standard error, Method IV has better power than the other procedures. Among Methods II, III, and V, the proposed method has smallest SVAR and MSE. Also the proposed method shows some gain in power over Methods II and III.

Hardy-Weinberg disequilibrium

To study the robustness of the proposed method when HWE is violated in the control population we do the following. We generate diplotypes for a cohort with the following probability

$$\text{pr}\{D = (h_r, h_s)\} = \begin{cases} (1-f)\pi_r^2 + f\pi_r & \text{if } r = s \\ 2(1-f)\pi_r\pi_s & \text{otherwise,} \end{cases}$$

where $\sum_{s=1}^{m} \pi_s = 1$. Here f is called the fixation index which is the probability that both the pair of haplotypes trace back to a common ancestor. Generally this inbreeding is very low in human population (Tzeng and Roeder, 2006), therefore for the simulation purpose we consider $f = 0.02$ and $f = 0.05$. One should note that if the HWE is violated in the target population, and the same holds for the control population as well. The results are presented in Table 5. Among Methods II, III, IV, and V, the proposed method shows gain in power when $\beta = 0.3$ and $\beta = 0.6$. MSE and simulation variance are smaller in the proposed method than Method II and III. When f increases from 0.02 to 0.05, the MSE increases for all four methods for $\beta = 0$, 0.3 and 0.6.

Mixture of populations but no admixture

For unrelated case-control studies it is often the case that a population consists of several subpopulations with varying genetic characteristics. Therefore, we assume that the target population consists of three subpopulations which have different haplotype frequencies. The haplotype frequencies of the three subpopulations are given by columns 2, 3, and 4 of Table (3), and an individual has equal probability of being one of the three subpopulations. Also we assume that the disease risk model for the three subpopulations are different and they are $H\{-3.2 + 0.51S + \beta s(D)\}$, $H\{-3 + 0.51S + \beta s(D)\}$, and $H\{-2.7 + 0.51S + \beta s(D)\}$ respectively. We assume that HWE holds within

Method		$\beta = 0$	$\beta = 0.3$	$\beta = 0.6$
I	EST	-0.0081	0.2846	0.5831
(known	SVAR	0.0342	0.0352	0.0389
phase)	AEV	0.0383	0.0361	0.0404
	MSE	0.0343	0.0354	0.0392
	CP	94.3	93.0	91.0
	PWR	5.7	35.7	85.7
II	EST	-0.0015	0.2852	0.5812
	SVAR	0.0372	0.0367	0.0424
	AEV	0.0404	0.0384	0.0429
	MSE	0.0372	0.0369	0.0428
	CP	94.7	92.7	92.3
	PWR	5.3	30.7	80.7
III	EST	0.0043	0.2891	0.5831
	SVAR	0.0372	0.0400	0.0423
	AEV	0.0421	0.0373	0.0443
	MSE	0.0372	0.0401	0.0426
	CP	95.0	93.0	92.0
	PWR	5.0	31.3	81.0
IV	EST	-0.0183	0.2649	0.5505
	SVAR	0.0362	0.0365	0.0368
	AEV	0.0393	0.0362	0.0363
	MSE	0.0365	0.0377	0.0393
	CP	93.0	92.0	92.5
	PWR	7.0	39.0	84.0
V	EST	-0.0010	0.2832	0.5865
	SVAR	0.0365	0.0358	0.0409
	AEV	0.0396	0.0376	0.0413
	MSE	0.0365	0.0361	0.0411
	CP	94.7	93.0	92.7
	PWR	5.3	34.0	83.3

Table 4: Results of the simulation study where the haplotype frequencies are given by column 2 of Table (3), and the HWE holds (f=0).

each subpopulation. We simulate $ns = 300$, 1:1 matched case-control data using the matching variable S, and then analyze them by using all five meth-

Method		f=0.02			f=0.05		
		$\beta = 0$	$\beta = 0.3$	$\beta = 0.6$	$\beta = 0$	$\beta = 0.3$	$\beta = 0.6$
I (known phase)	EST	-0.0015	0.3109	0.6152	0.0028	0.3042	0.6154
	SVAR	0.0362	0.0408	0.0349	0.0407	0.0386	0.0373
	AEV	0.0358	0.0370	0.0379	0.0366	0.0360	0.0379
	MSE	0.0362	0.0409	0.0351	0.0407	0.0386	0.0375
	CP	91.3	91.0	92.3	91.7	93.3	95.3
	PWR	8.7	38.7	90.7	8.3	40.3	90.7
II	EST	-0.0020	0.3122	0.6175	0.0023	0.3027	0.6188
	SVAR	0.0399	0.0415	0.0365	0.0442	0.0428	0.0392
	AEV	0.0381	0.0391	0.0409	0.0393	0.0385	0.0406
	MSE	0.0399	0.0416	0.0368	0.0442	0.0428	0.0396
	CP	91.7	93.0	93.7	92.7	92.7	96.3
	PWR	8.3	39.3	90.0	7.3	36.7	92.0
III	EST	0.0009	0.3161	0.6209	0.0069	0.3082	0.6209
	SVAR	0.0404	0.0426	0.0367	0.0446	0.0437	0.0388
	AEV	0.0394	0.0403	0.0422	0.0407	0.0399	0.0416
	MSE	0.0404	0.0429	0.0371	0.0446	0.0438	0.0392
	CP	92.7	93.7	94.7	91.3	93.7	96.0
	PWR	7.3	37.0	90.7	8.7	37.3	90.3
IV	EST	-0.0178	0.2940	0.6087	-0.0106	0.2936	0.6208
	SVAR	0.0397	0.0396	0.0337	0.0482	0.0429	0.0350
	AEV	0.0376	0.0378	0.0374	0.0429	0.0388	0.0377
	MSE	0.0400	0.0396	0.0338	0.0483	0.0429	0.0354
	CP	92.0	92.0	94.3	92.3	91	95.7
	PWR	8.0	36.0	87.0	7.7	35.0	90.7
V	EST	-0.0010	0.3099	0.6120	0.0027	0.3010	0.6130
	SVAR	0.0392	0.0405	0.0353	0.0435	0.0419	0.0379
	AEV	0.0373	0.0381	0.0393	0.0386	0.0375	0.0391
	MSE	0.0392	0.0406	0.0354	0.0435	0.0419	0.0381
	CP	91.7	93.0	94.7	92.0	93.0	96.0
	PWR	8.3	39.9	92.3	8.0	38.3	92.3

Table 5: Results of the simulation study where the haplotype frequencies are given by column 2 of Table (3), and the HWE is violated.

ods. The results are presented in the left panel of Table 6. The results show that type-I error probability has inflated for all methods. MSE increased and power of all the methods decreased for $\beta = 0.3$ and 0.6, and also bias increased compared to Table 4. The intuitive reason for this change is that though all subpopulations follow standard assumption of population genetics such as Hardy-Weinberg equilibrium, a pooled sample from many subpopulations violate that assumption. In addition, since these subpopulations have different risks of disease, the subpopulation membership acts as a confounder (Kleinbaum et al. 1982). Therefore haplotype-disease association is over estimated for not properly accounting the population structure. However, Method-V produces less biased estimate than the other alternative procedures.

The fact is that there are lot of uncertainties in the haplotype frequencies (it is neither homogeneous across the matched sets nor within a matched set), and that uncertainty has taken into account in the estimation of β by employing the ECM algorithm of simultaneous estimation. The right panel of Table 6 presents the simulation results when the haplotype frequency varies across the subpopulations but the disease risk model remains the same. Compared to the left panel of the table, the amount of bias is significantly less in this situation.

In summary we say that each of the four Methods II, III, IV, and V has some advantages and disadvantages, and under certain assumptions one is better than the others. However, in Tables 3, 4, and 5 MSE due to Method V is smaller than that of Methods II and III, and in many instances the MSE due to Method V is smaller than that of Method IV. Also, in Tables 4 and 5, among the methods II, III, IV, and V, Method V has uniformly better power than the alternatives procedures. Under the violation of model assumptions all the methods are affected, however the proposed method seems to be least affected in terms of bias and MSE, and gain in power is observed for $\beta = 0.3$ and 0.6.

Finally, we study certain other robustness aspects of the methods. We simulate data according to the simulation scenario which corresponds to Table (4), but using a recessive model. However, we analyze the simulated data by assuming an additive model. The estimate and MSE are presented in Table (7). It is clearly seen that all the methods underestimate the β parameter, except the situation when the true value of β is zero. In an alternative scenario we simulate data from an additive model keeping everything else remain the same as before but analyze them using a recessive model. The results are presented in Table (8). It is obvious that all the methods overestimate the parameter except in the situation when the true parameter is zero, i.e. under the null model.

The proposed method and other methods did not converge for approx-

Method		$\beta = 0$	$\beta = 0.3$	$\beta = 0.6$	$\beta = 0$	$\beta = 0.3$	$\beta = 0.6$
I	EST	0.0733	0.3427	0.6743	0.0138	0.2919	0.5916
(known	SVAR	0.056	0.0548	0.0616	0.0618	0.0578	0.0540
phase)	AEV	0.0555	0.0529	0.0564	0.0566	0.0534	0.0554
	MSE	0.0614	0.0566	0.0671	0.0620	0.0579	0.0541
	CP	91.1	92.9	92.7	93.6	93.6	94.2
	PWR	6.7	34.7	83.6	6.4	25.7	73.1
II	EST	0.1152	0.3584	0.7053	0.0298	0.3057	0.5989
	SVAR	0.0750	0.0740	0.0819	0.0782	0.0725	0.0733
	AEV	0.0784	0.069	0.0782	0.0786	0.0753	0.0762
	MSE	0.0883	0.0774	0.0930	0.0791	0.0725	0.0733
	CP	91.5	92.6	92.0	93.6	93.9	94.9
	PWR	7.4	29.3	75.9	6.4	20.3	57.9
III	EST	0.1170	0.3606	0.7040	0.0349	0.306	0.6032
	SVAR	0.0760	0.0775	0.0848	0.0784	0.0736	0.0759
	AEV	0.0813	0.0734	0.0796	0.0827	0.0793	0.0804
	MSE	0.0897	0.0812	0.0956	0.0796	0.0736	0.0759
	CP	92.6	93.6	93.7	94.0	95.9	94.9
	PWR	6.4	28.9	73.5	6.0	22.0	57.2
IV	EST	0.1047	0.3597	0.7115	0.0172	0.3095	0.623
	SVAR	0.0818	0.0849	0.0824	0.0854	0.0808	0.0817
	AEV	0.0868	0.0767	0.0789	0.0885	0.0832	0.0837
	MSE	0.0928	0.0885	0.0948	0.0857	0.0809	0.0822
	CP	91.8	91.9	91.3	93.6	95.3	94.6
	PWR	7.2	26.7	75.3	6.4	18.9	61.3
V	EST	0.1061	0.3526	0.6790	0.0198	0.2998	0.5986
	SVAR	0.0767	0.0679	0.0713	0.0747	0.0673	0.0623
	AEV	0.0834	0.0722	0.0785	0.0818	0.0752	0.0771
	MSE	0.0880	0.0707	0.0775	0.0751	0.0673	0.0623
	CP	92.4	92.9	94.8	93.0	96.2	93.6
	PWR	6.6	33.7	78.6	6.1	22.3	64.3

Table 6: Results of the simulation study where the target population consists of three subpopulations with different haplotype frequencies. We assume HWE holds in each subpopulation. Left panel: the disease prevalence varies across the subpopulations. Right panel: the disease prevalence does not vary across the subpopulations.

Method		$\beta = 0$	$\beta = 0.3$	$\beta = 0.6$
I	EST	0.0049	0.0458	0.1038
	MSE	0.0590	0.1313	0.3104
II	EST	0.0050	0.0571	0.1114
	MSE	0.0730	0.1384	0.3106
III	EST	0.0072	0.0622	0.1175
	MSE	0.0748	0.1388	0.3063
IV	EST	-0.0097	0.0477	0.1090
	MSE	0.0700	0.1441	0.3179
V	EST	0.0045	0.0589	0.1114
	MSE	0.0782	0.1413	0.3012

Table 7: Results of the simulation study where the haplotype frequencies are given by column 2 of Table (3), and the HWE holds (i.e., f=0). Here data are simulated from a recessive model but they are analyzed by an additive model.

Method		$\beta = 0$	$\beta = 0.3$	$\beta = 0.6$
I	EST	0.0509	0.5474	1.0843
	MSE	0.6730	0.6708	0.6921
II	EST	0.0509	0.5476	1.0942
	MSE	0.6730	0.6722	0.6934
III	EST	0.0665	0.5481	1.0946
	MSE	0.6728	0.6688	0.6975
IV	EST	-0.1006	0.4421	0.9674
	MSE	0.4795	0.3523	0.3686
V	EST	0.0508	0.5473	1.0756
	MSE	0.6730	0.6695	0.6915

Table 8: Results of the simulation study where the haplotype frequencies are given by column 2 of Table (3), and the HWE holds. Here data are simulated from an additive model but they are analyzed by a recessive model.

imately $2-3\%$ datasets. This is only due to small sample size. For our computations, subroutines were written in `Fortran` and we used `nlm()` function of `R` for optimization purposes. The computer codes are available at `http://www.stat.tamu.edu/~sinha/research.html`.

7 Discussion

This article proposes a likelihood based approach for analyzing matched case-control data in presence of unphased genotype data. For the estimation of the parameter of interest, we develop an ECM algorithm which allows to estimate the disease-haplotype association parameter and the haplotype frequency simultaneously. Consequently it is easy to calculate the standard error of the parameter by using the observed information matrix. As always the proposed method relies on some sort of model assumptions. In order to estimate the diplotype frequency, we impose HWE only among the control population, and we do not assume that the disease is rare. The proposed likelihood properly accounts the sampling strategy of the data collection, thus the estimate should be free from any design bias. Though the simulation and data example focus only on the effect of haplotype, the proposed method can easily accommodate other covariate effect in the model. The proposed method presumed that controls are unrelated to the cases. Thus, for family based case-control studies, such as case-parent or case-sib studies, the method need to be modified to account for family-wise association.

As the proposed method is a likelihood based approach, one can easily calculate some model diagnostic statistics such as AIC or BIC to have a notion of goodness of the fitted model. For instance, in absence of prior knowledge of haplotype effect on the disease, one may fit additive, dominant, and recessive model, and check which one yields the best fit for the data.

The extensive simulation study shows that in terms of bias and MSE the proposed method outperforms the existing alternatives in almost all situations, and it has significantly better power in many situations. With a moderate degree of model violation, the proposed method works quite well in terms of bias and MSE. Hence the method is robust under departures from model assumptions. Robustness is always an issue of the likelihood based approach which uses the HWE, hence in order to make the inference free from the HWE assumption, one may adopt the strategy of Zhao et al. (2003) for estimating the haplotype frequencies.

All the methods essentially assume that haplotype frequencies do not vary across the matched sets, however it is likely that the frequencies may vary across the matched sets if the matched sets are coming from different ethnic, or racial groups. Even if all subjects come from the same ethnic class, they might have different origins which may result in varying genetic structure within a population. The concern is then how to handle this heterogeneity in the genetic structure of the population. Although, through simulation study we demonstrate the robustness of the proposed method, a more rigorous approach is needed to handle the issues related to population stratification.

References

Breslow, N. E. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association* **91**, 14–28.

Chen, J. and Chatterjee, N. (2006). Haplotype-based association in cohort and nested case-control studies. *Biometrics* **62**, 28–35.

Clark, A. G. (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular Biology and Evolution* **7**, 111–122.

Excoffer, L. and Slatkin, M. (1995). Maximum likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution* **12**, 921–927.

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.

Fallin, D. and Schork, N. (2000). Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diplotype data. *American Journal of Human Genetics* **67**, 947–959.

Greenspan, G. and Geiger, D. (2006). Modeling Haplotype Block Variation Using Markov Chains. *Genetics* **172**, 2583-2599.

Kleinbaum, D. G., Kupper, L., and Chambless, L. E. (1982). Logistic regression analysis of epidemiologic data: Theory and practice. *Communications in Statistics: Theory and Methods* **11**, 485–547.

Kraft, P. K. and Thomas, D. C. (2000). Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. *American Journal of Human Genetics* **66**, 1119-1131.

Kraft, P. K., Cox, D. G., Paynter, R. A., Hunter, D. and DeVivo, I. (2005). Accounting for haplotype uncertainty in matched association studies: A comparison of simple and flexible technique. *Genetic Epidemiology* **28**, 261–272.

Lin, D. Y. and Zeng, D. (2006). Likelihood-based inference on haplotype effects in genetic association studies. *Journal of the American Statistical Association* **101**, 89–104.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **44**, 226–233.

Meng, X l. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80**, 267–278.

Niu, T., Qin, Z. S., Xu, Xiping., and Liu, j. S. (2002). Bayesian haplotype inference for multiple linked single-nucleotide polymorphism. *American Journal of Human Genetics* **70**, 157–169.

Pujana, A., Han, J-D. J., Starita, L. M., Stevens, K. N., Tewari, M., Ahn, J. S., Rennert, G., Moreno, V., Kirchhoff, T., Gold, B., Assmann, V., ElShamy, W. M., Rual, J-F., Levine, D., Rozek, L. S., Gelman, R. S., Gunsalus, K. C., Greenberg, R. A., Sobhian, B., Bertin, N., Venkatesan, K., Ayivi-Guedehoussou, N., Sole, X., Hernández, P., Lázaro, C., Nathanson, K. L., Weber, B. L., Cusick, M. E., Hill, D. E., Offit, K., Livingston, D. M., Gruber, S. B., Parvin, J. D., and Vidal, M. (2007). Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nature Genetics* **39**, 1338–1349.

Qin, Z., Niu, T., Liu, J. (2002). Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *American Journal of Human Genetics* **70**, 157–169.

Satten, G. A. and Carroll, R. J. (2000). Conditional and unconditional categorical regression models with missing covariates. *Biometrics* **56**, 384–388.

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.

Satten, G. A. and Kupper, L. L. (1993). Inferences about exposure-disease associations using probability-of-exposure information. *Journal of the American Statistical Association* **88**, 200–208.

Tzeng, J.Y. and Roeder, K. (2006). Invited Discussion of Likelihood-Based Inference on Haplotype Effects in Genetic Association Studies by Lin and Zeng. *Journal of the American Statistical Association* **101**, 111–114.

Woodson, K., Tangrea, J. A., Lehman, T. A., Modali, R., Taylor, K. M., Snyder, K., Taylor, P. R., Virtamo, J., and Albanes, D. (2003). Manganese superoxide dismutase (MNSOD) polymorphism, alpha-tocopherol supplementation and prostate cancer risk in the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (Finland). *Cancer Causes and Controls* **14**, 513–518.

Zhang, H., Zheng, G., and Li, Z. (2006). Statistical analysis for haplotype-based matched case-control studies. *Biometrics* **62**, 1124–1131.

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.