Biometrics 64, 1–22 DOI: 10.1111/j.1541-0420.2005.00454.x
December 2008

Semiparametric estimation exploiting covariate independence in two-phase randomized trials

James Y. Dai [2,*] , Michael LeBlanc [1,2] and Charles Kooperberg [1,2]

[1] Department of Biostatistics, University of Washington, Seattle, Washington, U.S.A.

[2] Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, U.S.A.

email: ydai@fhcrc.org

Summary: Recent results for case-control sampling suggest when the covariate distribution is constrained by gene-environment independence, semiparametric estimation exploiting such independence yields a great deal of efficiency gain. We consider the efficient estimation of the treatment-biomarker interaction in two-phase sampling nested within randomized clinical trials, incorporating the independence between a randomized treatment and the baseline markers. We develop a Newton–Raphson algorithm based on the profile likelihood to compute the semiparametric maximum likelihood estimate (SPMLE). Our algorithm accommodates both continuous phase-one outcomes and continuous phase-two biomarkers. The profile information matrix is computed explicitly via numerical differentiation. In certain situations where computing the SPMLE is slow, we propose a maximum estimated likelihood estimator (MELE), which is also capable of incorporating the covariate independence. This estimated likelihood approach uses a one-step empirical covariate distribution, thus is straightforward to maximize. It offers a closed-form variance estimate with limited increase in variance relative to the fully efficient SPMLE. Our results suggest exploiting the covariate independence in two-phase sampling increases the efficiency substantially, particularly for estimating treatment-biomarker interactions.

Key words: case-only estimator; estimated likelihood; gene-environment independence; Newton–Raphson algorithm; profile likelihood; treatment-biomarker interactions.

1. Introduction

In clinical and epidemiological studies, the effect of an intervention is often influenced by variables reflecting individual susceptibility. In pharmacogenetic studies, there is a growing body of evidence supporting interactions between genetic polymorphisms and antihypertensive treatments (Arnett et al., 2005). Identifying the effect-modifying genotypes, or other types of biomarkers helps to disclose the etiology of diseases, and to understand the mechanism of the intervention effect. Since bioassays are often expensive and there may be many candidate markers, it is common to measure biomarkers only in a case-control sample from the study cohort (Breslow and Day, 1980), or in a stratified case-control sample if additional covariates are involved in forming the strata (White, 1982; Scott and Wild, 1991). This often constitutes a two-phase outcome-dependent sampling design, in that the first-phase data contain the response ascertained for every subject, and perhaps a collection of "cheap" covariates (for example, treatment assignment, demographic factors); the second-phase data contain the biomarker data for a case-control subsample. If the phase-one cohort is a randomized clinical trial, the treatment assignment is independent of the phase-two biomarkers measured from baseline-stored blood. This paper pertains to the potential efficiency gain when exploiting this independence in two-phase randomized trials.

Statistical methods for two-phase sampling have been studied by many authors. When the covariate partially observed in the second phase is discrete, Ibrahim (1990) uses a weighted EM algorithm to estimate the parameters in a generalized linear model. Difficulty arises when the missing covariates are continuous: numerical integration or Monte Carlo methods are needed when a parametric covariate distribution is assumed (Ibrahim, Chen and Lipsitz, 1999). To avoid model misspecification and ease the computation, a number of pseudo-likelihood methods have been proposed, including the inverse probability weighted estimator (Flanders and Greenland, 1991; Lipsitz, Ibrahim and Zhao, 1999), conditional likelihood (Breslow and

Cain, 1988), estimated likelihood (Pepe and Fleming, 1991; Carroll and Wand, 1991), mean score estimators (Reilly and Pepe, 1995), and pseudoscore estimators (Chatterjee, Chen and Breslow, 2003). These approaches yield a consistent estimator of the regression parameters, yet they are not efficient in general. Robins, Rotnitzky and Zhao (1994) introduced a class of semiparametric estimators based on inverse probability weighted estimating equations, and obtained an efficient estimator that attains the semiparametric variance bound. However, the implementation is difficult. When the first-phase data can be reduced to discrete stratum labels, a profile likelihood approach has been proposed to obtain the semiparametric maximum likelihood estimates, with the covariate distributions left completely nonparametric (Scott and Wild, 1997; Lawless, Kalbfleisch and Wild, 1999).

Recent work in case-control studies suggests that exploiting the gene-environment independence improves the estimation efficiency of regression parameters substantially (Chatterjee and Carroll, 2005; Chatterjee and Chen, 2007). Moreover, for rare diseases, the gene-environment interaction in a logistic regression can be estimated by the odds ratio between the gene and the environmental variable in cases only (Piegorsch, Weinberg and Taylor, 1994; Umbach and Weinberg, 1997). Despite the efficiency advantage, these analyses are generally sensitive to departures from the gene-environment independence assumption (Albert et al., 2001). In two-phase sampling nested within randomized trials, however, there exists indisputable independence between the treatment and baseline covariates by design, including markers ascertained in the phase-two sample. Ignoring such design-based independence is a waste of information. We present two examples that motivate our research:

Example 1: The randomized clinical trial of estrogen plus progestin in the Women's Health Initiative (WHI) was terminated early in 2002 because of an increased risk of stroke, breast cancer, and cardiovascular diseases in the treatment arm (Rossouw et al., 2002). To determine whether the adverse effect of conjugated equine estrogen and medroxyprogesterone acetate on

stroke was modified by selected baseline blood biomarkers and genotypes, WHI analyzed baseline blood samples from cases and controls. Twenty-nine biomarkers were measured, encompassing inflammatory markers, lipid levels, thrombosis factors, blood cell counts and several Single Nucleotide Polymorphisms (SNPs). All markers except the SNPs yield continuous measurements. More recently in a similar attempt, the WHI is undertaking a genome-wide association study in which hundreds of thousands of SNPs are sequenced in a sample.

Example 2: The Genetics of Hypertension-Associated Treatment (GenHAT) Study is a large-scale, double-blind, randomized trial attempting to test the interaction between the insertion/deletion polymorphism in antiotension-converting enzyme and four antihypertensive treatments (Arnett et al., 2005). No significant association was found. In a hypothetical secondary study, a number of SNPs on several genes in renin-angiotensin-aldosterone system are further investigated. The outcome is the blood pressure (BP) change after 6 months treatment. While the outcome and the treatment are collected for every participant, genetic variants are only measured in 3 outcome-dependent subsamples: one from the stratum with BP change lower than 10% percentile, one from the stratum with BP change higher than 90% percentile, and one from the stratum with the rest of participants.

Both studies employ two-phase sampling to identify the effect-modifying biomarkers. The independence between the treatment and the biomarkers is dictated by randomization. The outcome of interest can be categorical (stroke in **Example 1**), or continuous (BP change in **Example 2**). In fact many clinical outcomes are continuous, such as cholesterol levels, HIV viral load, and child's IQ. It is important to account for continuous outcomes in analysis. In addition, biomarkers can be categorical (genotype) or continuous (blood assay); and as many as hundreds of thousands may be measured (**Example 1**). To our knowledge, the efficient estimation exploiting covariate independence in two-phase randomized trials has not been addressed. Chatterjee and Chen (2007) extended the profile technique to allow gene-

environment independence in studies with two-phase sampling. However, their method only considers binary outcomes. For continuous outcomes, the profile approach often entails a substantial loss of efficiency, since it has to reduce the outcome to discrete stratum labels (Chatterjee, Chen and Breslow, 2003). Moreover, computing the variance matrix when exploiting covariate independence using the profile technique can be algebraically cumbersome.

In this article, we propose two semiparametric methods that exploit covariate independence in two-phase sampling. They are semiparametric since the distribution of the missing covariates is treated nonparametrically, while the association between the outcome and covariates remains parametric. Both methods use full information from continuous outcomes and provide a straightforward computation of variance estimates. We first develop a profile likelihood based Newton–Raphson algorithm that can be used to compute the semiparametric maximum likelihood estimate (SPMLE). The independence between covariates is incorporated transparently. The novelty is that, instead of replacing the high dimensional nuisance parameters by a few low dimensional parameters as in Scott and Wild (1997), we profile out the nuisance parameters completely, and explicitly compute the information matrix through numerical differentiation, thus generating the variance as a by-product. When very many biomarkers are investigated, computing the SPMLE for every marker may be time-consuming and unnecessary. This is particularly the case if the disease is relatively rare and there are many covariates to be adjusted. We develop a maximal estimated likelihood estimator (MELE) that is much faster to compute and accounts for the covariate independence, yet it does not lose much efficiency. In essence it plugs a consistent empirical estimator of the covariate distribution in the likelihood. The amount of variance reduction by imposing the independence can be derived explicitly.

In Section 2, we define the sampling scheme and likelihood. In Section 3, we describe the profile Newton–Raphson algorithm that computes the SPMLE and the estimated variances. In Section 4, we derive the estimated likelihood and asymptotic distribution of the MELE. In

Section 5, we assess the finite sample properties of the proposed estimators in simulations. In Section 6, we show an application to a biomarker dataset from WHI. We end with a discussion. Technical details are available as on-line supplementary materials.

2. Sampling scheme and likelihood

Let Y be an outcome of interest, X be the treatment that is randomized, and Z be a collection of covariates which includes the expensive biomarker, and potentially other important predictors. Throughout this work, Y and Z can be continuous or categorical, but X is assumed to be categorical. Suppose without missing data, N subjects with i.i.d. random variables (Y_i, X_i, Z_i), are generated from the joint probability density $f_\beta(Y|X,Z)g(X,Z)$, where $f_\beta(Y|X,Z)$ is the parametric regression model with parameters β, which often takes the form of a generalized linear model; $g(X,Z)$ is the joint density function for (X,Z). We assume sampling takes place so that only a subset of subjects have Z measured. Let R_i denote the indicator of whether a subject has complete data. The observed data, therefore, contain (Y_i, X_i, Z_iR_i, R_i), $i = 1, ...N$. We assume that $\Pr(R_i = 1|Y_i, X_i, Z_i) = \Pr(R_i = 1|Y_i, X_i)$, that is, Z is missing at random (MAR) in the sense of Rubin (1976). Let $\mathcal{Y}$ and $\mathcal{X}$ be the sample spaces of the random variable Y and X. Let $\{\mathcal{S}_k\}$, $k = 1, ..., K$, be K mutually exclusive partitions of $\mathcal{Y} \times \mathcal{X}$ so that $\mathcal{Y} \times \mathcal{X} = \cup_{k=1}^{K} \mathcal{S}_k$. For a binary outcome, $\mathcal{S}_k$ may be solely defined by case-control status. For a continuous outcome such as birth weight, categorization of outcome by quantiles may be involved. Subjects are inspected sequentially as they arise from the joint density and the (Y, X) are observed. When $(y_i, x_i) \in \mathcal{S}_k$, the i^{th} subject is selected for observing Z with prespecified positive probabilities p_k, hence, $\Pr(R_i = 1|Y_i, X_i) = \sum_{k=1}^{K} p_k 1_{[(y_i,x_i)\in\mathcal{S}_k]}$. This is i.i.d. Bernoulli sampling (Lawless, Kalbfleisch and Wild, 1999).

Let $V = \{i : R_i = 1\}$ and $\bar{V} = \{j : R_j = 0\}$ be the sets containing subjects with complete and incomplete data, respectively. The likelihood for those with missing Z involves integration of $f_\beta(y|x,Z)$ by $dG(Z|x)$, where $G(Z|x)$ is the conditional cumulative distribution function of

Z given X. Parametric modeling of G is subject to model misspecification. A semiparametric approach is to treat G nonparametrically, that is, maximizing G over distributions whose support consists of the observed z. For well behaved univariate $g(Z)$, smoothing methods with appropriately selected bandwidths can estimate $g(Z)$ more efficiently than estimating $g(Z)$ completely nonparametrically; yet nonparametric estimation provides flexibility and robustness given that $g(Z)$ is often not of interest in inference. Let $\mathcal{Z}$ be the set of z in the observed sample space, and let $\mathcal{Z}_x$ be the restricted set of observed z with $X = x$. This leads to an empirical likelihood which contains the point mass $g_{z|x} = \Pr(Z = z|X = x)$ for $z \in \mathcal{Z}_x$, with the constraint $\sum_{z\in\mathcal{Z}_x} g_{z|x} = 1$. Note that Z can be a collection of covariates, so that $g(z)$ is a point mass on a combination of several covariate values. If $X \perp Z$, the conditional part in the point mass vanishes so that $g_{z|x} = g_z = \Pr(Z = z)$ for $z \in \mathcal{Z}$. The semiparametric likelihoods without and with using the independence $X \perp Z$ are

$$L(\beta, G) = \prod_{i\in V} f_\beta(y_i|x_i, z_i) g_{z_i|x_i} \prod_{j\in\bar{V}} \left(\sum_{z_l\in\mathcal{Z}_{x_j}} g_{z_k|x_j} f_\beta(y_j|x_j, z_l) \right), \quad (1)$$

$$L^\perp(\beta, G) = \prod_{i\in V} f_\beta(y_i|x_i, z_i) g_{z_i} \prod_{j\in\bar{V}} \left(\sum_{z_l\in\mathcal{Z}} g_{z_l} f_\beta(y_j|x_j, z_l) \right), \quad (2)$$

respectively. Throughout this article, we use the superscript $^\perp$ to indicate expressions that employ the independence between X and Z. Intuitively, imposing independence shrinks the dimension of the nuisance parameter G, since we only need to estimate the marginal distribution of Z, thereby improving the estimation of β. Note that the dimension of $g_{z|x}$ and g_z increases with the number of the phase-two subjects for continuous Z. Maximization of g and β simultaneously using an EM algorithm was considered in Lawless (1997), though the convergence is extremely slow if the dimension of g is large and the proportion of missing data is large. Also, when Y is continuous, the existing profile likelihood method has to categorize Y into strata, and therefore loses information (Lawless, Kalbfleisch and Wild, 1999).

3. A profile likelihood based Newton–Raphson algorithm to compute SPMLE

To maximize likelihoods with a Euclidian parameter θ, a Newton–Raphson algorithm iteratively updates $\hat{\theta}$ by $\theta^{(m+1)} = \theta^{(m)} + I(\theta^{(m)})S(\theta^{(m)})$ until convergence, where $S(\theta^{(m)})$ is the score function and $I(\theta^{(m)})$ is the observed information evaluated at the current $\theta^{(m)}$. Near the solution the convergence of Newton–Raphson algorithm is always fast (exponential), and it automatically leads to variance estimates. When maximizing the semiparametric likelihoods (1) and (2), the interest is in inference on β, not in the infinite dimensional nuisance parameter g. A profile likelihood can be derived in which g is maximized first for a fixed β, then maximized with respect to β using a Newton–Raphson algorithm.

Specifically, let $\ell_p(\beta, \hat{g}_\beta)$ denote the profile log likelihood for β, where $\hat{g}_\beta$ is the maximizer of g given β. Let $S_p(\beta, \hat{g}_\beta) = \partial \ell_p(\beta, \hat{g}_\beta)/\partial\beta$ and $I_p(\beta, \hat{g}_\beta) = \partial S_p(\beta, \hat{g}_\beta)/\partial\beta$. The Newton–Raphson algorithm iterates the following steps: (1) Given $\beta^{(m)}$, compute $\hat{g}_{\beta^{(m)}}$. (2) Compute $S_p(\beta^{(m)}, \hat{g}_{\beta^{(m)}})$. (3) Compute $I_p(\beta^{(m)}, \hat{g}_{\beta^{(m)}})$ via numerical differentiation. (4) Update β by $\beta^{(m+1)} = \beta^{(m)} + I_p(\beta^{(m)}, \hat{g}_{\beta^{(m)}})S_p(\theta^{(m)}, \hat{g}_{\beta^{(m)}})$. Go to step (1).

This algorithm relies on a fast computation of $\hat{g}_{\beta^{(m)}}$ for any fixed $\beta^{(m)}$. By introducing a Lagrange multiplier that respects the fact that g sum to 1, we can show that $\hat{g}_\beta(z|x)$ satisfies

$$\sum_{i\in V} 1_{[x_i=x, z_i=z]} + \sum_{j\in\bar{V}} \frac{\sum_{z_k\in\mathcal{Z}_x} f_\beta(y_j|x_j, z_k)g(z_k|x_j)1_{[x_j=x, z_k=z]}}{\sum_{z_k\in\mathcal{Z}_x} g(z_k|x)f_\beta(y_j|x_j, z_k)1_{[x_j=x]}} = N_x g(z|x), \tag{3}$$

where $N_x = \sum_{l=1}^{N} 1_{[x_l=x]}$, the total number of subjects with the covariate value x. Note that X is discrete and observed for everyone. The second term of (3) on the left hand side is essentially $g(z|y_j, x_j = x)$, hence the left hand side and the right hand side are both the expected number of subjects with covariate value (x, z) in the phase-one data. Solving (3) for $\hat{g}(z|x)$ is not immediate, because the denominator of the second term in (3) involves all $g(z_k|x)$. However, note that $\sum_{z_k\in\mathcal{Z}_x} g(z_k|x)f_\beta(y_j|x, z_k) = f(y_j|x)$, a quantity that can be approximated using the phase-one data. Let $\hat{f}^0(y_j|x_j)$ be the estimated probability of y_j given x_j in the phase-one

data. In the on-line supplemental material we describe a fast computation of $\hat{g}_\beta(z|x)$ based on the approximation by $\hat{f}^0(y_j|x_j)$.

After $\hat{g}_\beta(z|x)$ is computed, $S_p(\beta, \hat{g}_\beta)$ is readily obtainable. Observe that

$$S_p(\beta, \hat{g}_\beta) = \left\{ \sum_{i=1}^{N} \frac{\partial}{\partial \hat{g}_\beta} \ell(\beta, \hat{g}_\beta | x_i, y_i, r_i z_i) \right\} \frac{\partial \hat{g}_\beta}{\partial \beta} + \sum_{i=1}^{N} \frac{\partial}{\partial \beta} \ell(\beta, \hat{g}_\beta | x_i, y_i, r_i z_i, \hat{g}_\beta). \quad (4)$$

Since $\hat{g}_\beta$ is the maximizer for every fixed β, $\sum_{i=1}^{N} \frac{\partial}{\partial \hat{g}_\beta} \ell(\beta, \hat{g}_\beta | x_i, y_i, r_i z_i) = 0$. Hence the first term of (4) equals to 0. Therefore

$$S_p(\beta, \hat{g}_\beta) = \sum_{i \in V} S(y_i|x_i, z_i) + \sum_{j \in \bar{V}} \sum_{z_k \in \mathcal{Z}_{x_j}} \frac{\hat{g}_\beta(z_k|x_j) f_\beta(y_j|x_j, z_k)}{\sum_{z_k \in \mathcal{Z}_{x_j}} \hat{g}_\beta(z_k|x_j) f_\beta(y_j|x_j, z_k)} S(y_j|x_j, z_k). \quad (5)$$

Once the profile score is computed, we use numerical differentiation to approximate the profile information matrix:

$$I_p(\beta, \hat{g}_{\beta^{(m)}}) = \frac{\partial S_p(\beta^{(m)}, \hat{g}_{\beta^{(m)}})}{\partial \beta^{(m)}} = \frac{S_p(\beta^{(m)} + \epsilon, \hat{g}_{\beta^{(m)}+\epsilon}) - S_p(\beta^{(m)} - \epsilon, \hat{g}_{\beta^{(m)}-\epsilon})}{2\epsilon}.$$

This involves perturbing each element of $\beta^{(m)}$ in both directions, computing two new $\hat{g}_\beta$, and evaluating their profile scores. A highly accurate I_p can be achieved with ϵ in the order of $1/n$.

When $X \perp Z$, we only need to maximize g on the pooled sample space $\mathcal{Z}$, rather than on the restricted sample space $\mathcal{Z}_x$. We can now compute $\hat{g}^\perp_\beta(z|x)$ using

$$\hat{g}^\perp_\beta(z|x) = \hat{g}_\beta(z) \approx \frac{\sum_{i \in V} 1_{[z_i = z]}}{N - \left(\sum_{j \in \bar{V}} \sum_{z_k \in \mathcal{Z}} \frac{f_\beta(y_j|x_j, z_k) 1_{[z_k = z]}}{\sum_{z_k \in \mathcal{Z}} \hat{g}_\beta(z_k) f_\beta(y_j|x_j, z_k)} \right)},$$

and (5) becomes

$$S^\perp_p(\beta, \hat{g}_\beta) = \sum_{i \in V} S_\beta(y_i|x_i, z_i) + \sum_{j \in \bar{V}} \sum_{z_k \in \mathcal{Z}} \frac{\hat{g}_\beta(z_k) f_\beta(y_j|x_j, z_k)}{\sum_{z_k \in \mathcal{Z}} \hat{g}_\beta(z_k) f_\beta(y_j|x_j, z_k)} S_\beta(y_j|x_j, z_k). \quad (6)$$

A slight modification of the Newton–Raphson algorithm suffices.

Starting from a naive estimator of β, for example the inverse probability weighted estimator (Flanders and Greenland, 1991; Lipsitz, Ibrahim and Zhao, 1999), the profile Newton–Raphson algorithm usually takes 3-4 iteration to achieve 1e-5 accuracy. At convergence, we obtain the variance estimates of $\hat{\beta}$ by inverting the information matrix of the profile likelihood $I_p(\hat{\beta}, \hat{g}_{\hat{\beta}})$ as a by-product of the algorithm. When Z is discrete, the parameter space (β, g)

is of fixed dimension. The usual large sample theory for maximum likelihood estimates applies with standard regularity conditions (Cox and Hinkley, 1974, Chapter 9). When Z contains continuous covariates, the parameter space (β, g) is of infinite dimension, so that modern semiparametric inference theory is required to prove the consistency and asymptotic normality. The proofs of asymptotic theories follow Murphy and van der Vaart (2000); Breslow, Robins and Wellner (2003) and they are not presented here.

4. Estimated likelihood

The main computational burden of obtaining the SPMLE lies in updating $\hat{g}_{\hat{\beta}}$ and the numerical differentiation to get $I_p(\hat{\beta}, \hat{g}_{\hat{\beta}})$. When the disease is rare, updating $\hat{g}_{\hat{\beta}}$ can be time-consuming; when there are many covariates to be adjusted, numerical differentiation can slow down the algorithm. Much computation may not be needed in a genome-wide study where most markers do not exhibit signal. An estimated likelihood approach may be a good alternative to exploit the independence in these situations. Pepe and Fleming (1991) propose an estimated likelihood approach which first plugs an empirical estimator $\widehat{G}(Z|x)$ into the likelihood for the incomplete data, i.e., $\hat{f}_\beta(y|x) = \int f_\beta(y|x, Z) d\widehat{G}(Z|x)$, and then maximizes the estimated likelihood solely with respect to β. Robins, Rotnitzky and Zhao (1994) and Lawless, Kalbfleisch and Wild (1999) compared a variety of methods used in two-phase studies. The estimated likelihood approach was found to perform closely to the SPMLE in efficiency. Pepe and Fleming (1991) assumes the validation sample is a random sample from the cohort, i.e., missing completely at random (MCAR). Weaver and Zhou (2005) extends the methodology to outcome-dependent sampling schemes, though they consider a slightly different scenario of a simple random sample (SRS) in addition to the outcome-dependent sample. Here we derive the estimated likelihood based estimator under the two-phase sampling scheme as specified in Section 2.

When the phase-two sampling is outcome-dependent, a consistent estimator of $G(z|x)$ can

be formulated as a weighted average of the empirical distribution of $Z|x$ in each stratum $\mathcal{S}_k$ (Hu and Lawless, 1996; Lawless, Kalbfleisch and Wild, 1999). Observe that

$$G(z|x) = \sum_{k=1}^{K} \Pr(Z < z|\mathcal{S}_k, x)\Pr(\mathcal{S}_k|x).$$

Because the probability of observing Z is assumed constant in each stratum, we observe that $\Pr(Z < z|\mathcal{S}_k, x) = \Pr(Z < z|\mathcal{S}_k, x, R = 1)$.

Therefore, we obtain an empirical estimate of $G(z|x)$ using Z in the validation data,

$$\widehat{G}(z|x) = \sum_{k=1}^{K} \frac{N_{kx}}{N_x} \sum_{i\in\mathcal{S}_k} \frac{1_{[z_i<z, x_i=x, r_i=1]}}{\sum_{i\in\mathcal{S}_k} 1_{[x_i=x, r_i=1]}},$$

where $N_{kx} = \sum_{i=1}^{N} 1_{[i\in\mathcal{S}_k, x_i=x]}$, $N_x = \sum_{i=1}^{N} 1_{[x_i=x]}$. The estimated likelihood for incomplete data becomes

$$\hat{f}_\beta(y|x) = \sum_{k=1}^{K} \frac{N_{kx}}{N_x} \sum_{i\in\mathcal{S}_k} \frac{f_\beta(y|x, z_i)1_{[x_i=x, r_i=1]}}{\sum_{i\in\mathcal{S}_k} 1_{[x_i=x, r_i=1]}}.$$

Let $\hat{L}_N(\beta)$ denote the estimated likelihood. We want to maximize

$$\hat{L}_N(\beta) = \prod_{i\in V} f_\beta(y_i|x_i, z_i) \prod_{j\in\bar{V}} \hat{f}_\beta(y_j|x_j). \tag{7}$$

Using $X \perp Z$ we can improve the estimation of the likelihood. Following the same derivation for $\widehat{G}(z|x)$ with $X \perp Z$, the estimated empirical distribution is simplified to

$$\widehat{G}^{\perp}(z|x) = \widehat{G}(z) = \sum_{k=1}^{K} \frac{N_k}{N} \sum_{i\in\mathcal{S}_k} \frac{1_{[z_i<z, r_i=1]}}{\sum_{i\in\mathcal{S}_k} 1_{[r_i=1]}},$$

where $N_k = \sum_{i=1}^{N} 1_{[i\in\mathcal{S}_k]}$. Essentially we are able to use all observed Z to estimate the empirical distribution, not constrained by a particular x. Hence the estimated likelihood becomes

$$\hat{f}^{\perp}_\beta(y|x) = \sum_{k=1}^{K} \frac{N_k}{N} \sum_{i\in\mathcal{S}_k} \frac{f_\beta(y|x, z_i)1_{[r_i=1]}}{\sum_{i\in\mathcal{S}_k} 1_{[r_i=1]}}.$$

It is immediate that applying the independence assumption reduces the variability of the estimated likelihood, $\mathrm{Var}[\hat{f}^{\perp}_\beta(y_j|x_j)] < \mathrm{Var}[\hat{f}_\beta(y_j|x_j)]$, because the former involves an average of more terms. The estimated likelihood using the independence assumption is

$$\hat{L}^{\perp}(\beta) = \prod_{i\in V} f_\beta(y_i|x_i, z_i) \prod_{j\in\bar{V}} \hat{f}^{\perp}_\beta(y_j|x_j). \tag{8}$$

Let $\tilde{\beta}$ denote the estimator maximizing (7), and $\tilde{\beta}^{\perp}$ the estimator maximizing (8). We describe

the asymptotic properties of $\tilde{\beta}^{\perp}$ in the following two theorems. The derivation of the large sample properties is somewhat similar to Weaver and Zhou (2005). The emphasis here is on the efficiency gain when using $\tilde{\beta}^{\perp}$ instead of $\tilde{\beta}$. Let ρ_V be the probability of a subject falling in the validation sample, and let ρ_k be the probability of a subject falling in the stratum k. We assume ρ_V and ρ_k are strictly positive, so that $\sum_{i=1}^{N} 1_{[r_i=1]}/N \rightarrow \rho_V > 0$ and $N_k/N \rightarrow \rho_k > 0$ for every k. Let $\{x\}$ be the set of unique values attainable by X. With the main assumptions of missing at random and the covariate independence ($X \perp Z$), we derive the following theorems:

Theorem 1. (consistency) $\tilde{\beta}$ and $\tilde{\beta}^{\perp}$ are both consistent w.r.t. true parameter β.

Theorem 2. (Asymptotic Normality)

$$\begin{aligned}
\sqrt{N}(\tilde{\beta} - \beta) &\rightarrow_d N\left(0, I(\beta)^{-1} + I(\beta)^{-1}\Sigma I(\beta)^{-1}\right). \\
\sqrt{N}(\tilde{\beta}^{\perp} - \beta) &\rightarrow_d N\left(0, I(\beta)^{-1} + I(\beta)^{-1}\Sigma^{\perp} I(\beta)^{-1}\right),
\end{aligned}$$

where

$$\begin{aligned}
I(\beta) &= \rho_V E[-\frac{\partial^2 \log f_\beta(Y|X,Z)}{\partial \beta^2}] + (1-\rho_V)E[-\frac{\partial^2 \log f_\beta(Y|X)}{\partial \beta^2}], \\
\Sigma &= \sum_{k=1}^{K}\sum_{\{x\}} \frac{\Pr(\mathcal{S}_k, X=x)\left[\Pr(R=0|X=x)\right]^2}{p_k} \mathrm{Var}_{\boldsymbol{Z}}\left[\mathrm{E}_{\boldsymbol{Y}}[W|x,z,R=0]|x,z\in\mathcal{S}_k\right], \\
\Sigma^{\perp} &= \sum_{k=1}^{K} \frac{\rho_k(1-\rho_V)^2}{p_k} \mathrm{Var}_{\boldsymbol{Z}}\left[\mathrm{E}_{\boldsymbol{Y},\boldsymbol{X}}[W|z,R=0]|z\in\mathcal{S}_k\right], \\
W &= \frac{\partial f_\beta(Y|X,Z)/\partial\beta}{f_\beta(Y|X)} - \frac{\partial f_\beta(Y|X)/\partial\beta}{f_\beta^2(Y|X)} f_\beta(Y|X,Z).
\end{aligned}$$

We assume the usual regularity conditions for maximum likelihood holds for $f_\beta(Y|X,Z)$ and $f_\beta(Y|X)$ (Cox and Hinkely, 1974, Chapter 9) and that the sampling probability in each stratum is strictly positive. The proof of existence, uniqueness and consistency of an estimated likelihood estimator follows the results of Foutz (1977). The key step is that the second derivative of the estimated likelihood converges to a positive definite information matrix $I(\beta)$, i.e., $-\frac{1}{N}\frac{\partial^2 \log \hat{L}(\beta)}{\partial \beta^2} \rightarrow_p I(\beta)$. Consistent estimators for $I(\beta)$, Σ and $\Sigma^{\perp}$ can be formulated using empirical terms. See on-line supplementary material. Unlike the SPMLE in Section 3, $\hat{I}(\hat{\beta})$ can

be computed explicitly. Therefore $\tilde{\beta}$ and $\tilde{\beta}^{\perp}$ can be obtained much faster than the SPMLE. Clearly $\Sigma^{\perp} < \Sigma$. The asymptotic variance reduction when exploiting the independence is $I(\beta)^{-1}(\Sigma - \Sigma^{\perp})I(\beta)^{-1}$.

5. Simulation

We conducted a series of simulations to evaluate the proposed estimators, and to investigate the efficiency gain when exploiting covariate independence. We consider a two-phase sampling scheme with the following features: the outcome Y may be binary or continuous; a binary covariate X indexing the treatment assignment, which takes the distribution form of Bernoulli(0.5); both Y and X are observed for everyone, thus form the phase-one data; in the second phase, a case-control sample is identified from the cohort by Bernoulli sampling, independent of X. A continuous biomarker is measured for subjects in the case-control sample; In all simulations, X is independent of Z.

5.1 *Binary outcome*

We generated data for 10,000 subjects using the logistic model

$$\text{logit}(\Pr(y = 1|x, z)) = \exp(\beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz)$$

Set $\beta_0 = -3$, $\beta_1 = 0.2$, $\beta_2 = 0.1$, and vary β_3 to achieve different amounts of interaction between X and Z. We generated the biomarker data (Z) from a log normal distribution with a ceiling of 10 for Z, that is, $Z \sim \min(e^{N(0,1)}, 10)$. The Bernoulli sampling probabilities for cases and controls are such that on average 800 cases and 800 controls were selected; 5,000 datasets are simulated. Table 1 summarizes the biases and the sample variances of the various estimators. The complete-case estimator (CC) only uses the subjects with complete data in a logistic regression, ignoring the sampling. This estimator is consistent for the slope parameters (Prentice and Pyke, 1979). The weighted estimator (WE) uses the estimated sampling probabilities from the 4 strata defined by Y and X to weight the estimation equa-

tions. SPMLE and SPMLE exploiting independence (SPMLE$^{\perp}$) are computed as in Section 3. MELE and MELE$^{\perp}$ are computed as in Section 4. All estimators are unbiased for the three slope parameters. The efficiency is relative to the SPMLE ignoring independence. Among the methods ignoring independence, the SPMLE achieves the lowest variance. The complete-case estimator yields the same efficiency as SPMLE in estimating β_2 and β_3, but it has a much lower efficiency in estimating β_1 since it does not utilize the phase-one data. The WE is the worst in terms of estimating the interaction, while the MELE is very close to the SPMLE in efficiency, especially when β_3 is small. The trend observed here is consistent with Lawless, Kalbfleisch and Wild (1999). Interestingly, when using the independence (MELE$^{\perp}$ and SPMLE$^{\perp}$), the sampling variances drop markedly. The efficiency gain in estimating β_3 ranges from over 100% ($\beta_3 = 0$) to 20% ($\beta_3 = 1$). The efficiency gain in the main effects are 10%-50% depending on the effect size. In the parameter values we considered, the estimated likelihood approach performs closely to the semiparametric likelihood approach. We expect the relative efficiency of the MELE to decrease when parameter values get larger (Lawless, Kalbfleisch and Wild, 1999).

[Table 1 about here.]

Table 2 evaluates the validity of the estimated variances for the SPMLE, MELE, SPMLE$^{\perp}$ and MELE$^{\perp}$ based on 5000 simulations. The mean of the estimated variances agree very well to the corresponding sample variances. The empirical 95% coverage probabilities are close to the nominal 95%. We also studied many different parameter choices, as well as the situation where Z is categorical. The results are similar to those in Table 1 and 2 (results not shown).

[Table 2 about here.]

5.2 *Continuous outcome*

We simulated a continuous outcome variable $Y \sim N(\mu, \sigma^2)$, with $\mu = -1 + 0.2x + 0.1z + \beta_3 xz$, $\sigma^2 = 1$, X and Z as in Section 5.1. In the first phase, 2000 subjects are generated with X and Y observed. To create the phase-one strata, Y is divided at its 90th quantile. Individuals in the upper stratum are "cases", and those in the lower stratum are "controls". All cases and a random sample of about 200 controls enter the second phase and their Zs are measured. In Table 3 we compare the efficiencies of WE (using the estimated sampling probability), MELE, MELE$^{\perp}$, SPMLE, and SPMLE$^{\perp}$. We also include the SPMLE with reduced phase-one data by Lawless, Kalbfleisch and Wild (1999) (referred to as "LKW") to see how much information is lost in categorizing the phase-one continuous outcome. Note both the WE and the LKW reduce the phase-one data: the WE estimates the sampling probabilities using observed counts in each strata and uses it as if it is fixed in the weighted log-likelihood; the LKW treats the phase one data as stratum labels, but maximizes the likelihood with respect to β and g simultaneously, therefore the weight is updated in the iterations.

The biases of all estimators are negligible and therefore omitted in Table 3. As expected, the WE and the LKW have much bigger variances than the SPMLEs and the MELEs because the latter use more information from the first phase. Exploiting the independence between X and Z yields an additional efficiency gain, ranging from 10% to 50%. Consistent with the results in Table 1, the efficiency gain decreases with β_3. The estimated likelihood approach yields a second-best performance with respect to efficiency, next to the likelihood based methods. Note when $\beta_3 = 1$ the performance of WE is better than that of LKW. This may be caused by the relatively small sample size (2000 phase-one subjects, 400 phase-two subjects). To improve the performance of the LKW method, one can create more strata for the phase-one data (poststratification) hence gain more precision (Lawless, Kalbfleisch and Wild, 1999). However it will still be inferior to methods using continuous outcomes from the first place.

[Table 3 about here.]

6. Data application

We took the WHI biomarker study as in Example 1 to illustrate our methods. The aforementioned 29 biomarkers were picked by WHI investigators as markers that are possibly associated with either stroke, venous thrombotic disease, or myocardial infarction (MI). A comprehensive analysis of these samples is published (Kooperberg et al., 2007). In our terminology, this biomarker study is a two-phase study. The first-phase data consist of the randomized treatment assignment and stroke outcome for 16,608 study participants. The second phase consists of 124 cases and 504 controls from whom blood was analyzed. All blood biomarkers are continuous and were logarithm (base 10) transformed. To eliminate potential confounding factors, we included a number of important clinical characteristics in the second phase data, such as age, physical activity levels, diabetes, hypertension, systolic and diastolic blood pressure, and waist:hip ratio. We are interested in the interaction between the hormone treatment and each of biomarkers in a logistic regression adjusting for both main effects and aforementioned clinical characteristics. We compared four different methods to analyze two-phase data: the standard method ignoring the missing data (complete-case only) and the independence between the treatment and biomarkers, namely complete-case analysis, the proposed estimated likelihood estimator with or without exploiting independence, and the proposed SPMLE with or without exploiting covariate independence.

Table 4 shows the estimates of the interaction between the treatment and the thrombosis biomarkers using the aforementioned five methods. The model used in this table is

$$\text{logit}(P(\text{stroke}|\ldots)) = \beta_0 + \beta_1 HT + \beta_2 \log(B) + \beta_3 HT \times \log(B) + \sum_i \beta_{i+3} X_i,$$

where stroke is the event of a WHI participant having a stroke, HT an indicator whether this participant was assigned to Hormone Therapy, B the (continuous) biomarker, and the X_i

are other confounding factors. For ease of exposition, we only show the results of one class of biomarkers. Without exploiting the independence, there is essentially no improvement in efficiency over the complete-case analysis using either the MELE or the SPMLE. On the other hand, the standard errors of the two proposed methods exploiting independence is markedly smaller than the standard complete-case analysis and two other semiparametric methods. We found three thrombosis markers show a markedly increased significance. In particular, the p-values for PAP (plasmin-antiplasmin complex) are significant after adjusting for 29 tests by Bonferroni correction. A similar pattern of variance reduction, though in a smaller magnitude, is observed for the main effects - the treatment effect and the biomarker effect. The results of the MELE and the SPMLE exploiting independence are mostly indistinguishable. This is probably because the stroke is rather rare event (124 cases out of 16608 participants in this study) in the study cohort, the one-step estimated covariate distribution in the MELE is close to the iteratively estimated one in the SPMLE. These results demonstrate the advantage of our methods: exploiting the independence between the randomized treatment and the biomarker improves the power of detecting the interaction between them.

[Table 4 about here.]

7. Discussion

With the rapid advance of biotechnologies such as gene chips and proteomics, it is increasingly popular to employ some form of two-phase sampling to measure the expensive biomarker data for a sample of the study cohort, while collecting the cheap covariates and outcomes for everyone. When the phase-one cohort is a randomized clinical trial, there is independence between the treatment and baseline covariates. In this article, we show that exploiting this independence substantially improves the estimation efficiency, particularly for treatment-biomarker interactions. This is especially relevant to many pharmacogenetic studies where

drug-genotype interactions are under investigation. In an observational two-phase study, however, unless we have strong *a priori* evidence about the gene-environment independence, we should exert extra caution when exploiting it in estimation as deviation from the independence can draw a substantial bias (Albert et al., 2001).

Beyond exploiting the independence caused by randomization, our methods have general merits in semiparametric estimation. Existing methods to compute the SPMLE, such as the EM algorithm and the profile likelihood, all have limitations. We propose a profile likelihood based Newton–Raphson algorithm that computes the SPMLE for a wide range of data forms including continuous outcomes, as long as the phase-one covariate is discrete and the phase-two data is missing at random. An innovation in our algorithm is the usage of numerical differentiation to compute the profile information matrix, thus avoiding the complicated algebraic derivation of Lawless, Kalbfleisch and Wild (1999). In situations where computing the SPMLE is time-consuming, the proposed estimated likelihood approach may help. A contribution of this article is to derive the asymptotics for the estimated likelihood in two-phase sampling and work out the efficiency gain when using the covariate independence. In our simulations and data application, the MELE performs almost as good as the SPMLE. Simulations in Lawless, Kalbfleisch and Wild (1999) suggest that the relative efficiency of MELE may decline when the effect size increases. In genetic association studies with many markers, the majority of markers have no effect and some may have weak effect, the estimated likelihood will be time-efficient in screening for a subset of interesting markers. The SPMLE can be used subsequently to get a more precise estimate for the biggest hits.

Semiparametric efficient estimators can be alternatively derived in the framework of augmented inverse probability weighted estimators (Robins, Rotnitzky and Zhao, 1994). When parametric models involved are correctly specified, the estimates are asymptotically equivalent to the SPMLE derived under the likelihood framework, but are harder to implement

(Carpenter, Kenward and Vansteelandt, 2006). The augmented inverse probability weighted approach has its appeal in that it remains consistent if either the selection probability or the conditional distribution of missing data given observed data is correctly modeled. In randomized clinical trials, when baseline covariates (Z) to be adjusted are continuous and high-dimensional, it is almost impossible to correctly specify the parametric distribution of $Y|X,Z$. Then the SPMLE assuming the wrong model of $Y|X,Z$ will not be consistent. However, since the sampling probabilities are precisely controlled by the investigators, it is possible to construct a doubly-robust estimator that exploits the independence introduced by randomization, yet still yields consistent estimators of treatment-biomarker interactions. Indeed, Robins and Ritov (1997) shows that in this setting any estimator that fails to use the knowledge of sampling probabilities can perform poorly in moderate samples. It remains interesting for future methodological studies to exploit the independence in the framework of doubly-robust estimators.

Supplementary Materials

Web Appendices referenced in Section 3 and 4 are available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.

Acknowledgments

This work was supported in part by NIH grant R01 CA 74841, P01 CA53996 and U01 CA125489. The Women's Health Initiative program is funded by the National Heart, Lung, and Blood Institute, US Department of Health and Human Services. The authors are grateful to the associate editor and the referee for their helpful comments which led to improved clarity of our presentation.

References

Albert, P. S., Ratnasinghe, D., Tangrea, J., and Wacholder, S. (2001). Limitations of the case-only design for identifying gene-environment interactions. *American Journal of Epidemiology* **154,** 587–693.

Arnett, D. K., Davis, B. R., Ford, C. E., Boerwinkle, E., Leiendecker-Foster, C., B., M. M., Black, H., and Eckfeldt, J. H. (2005). Pharmacogenetic association of the angiotensin-converting enzyme insertion/deletion polymorphism on blood pressure and cardiovascular risk in relation to antihypertensive treatment: the genetics of hypertension-associated treatment (GenHAT) study. *Circulation* **111,** 3374–3383.

Breslow, N. E. and Cain, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika* **75,** 11–20.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research I. The Analysis of Case-control Studies*. International Agency for Research on Cancer: Lyon, France.

Breslow, N. E. and Robins, J. M. and Wellner, J. M (2003). Large sample theory for semiparametric regression models with two-phase, outcome-dependent sampling. *Annals of Statistics* **31,** 1110–1139.

Carpenter, J. R. and Kenward, M. G. and Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society, Ser. A* **169,** 571–584.

Carroll, R. J. and Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society, Ser. B* **53,** 573–585.

Chatterjee, N. and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* **92,** 399–418.

Chatterjee, N. and Chen, Y. H. (2005). Maximum likelihood inference on a mixed conditionally

and marginally specified regression model for genetic epidemiologic studies with two-phase sampling. *Journal of the Royal Statistical Society, Ser. B* **69,** 123–142.

Chatterjee, N., Chen, Y. H., and Breslow, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association* **98,** 158–168.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. New York: Chapman and Hall.

Flanders, W. D. and Greenland, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine* **10,** 739–747.

Foutz, R. V. (1977). On the unique consistent solution to the likelihood equations. *Journal of the American Statistical Association* **72,** 147–148.

Hu, X. J. and Lawless, J. F. (1996). Estimation from truncated lifetime data with supplementary information on covariates and censoring times. *Biometrika* **83,** 747–761.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association* **85,** 765–769.

Ibrahim, J. G., Chen, M., and Lipsitz, S. R. (1999). Monte carlo em for missing covariates in parametric regression models. *Biometrics* **55,** 591–596.

Kooperberg, C., Cushman, M., Hsia, J., Robinson, J. G., Aragaki, A. K., Lynch, J. K., Baird, A. E., Johnson, K. C., Kuller, L. H., Beresford, S. A., and Rodriguez, B. Can biomarkers identify women at increased stroke risk? *PLoS clinical trails* **15:**e28.

Lawless, J. F. (1997). *Likelihood and pseudo likelihood estimation based on response-biased observation*. Proc. Georgia Symp. Estimation Functions, Hayward: Institute of Mathematical Statistics.

Lawless, J. F., Kalbfleisch, J. D., and Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society, Ser. B* **61,** 413–438.

Lipsitz, S. R., Ibrahim, J. G., and Zhao, L. P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association* **94,** 1147–1160.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood (with discussion). *Journal of the American Statistical Association* **95,** 449–485.

Pepe, M. S. and Fleming, T. R. (1991). A non-parametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association* **86,** 108–113.

Piegorsch, W. W., Weinberg, C. R., and Taylor, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population based case-control studies. *Statistics in Medicine* **13,** 153–162.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-controls studies. *Biometrika* **66,** 403–411.

Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* **82,** 299–314.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89,** 846–866.

Robins, J. M., and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semiparametric models. *Statistics in Medicine* **16,** 285–319.

Rossouw, J. E., Anderson, G. L., Prentice, R. L., LaCroix, A. Z., Kooperberg, C., Stefanick, M. L., Jackson, R. D., Beresford, S. A., Howard, B. V., Johnson, K. C., Kotchen, J. M., Ockene, J. (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women's health initiative randomized controlled trial. *Journal of American Medical Association* **288,** 321–333.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63,** 581–592.

Scott, A. J. and Wild, C. J. (1991). Fitting logistic regression models in stratified case-control studies. *Biometrics* **47,** 497–510.

Scott, A. J. and Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84,** 57–71.

Umbach, D. M. and Weinberg, C. R. (1997). Designing and analyzing case-control studies to exploit independence of genotype and exposure. *Statistics in Medicine* **16,** 1731–1743.

Weaver, M. A. and Zhou, H. (2005). An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *Journal of the American Statistical Association* **100,** 459–469.

White, J. E. (1982). A two-stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology* **115,** 119–128.

Table 1

Binary Y: a comparison of weighted estimation equation estimator (WE), semiparametric maximum likelihood estimator (SPMLE), the maximal estimated likelihood estimator (MELE), with ($^{\perp}$) and without exploiting independence assumption in 5000 simulations.

	$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_3$		
Method	Bias	SD	RE[a]	Bias	SD	RE[a]	Bias	SD	RE[a]
$\beta_3 = 0$									
CC	−0.0009	0.135	65	0.0022	0.039	100	0.0001	0.054	100
WE	0.0009	0.110	97	0.0027	0.040	96	0.0001	0.055	96
MELE	0.0009	0.110	98	0.0025	0.040	97	0.0001	0.055	97
SPMLE	0.0008	0.109	100	0.0021	0.039	100	0.0001	0.054	100
MELE$^{\perp}$	0.0006	0.091	144	0.0010	0.034	135	0.0003	0.037	215
SPMLE$^{\perp}$	0.0006	0.091	144	0.0009	0.033	137	0.0003	0.037	214
$\beta_3 = 0.5$									
CC	−0.0025	0.161	67	−0.0004	0.045	100	0.0048	0.071	100
WE	−0.0021	0.141	87	0.0001	0.046	97	0.0074	0.076	87
MELE	−0.0047	0.133	98	− 0.0001	0.045	98	0.0047	0.072	98
SPMLE	−0.0050	0.132	100	−0.0004	0.045	100	0.0048	0.071	100
MELE$^{\perp}$	−0.0041	0.121	118	− 0.0002	0.039	128	0.0046	0.058	148
SPMLE$^{\perp}$	−0.0063	0.120	120	−0.0027	0.038	137	0.0055	0.057	154
$\beta_3 = 1$									
CC	−0.0067	0.189	68	−0.0010	0.049	100	0.0093	0.101	100
WE	−0.0122	0.164	89	−0.0008	0.050	99	0.0117	0.107	88
MELE	−0.0079	0.156	99	− 0.0010	0.050	99	0.0090	0.102	98
SPMLE	−0.0083	0.155	100	−0.0010	0.049	100	0.0093	0.101	100
MELE$^{\perp}$	−0.0052	0.149	108	−0.0023	0.046	118	0.0067	0.091	122
SPMLE$^{\perp}$	−0.0083	0.147	110	−0.0027	0.045	124	0.0086	0.089	127

Note: [a] - Relative efficiency comparing to the SPMLE ignoring the independence. The phase-one cohort size is 10,000, 800 cases and 800 controls are selected in the phase-two. The data is generated by $\text{logit}[P(Y = 1|X, Z)] = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ$, where $\beta_0 = -3$, β_1=0.2, β_2=0.1. $X \sim Ber(0.5)$, $Z \sim min(10, e^{N(0,1)})$.

Table 2
Binary Y: the performance of variance estimators based on the profile information: 5000 simulations

	SPMLE			**SPMLE** $^{\perp}$		
Parameter	$Var(\hat{\beta})$	$\widehat{Var}(\hat{\beta})$	95% C.I.	$Var(\hat{\beta})$	$\widehat{Var}(\hat{\beta})$	95% C.I.
$\beta_3 = 0$						
$\hat{\beta}_1$	0.0118	0.0119	95.1%	0.0082	0.0082	94.9%
$\hat{\beta}_2$	0.0015	0.0015	94.7%	0.0011	0.0011	95.4%
$\hat{\beta}_3$	0.0029	0.0029	94.6%	0.0014	0.0013	94.9%
$\beta_3 = 0.5$						
$\hat{\beta}_1$	0.0173	0.0171	94.6%	0.0144	0.0142	95.2%
$\hat{\beta}_2$	0.0020	0.0020	95.0%	0.0015	0.0014	95.1%
$\hat{\beta}_3$	0.0051	0.0050	94.8%	0.0033	0.0033	94.9%
$\beta_3 = 1$						
$\hat{\beta}_1$	0.0241	0.0244	95.3%	0.0217	0.0224	95.4%
$\hat{\beta}_2$	0.0025	0.0024	95.4%	0.0020	0.0020	95.6%
$\hat{\beta}_3$	0.0101	0.0101	95.5%	0.0079	0.0081	95.0%
	MELE			**MELE**$^{\perp}$		
Parameter	$Var(\tilde{\beta})$	$\widehat{Var}(\tilde{\beta})$	95% C.I.	$Var(\tilde{\beta})$	$\widehat{Var}(\tilde{\beta})$	95% C.I.
$\beta_3 = 0$						
$\tilde{\beta}_1$	0.0121	0.0121	95.2%	0.0082	0.0082	94.8%
$\tilde{\beta}_2$	0.0016	0.0015	94.7%	0.0011	0.0011	95.3%
$\tilde{\beta}_3$	0.0030	0.0029	94.7%	0.0014	0.0013	95.0%
$\beta_3 = 0.5$						
$\tilde{\beta}_1$	0.0176	0.0176	94.7%	0.0147	0.0146	95.1%
$\tilde{\beta}_2$	0.0020	0.0020	95.1%	0.0016	0.0015	94.8%
$\tilde{\beta}_3$	0.0052	0.0053	94.9%	0.0034	0.0034	95.4%
$\beta_3 = 1$						
$\tilde{\beta}_1$	0.0244	0.0248	95.5%	0.0222	0.0229	95.3%
$\tilde{\beta}_2$	0.0025	0.0024	95.5%	0.0021	0.0021	95.5%
$\tilde{\beta}_3$	0.0103	0.0108	95.4%	0.0083	0.0085	96.0%

Note: The phase-one cohort size is 10^4, 800 cases and 800 controls are selected in the phase-two. The data is generated by $\text{logit}[P(Y = 1|X, Z)] = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ$, where $\beta_0 = -3$, β_1=0.2, β_2=0.1. $X \sim Ber(0.5)$, $Z \sim min(10, e^{N(0,1)})$. $Var(\hat{\beta})$ is the sample variance of $\hat{\beta}$ in 5000 simulations. $\widehat{Var}(\hat{\beta})$ is the sample mean of the estimated variance of $\hat{\beta}$ in 5000 simulations.

Table 3

Continuous Y: a comparison of the weighted estimator (WE), the reduced semiparametric maximum likelihood estimator as in the (Lawless, Kalbfleisch and Wild, 1999) (LKW), the semiparametric maximum likelihood estimator (SPMLE), the maximal estimated likelihood estimator (MELE), with ($^{\perp}$) and without exploiting independence assumption in 1000 simulations.

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_3$		$\hat{\sigma}$	
	SD	RE[a]	SD	RE[a]	SD	RE[a]	SD	RE[a]
$\beta_3 = 0$								
WE	0.152	25	0.045	57	0.064	54	0.036	22
LKW	0.149	26	0.038	81	0.054	73	0.036	22
MELE	0.078	94	0.037	86	0.050	88	0.017	99
SPMLE	0.076	100	0.034	100	0.047	100	0.017	100
MELE$^{\perp}$	0.070	117	0.032	110	0.038	154	0.017	99
SPMLE$^{\perp}$	0.070	117	0.031	121	0.038	153	0.017	100
$\beta_3 = 0.5$								
WE	0.160	25	0.051	53	0.060	49	0.039	21
LKW	0.158	26	0.041	79	0.052	67	0.041	19
MELE	0.084	92	0.038	93	0.044	95	0.018	99
SPMLE	0.080	100	0.037	100	0.042	100	0.018	100
MELE$^{\perp}$	0.077	109	0.034	117	0.039	116	0.018	101
SPMLE$^{\perp}$	0.075	113	0.033	122	0.039	120	0.018	101
$\beta_3 = 1$								
WE	0.175	35	0.056	73	0.064	66	0.044	19
LKW	0.188	30	0.060	65	0.069	57	0.046	18
MELE	0.109	89	0.048	101	0.052	99	0.019	99
SPMLE	0.103	100	0.048	100	0.052	100	0.019	100
MELE$^{\perp}$	0.098	111	0.046	113	0.049	112	0.019	100
SPMLE$^{\perp}$	0.096	117	0.045	116	0.049	114	0.019	100

Note: [a] - Relative efficiency comparing to SPMLE ignoring independence. The phase-one cohort size is 2000, 200 cases and 200 controls are selected in the phase-two. The data is generated by a normal distribution with the mean $E[Y = 1|X, Z] = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ$ and variance $\sigma^2 = 1$, $\beta_0 = -1$, β_1=0.2, β_2=0.1. $X \sim Ber(0.5)$, $Z \sim \min(10, e^{N(0,1)})$.

Table 4

WHI E+P trial: an investigation of the interactions between the hormone treatment and biomarkers.

	complete-case		**MELE**		**SPMLE**		**MELE$^{\perp}$**		**SPMLE$^{\perp}$**	
Biomarker	HT*BM	p-val	HT*BM	p-val	HT*BM	p-val	HT*BM	p-val	HT*BM	p-val
Thrombosis										
D-dimer	−1.418(0.671)	0.034	−1.318(0.663)	0.047	−1.418(0.669)	0.034	−1.499(0.579)	0.009	−1.578(0.588)	0.007
Factor VIII	−1.367(1.025)	0.182	−1.442(1.068)	0.177	−1.366(1.020)	0.180	−1.846(0.932)	0.047	−1.791(0.901)	0.047
Fibrinogen	−0.479(1.761)	0.785	−0.625(1.772)	0.724	−0.479(1.743)	0.783	−0.092(1.537)	0.952	−0.092(1.557)	0.953
PAI-1	0.481(0.572)	0.401	0.472(0.565)	0.403	0.480(0.570)	0.399	0.308(0.501)	0.539	0.320(0.507)	0.528
PAP	−3.815(1.327)	0.004	−3.755(1.331)	0.005	−3.815(1.313)	0.004	−3.796(1.156)	0.001	−3.871(1.135)	6e-04
TAFI	−0.630(1.253)	0.615	−0.751(1.254)	0.549	−0.630(1.243)	0.612	−0.473(1.116)	0.672	−0.491(1.132)	0.665
Vwf	−1.289(1.097)	0.239	−1.251(1.096)	0.254	−1.290(1.088)	0.236	−1.151(0.980)	0.240	−1.160(0.984)	0.238

Note: HT*BM: the interaction between the hormone treatment and biomarker; D-dimer: fibrin D-dimer; PAI-1: plasminogen activator inhibitor-1 antigen; PAP: plasmin-antiplasmin complex; TAFI: tissue factor pathway inhibitor ; Vwf: von Willebrand factor.