Statistica Sinica Preprint No: SS-2021-0214	
Title	Semiparametric Maximum Likelihood Estimation With Two-Phase Stratified Case-Control Sampling
Manuscript ID	SS-2021-0214
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0214
Complete List of Authors	Yaqi Cao, Lu Chen, Ying Yang and Jinbo Chen
Corresponding Author	Jinbo Chen
E-mail	jinboche@pennmedicine.upenn.edu
Notice: Accepted version subject to English editing.	

Semiparametric Maximum Likelihood Estimation with Two-phase Stratified Case-control Sampling

Yaqi Cao*, Lu Chen*, Ying Yang, Jinbo Chen**

University of Pennsylvania, Tsinghua University

Abstract: We study statistical inference methods for fitting logistic regression models to data arising from the two-phase stratified case-control sampling design, where a subset of covariates are available only for a portion of cases and controls who are selected based on the case-control status and fully collected covariates. We are additionally interested in characterizing the distribution of incomplete covariates conditional on fully observed ones. It is desirable to include all subjects in the analysis to achieve consistency of parameter estimation and optimal statistical efficiency. We develop a semiparametric maximum likelihood approach under the rare disease assumption, where parameter estimates are obtained through a novel reparametrized profile likelihood technique. We study the large sample distribution theory for the proposed estimator, and demonstrate through simulation studies that it performs well in finite samples and has improved statistical efficiency compared with existing approaches. We apply the proposed method to analyze a stratified case-control study of breast cancer nested

---

* Co-first authors

** Corresponding author

---

within the Breast Cancer Detection and Demonstration Project, where one breast cancer risk predictor, percent mammographic density, was measured only for a subset of study women.

Key words and phrases: Logistic regression model; Profile likelihood; Semiparametric maximum likelihood; Stratified case-control study; Two-phase sampling.

1. Introduction

In stratified case-control study design, cases and controls are selected via stratified matching from the population subgroups with or without the outcome of interest. The logistic regression model is most frequently adopted for relating the binary outcome status with interest covariates, where the stratified sampling is accounted for through adjustment of matching strata in the model. Specialized methods have been studied when covariates are fully observed and the effect of covariates that are involved in forming matching strata is of interest (Fears and Brown, 1986; Breslow and Cain, 1988; Scott and Wild, 1991). Here, we consider the two-phase stratified case-control design, where some covariates are collected in the first phase for all subjects in a stratified case-control sample, and the remaining covariates are collected in the second phase only for a subset of cases and controls. The resulting incomplete data conforms to a missing at random

---

mechanism (Little and Rubin, 1987). We develop statistical methods for achieving two analytical goals, to obtain consistent and efficient estimates of odds ratio (OR) association parameters for covariates collected in both phases and to estimate the distribution of incomplete covariates conditional on the observed ones. The second goal is important when it is of interest to quantify the added value of incompletely measured covariates, e.g., expensive biomarkers, for prediction.

The two-phase sampling design (Neyman, 1938; White, 1982) has been widely used as a cost-effective option for data collection. For studying binary outcomes, statistical methods for analyzing two-phase data when Phase I consists of a cross-sectional or unstratified case-control sample have been widely studied (Breslow and Chatterjee, 1999; Breslow and Holubkov, 1997b; Lawless et al., 1999; Chatterjee and Chen, 2007). These methods have been focused on improving statistical efficiency for consistent estimation of odds ratio association parameters without imposing structural constraints on the covariate distribution (Breslow and Cain, 1988; Schill et al., 1993; Scott and Wild, 1997; Breslow and Holubkov, 1997a). To date, the two-phase stratified case-control study design has received limited attention despite its frequent use. While data from the stratified case-control design is analyzed essentially in the same way as the standard case-control

---

data, available statistical methods for analyzing two-phase case-control data cannot accommodate stratified matching in Phase I. When the selection of Phase II subjects is performed within Phase I matching strata, the data can be analyzed using two-phase case-control methods as long as the matching strata are fully adjusted for in the model and their effects are not of interest. However, such analysis can be highly inefficient (Chen et al., 2008) and yield biased estimates when Phase II sampling does not fully comply with Phase I stratified matching.

In this work, we develop a novel semiparametric maximum likelihood (ML) approach when a parametric regression model is imposed for the conditional distribution of incomplete covariates. Our method yields consistent and efficient estimates of all parameters simultaneously. In the same spirit as the prospective analysis of the standard case-control data (Prentice and Pyke, 1979), our method does not impose structural constraints on the distribution of fully observed covariates. This nuisance distribution function is eliminated from the empirical retrospective likelihood function via a novel reparametrized profile likelihood technique. The computation of our estimates is therefore highly efficient. The rest of this article is organized as follows. In Section 2, we describe our method and present the large sample theory. We also describe an existing pseudo-likelihood (PL)

---

approach (Chen et al., 2008) that addressed the challenge that selection of Phase II subjects was stratified on variables other than those used in Phase I matching. PL requires enumeration of cases and controls in each Phase I matching stratum in the underlying cohort from which the Phase I sample was drawn. It does not provide estimates of the conditional distribution of the incomplete covariates. In Section 3, we illustrate our method using data from the Breast Cancer Detection and Demonstration Project (BCDDP). Results from extensive simulation studies for evaluating the finite sample performance of our method in comparison with PL are presented in Section 4. We make final remarks in Section 5.

2. Methods

2.1 Notation and model assumptions

Let Y denote the case-control status ($Y = 1$: case; $Y = 0$: control) and $(\mathbf{X}, Z)$ denote the covariates of interest, where $\mathbf{X}$ is a $p \times 1$ vector, $\mathbf{X} = (X_1, \dots, X_p)^T$, and Z is univariate. Let A denote the matching strata taking values a , $a = 1, \dots, S$, where A can be defined based on $\mathbf{X}$. In Phase I, n_{1a} cases and n_{0a} controls are sampled from the conditional distributions $\Pr(\mathbf{X} \mid Y = 1, A = a)$ and $\Pr(\mathbf{X} \mid Y = 0, A = a)$, respectively, $a = 1 \dots, S$, for whom $\mathbf{X}$ is observed. The Phase I sample then consists of

2.1 Notation and model assumptions

a total of $n \equiv \sum_{y=0}^1 \sum_{a=1}^S n_{ya}$ subjects with $n_1 \equiv \sum_{a=1}^S n_{1a}$ cases and $n_0 \equiv \sum_{a=1}^S n_{0a}$ controls. In Phase II, a subset is selected from n subjects to collect data on Z via Bernoulli sampling with success probability $\omega(Y, \mathbf{X}, A)$ depending on the observed data $(Y, \mathbf{X}, A)$. Let R denote the selection decision ($R = 1$: yes; $R = 0$: no). The observed data then consists of $\mathcal{O} = \bigcup_{y=0,1} \bigcup_{a=1,\dots,S} \mathbf{O}_{ya} \equiv \{(R_i, \mathbf{X}_i^T, Z_i R_i)^T, i = 1, \dots, n_{ya}\}$, which we refer to as two-phase stratified case-control data. Because the selection does not depend on Z , the incompleteness of Z occurs at random, $\omega(Y, \mathbf{X}, A) \equiv \Pr(R | Y, \mathbf{X}, A) = \Pr(R | Y, \mathbf{X}, Z, A)$.

We are interested in assessing the relationship between the outcome variable Y and covariates $(\mathbf{X}^T, Z, A)^T$ using a logistic regression model:

$$\begin{aligned} \Pr(Y = 1 | \mathbf{X}, Z; A = a) &\equiv p_{\text{risk}}(Y = 1, A = a, \mathbf{X}, Z; \alpha_a, \boldsymbol{\beta}) \\ &= \frac{\exp(\alpha_a + \boldsymbol{\beta}_1^T \mathbf{X} + \beta_2 Z)}{1 + \exp(\alpha_a + \boldsymbol{\beta}_1^T \mathbf{X} + \beta_2 Z)}, \quad a = 1, \dots, S, \end{aligned} \quad (2.1)$$

where $\boldsymbol{\beta}$ denotes $(\boldsymbol{\beta}_1^T, \beta_2)^T$. Below we use $\boldsymbol{\alpha}$ to denote the collection of stratum-specific intercept parameters $(\alpha_1, \dots, \alpha_S)^T$. It is also of interest to describe the distribution of Z conditional on $(\mathbf{X}, A)$, $\Pr(Z | \mathbf{X}, A)$. Because $(\mathbf{X}, A)$ may have a large number of values, nonparametric modeling of $\Pr(Z | \mathbf{X}, A)$ is practically challenging. Therefore, we adopt a parametric model for $\Pr(Z | \mathbf{X}, A)$, denoted as $p_{\boldsymbol{\theta}}(Z | \mathbf{X}, A)$, with $\boldsymbol{\theta}$ being the index parameter vector. We develop statistical methods for jointly estimating $\boldsymbol{\beta}$

2.2 Estimation of OR parameters and conditional distribution of Z

and $\boldsymbol{\theta}$ using the observed data $\mathcal{O}$. In subsequent development, we do not impose model structure on the distribution $\Pr(\mathbf{X} \mid A = a)$, because it is not of interest and also infeasible to impose reasonable multivariate parametric models for this whole vector of covariates. We assume that the outcome is rare (Spinka et al., 2005; Chen et al., 2007). Then $\Pr(\mathbf{X} \mid A; Y = 0)$ and $\Pr(Z \mid \mathbf{X}, A; Y = 0)$ can be approximated by $\Pr(\mathbf{X} \mid A)$ and $p_{\boldsymbol{\theta}}(Z \mid \mathbf{X}, A)$, respectively.

2.2 Estimation of OR parameters and conditional distribution of Z

Let $\delta_{\mathbf{x}}^a$ denote the probability mass of $\mathbf{X} = \mathbf{x}$ in the a^{th} matching stratum ($a = 1, \dots, S$), which satisfies $\sum_{\mathbf{x}} \delta_{\mathbf{x}}^a = 1$. Let $\boldsymbol{\delta}_a$ denote the collection of $\delta_{\mathbf{x}}^a$, $a = 1, \dots, S$. The totality of nuisance parameters $\{\boldsymbol{\delta}_a, a = 1, \dots, S\}$, denoted as $\boldsymbol{\delta}$, increases with sample size n . The retrospective log-likelihood function for the observed data $\mathcal{O}$ can be written as

$$\begin{aligned} \ell_0(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\delta}) = & \log \Pr\{R_i, i = 1, \dots, n \mid Y_i, \mathbf{X}_i, A_i, i = 1, \dots, n\} \\ & + \log \left\{ \prod_{i \in P_I/P_{II}} \Pr(\mathbf{X}_i \mid A_i, Y_i) \prod_{i \in P_{II}} \Pr(\mathbf{X}_i, Z_i \mid A_i, Y_i) \right\}, \end{aligned}$$

where P_I and P_{II} denote subjects in Phase I and Phase II, respectively.

Because the missingness is at random, we can obtain the semiparametric ML estimator of $(\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$ by maximizing the second part of $\ell_0(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\delta})$.

2.2 Estimation of OR parameters and conditional distribution of Z

Using a result in Satten and Kupper (1993), $\Pr(\mathbf{X}, Z \mid A; Y = 1)$ is related to $\Pr(\mathbf{X}, Z \mid A; Y = 0)$ as follows,

$$\Pr(\mathbf{X}, Z \mid A; Y = 1) = \frac{\exp(\boldsymbol{\beta}_1^T \mathbf{X} + \beta_2 Z) \Pr(\mathbf{X}, Z \mid A; Y = 0)}{\sum_{\mathbf{x}, z} \exp(\boldsymbol{\beta}_1^T \mathbf{x} + \beta_2 z) \Pr(\mathbf{x}, z \mid A; Y = 0)}.$$

The second part of $\ell_0(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\delta})$ for a rare outcome can then be written as

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\delta}) = & \sum_{i=1}^n \left[R_i \log \{ \exp(Y_i(\boldsymbol{\beta}_1^T \mathbf{X}_i + \beta_2 Z_i)) p_{\boldsymbol{\theta}}(Z_i \mid \mathbf{X}_i, A_i) \} \right. \\ & + (1 - R_i) Y_i \log \left\{ \sum_z \exp(\boldsymbol{\beta}_1^T \mathbf{X}_i + \beta_2 z) p_{\boldsymbol{\theta}}(Z = z \mid \mathbf{X}_i, A_i) \right\} + \log \delta_{\mathbf{X}_i}^{A_i} \left. \right] \\ & - \sum_a n_{1a} \log \left\{ \sum_{\mathbf{x}, z} \exp(\boldsymbol{\beta}_1^T \mathbf{x} + \beta_2 z) p_{\boldsymbol{\theta}}(Z = z \mid \mathbf{x}, a) \cdot \delta_{\mathbf{x}}^a \right\}. \end{aligned} \quad (2.2)$$

We seek to obtain estimates for the interest parameters $(\boldsymbol{\beta}, \boldsymbol{\theta})$, $(\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\theta}}^T)$, as the maximizer for $\ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\delta})$. If $\mathbf{X}$ only had a small number of unique values, then $(\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\theta}}^T, \hat{\boldsymbol{\delta}}^T)^T$ could be obtained simultaneously as estimates for $(\boldsymbol{\beta}^T, \boldsymbol{\theta}^T, \boldsymbol{\delta}^T)^T$, by applying the standard Expectation-Maximization (EM) algorithm. But $\mathbf{X}$ may have many values in reality, and the number of unique values increases with sample size when some components are continuous. It is well known that the EM algorithm often breaks down when the data is sparse relative to the number of parameters (Little and Rubin, 1987). We therefore first eliminate the high-dimensional nuisance parameter $\boldsymbol{\delta}$ by deriving the profile log-likelihood $\ell_p(\boldsymbol{\beta}, \boldsymbol{\theta}, \hat{\boldsymbol{\delta}}(\boldsymbol{\beta}, \boldsymbol{\theta})) \equiv \sup_{\boldsymbol{\delta}} \ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\delta})$

2.2 Estimation of OR parameters and conditional distribution of Z

with respect to $(\boldsymbol{\beta}, \boldsymbol{\theta})$.

Below, through the profile likelihood approach, we establish a key result that estimates $(\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\theta}}^T)$ can be obtained by maximizing a proper prospective likelihood function with desirable large sample properties. We introduce new parameters $\mu_a = \mu_a(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\delta}_a) = \exp(-\alpha_a) \cdot \{\Pr(Y = 1 \mid A = a) / \Pr(Y = 0 \mid A = a)\}$, $a = 1, \dots, S$. Let $\boldsymbol{\mu}$ denote $(\mu_1, \dots, \mu_S)$. Recall that the two-phase stratified case-control data arises from distribution $\Pr(R, \mathbf{X}, Z \mid Y = y, A = a)$ as defined in Section 2.1. We define a modified two-phase prospective study design where data conforms to distribution $P^*(R, Y, \mathbf{X}, Z \mid A = a)$ as follows. Define a modified logistic regression model as

$$P^*(Y = 1 \mid \mathbf{x}, z; A = a) = \frac{\exp(\alpha_a^* + \boldsymbol{\beta}_1^T \mathbf{x} + \beta_2 z)}{1 + \exp(\alpha_a^* + \boldsymbol{\beta}_1^T \mathbf{x} + \beta_2 z)}, \quad (2.3)$$

where $\alpha_a^* = \log(n_{1a}/n_{0a}) - \log \mu_a = \alpha_a + \log(n_{1a}/n_{0a}) - \log(\Pr(Y = 1 \mid A = a) / \Pr(Y = 0 \mid A = a))$. At Phase I, conventional predictors $\mathbf{X}$ is collected from $P^*(\mathbf{X} = \mathbf{x} \mid A = a)$, which is unspecified. At Phase II, Z is collected from a modified conditional distribution $P^*(Z = z \mid \mathbf{x}, a)$,

$$P^*(Z = z \mid \mathbf{x}, a) \equiv \frac{(1 + \exp(\alpha_a^* + \boldsymbol{\beta}_1^T \mathbf{x} + \beta_2 z)) p_{\boldsymbol{\theta}}(Z = z \mid \mathbf{x}; A = a)}{\sum_{z'} \{(1 + \exp(\alpha_a^* + \boldsymbol{\beta}_1^T \mathbf{x} + \beta_2 z')) p_{\boldsymbol{\theta}}(Z = z' \mid \mathbf{x}; A = a)\}}. \quad (2.4)$$

The modified and actual conditional distributions of Z turn out to be iden-

2.2 Estimation of OR parameters and conditional distribution of Z

tical,

$$P^*(Z = z \mid \mathbf{x}, a; Y = y) = \Pr(Z = z \mid \mathbf{x}, a; Y = y),$$

as shown in the Supplementary Materials. The modified and true missingness probabilities are the same as well, $P^*(R = 1 \mid Y, \mathbf{X}, Z, A) = \Pr(R = 1 \mid Y, \mathbf{X}, Z, A) = \omega(Y, \mathbf{X}, A)$. The corresponding two-phase prospective log-likelihood can be written as follows,

$$\begin{aligned} \ell_{\text{ML}}^* &= \sum_{i=1}^n \log P^*(R_i \mid Y_i, \mathbf{X}_i, A_i) \\ &\quad + \sum_{i=1}^n R_i \log P^*(Y_i, Z_i \mid \mathbf{X}_i, A_i) + (1 - R_i) \log P^*(Y_i \mid \mathbf{X}_i, A_i). \end{aligned}$$

We show in the lemmas and theorems below that the ML estimator $(\hat{\beta}^T, \hat{\theta}^T)$ can be obtained by maximizing the second part of ℓ_{ML}^* ,

$$\ell^*(\beta, \theta, \mu) \equiv \sum_{i=1}^n R_i \log P^*(Y_i, Z_i \mid \mathbf{X}_i, A_i) + (1 - R_i) \log P^*(Y_i \mid \mathbf{X}_i, A_i).$$

In this sense, the two-phase stratified case-control design is equivalent to the modified prospective two-phase case-control design. We formally establish this result in Lemma 1 below. Let $\eta_{\mathbf{x}z}^a = \exp(\beta_1^T \mathbf{x} + \beta_2 z) p_{\theta}(Z = z \mid \mathbf{X} = \mathbf{x}, A = a)$. From (2.3) and (2.4), we obtain

$$P^*(Y = 1, Z = z \mid \mathbf{x}, a) = \frac{n_{1a} \eta_{\mathbf{x}z}^a / \mu_a}{n_{0a} + n_{1a} (\sum_{z'} \eta_{\mathbf{x}z'}^a / \mu_a)},$$

$$P^*(Y = 0, Z = z \mid \mathbf{x}, a) = \frac{n_{0a} p_{\theta}(Z = z \mid \mathbf{x}, a)}{n_{0a} + n_{1a} (\sum_{z'} \eta_{\mathbf{x}z'}^a / \mu_a)}.$$

2.2 Estimation of OR parameters and conditional distribution of Z

Note that $\sum_z f(z)$, the summation of $f(z)$ with all possible values of Z , is a general notation, where $f(z)$ can be any measurable function of z . When Z is continuous, $\sum_z f(z) = \int f(z)dz$.

Lemma 1. *Let $n_{\mathbf{x}z}$ denote the total number of cases and controls with $\mathbf{X} = \mathbf{x}$ and $Z = z$. The profile log-likelihood $\ell_p(\boldsymbol{\beta}, \boldsymbol{\theta}, \hat{\boldsymbol{\delta}}(\boldsymbol{\beta}, \boldsymbol{\theta}))$ is equivalent to $\ell^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \hat{\boldsymbol{\mu}}(\boldsymbol{\beta}, \boldsymbol{\theta}))$ up to a constant term, where $\hat{\boldsymbol{\mu}}(\boldsymbol{\beta}, \boldsymbol{\theta})$ is defined by the solution to equations*

$$n_{1a} = \sum_{\mathbf{x}, z} n_{\mathbf{x}z} P^*(Y = 1, Z = z \mid \mathbf{x}, a), \quad a = 1, \dots, S.$$

Although $\ell^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\mu})$ is not a genuine log-likelihood, we show below that the maximum likelihood estimator of $(\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\theta}}^T)$ can be obtained as the root to the “pseudo” score function

$$S^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\mu}) = \left(\frac{\partial \ell^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\mu})}{\partial \boldsymbol{\beta}^T}, \frac{\partial \ell^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\mu})}{\partial \boldsymbol{\theta}^T}, \frac{\partial \ell^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T} \right)^T$$

corresponding to the prospective likelihood $\ell^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\mu})$, and that $\ell^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \hat{\boldsymbol{\mu}}(\boldsymbol{\beta}, \boldsymbol{\theta})) \equiv \sup_{\boldsymbol{\mu}} \ell^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\mu})$. Our theory generalizes a result in Scott and Wild (2001) and Chatterjee and Chen (2007) where Phase I is a cross-sectional sample and Phase II is a stratified subsample. As a result, for the a^{th} stratum, a in $1, \dots, S$, we reduce the number of nuisance parameters, that is, the

2.2 Estimation of OR parameters and conditional distribution of Z

number of unique values of $\mathbf{X}$, to 1. The form of $S^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\mu})$ is given in Lemma 2 below.

Lemma 2. *Let $\mathbf{W}$ denote the vector of risk factors $(\mathbf{X}^T, Z)^T$, and Φ denote the vector of parameters $(\boldsymbol{\beta}^T, \boldsymbol{\theta}^T, \boldsymbol{\mu}^T)^T$. Write*

$$h(Y, \mathbf{W}, A) = Y \log(n_{1A} \cdot \eta_{\mathbf{X}Z}^A / \mu_A) + (1 - Y) \log(n_{0A} p_{\theta}(Z | \mathbf{X}, A))$$

Then the “pseudo” score function $S^*(\Phi)$ can be written as

$$\begin{aligned} S^*(\Phi) &= \partial \ell^*(\Phi) / \partial \Phi \\ &= \sum_i \left\{ R_i \frac{\partial h(Y_i, \mathbf{W}_i, A_i)}{\partial \Phi} + (1 - R_i) E_Z^* \left(\frac{\partial h(\cdot)}{\partial \Phi} \mid Y_i, \mathbf{X}_i, A_i \right) - E^* \left(\frac{\partial h(\cdot)}{\partial \Phi} \mid \mathbf{X}_i, A_i \right) \right\} \\ &\equiv \sum_{i=1}^n \varphi(R_i, Y_i, \mathbf{W}_i, A_i), \end{aligned} \tag{2.5}$$

where $E_Z^*(\cdot \mid Y, \mathbf{X}, A)$ and $E^*(\cdot \mid \mathbf{X}, A)$ denote the expectation taken with respect to $P^*(Z \mid Y, \mathbf{X}, A)$ and $P^*(Y, Z \mid \mathbf{X}, A)$, respectively. Moreover, $\partial \ell^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\mu}) / \partial \boldsymbol{\mu}^T \equiv \mathbf{0}$ at $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}(\boldsymbol{\beta}, \boldsymbol{\theta})$, so that $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ can be obtained by treating $\boldsymbol{\mu}$ as an independent set of parameters when solving the “pseudo” score equations $S^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\mu}) = \mathbf{0}$.

Lemma 1 and Lemma 2 shows that the MLE of $\boldsymbol{\beta}$ can be obtained by maximizing the modified prospective likelihood of the form $\Pr(Y = 1 | \mathbf{x}, z; A = a, R_{\text{scc}} = 1) \equiv P^*(Y = 1 | \mathbf{x}, z; A = a)$ in model (2.3), where

2.3 Asymptotic theory

R_{scc} is the indicator variable for whether or not a subject was selected into the Phase I stratified case-control sample ($R_{\text{scc}} = 1$: yes; $R_{\text{scc}} = 0$: no). $\Pr(R_{\text{scc}} = 1|Y = y, A = a)$, the probability of a subject being selected into stratum $(Y = y, A = a)$ is fixed at its asymptotic value and reflected by the new parameter μ_a . Proof of the two Lemmas is provided in the Supplementary Materials.

2.3 Asymptotic theory

We develop the asymptotic theory when the total sample size $n = \sum_{y=0}^1 \sum_{a=1}^S n_{ya}$ goes to infinity and the sampling proportions in each matching stratum, namely n_{ya}/n , converges to a positive constant π_{ya} , $\sum_{y=0}^1 \sum_{a=1}^S \pi_{ya} = 1$. We will repeatedly use the theorem below that establishes the relationship between the expectation under the stratified case-control sampling design and that under the pseudo prospective distributions as characterized by (2.3) and (2.4).

Theorem 1. *Under the two-phase stratified case-control sampling design, for any measurable function $Q(R, Y, \mathbf{W}, A)$ that satisfies*

$$0 < \int E^*(Q(R, Y, \mathbf{W}, A) | \mathbf{x}, a) \cdot g_a(\mathbf{x}) dF_a(\mathbf{x}) < \infty, a = 1, \dots, S,$$

where $g_a(\mathbf{x}) = (n_{0a}/n) + (n_{1a}/n) \cdot (\sum_{z'} \eta_{\mathbf{x}z'}^a / \mu_a)$ and $F_a(\mathbf{x})$ is the distribution

2.3 Asymptotic theory

of $\mathbf{X}$ given $A = a$,

$$n^{-1} \sum_{i=1}^n Q(R_i, Y_i, \mathbf{W}_i, A_i) \xrightarrow{p} \sum_a \int E^*(Q(R, Y, \mathbf{W}, A) \mid \mathbf{x}, a) \cdot g_a(\mathbf{x}) dF_a(\mathbf{x}).$$

We obtain the following corollary from Theorem 1 pertaining to the two-phase sampling, which, together with Lemma 2, leads directly to the asymptotic unbiasedness of the “pseudo” score functions $S^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\mu})$. We note that the contribution to $ES^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\mu})$ by each individual subject is not equal to zero because l^* is not the true log-likelihood function, but the total contribution by cases cancels out that by controls.

Corollary 1. *For any measurable function $Q(Y, \mathbf{W}, A)$ that satisfies*

$$0 < \int E^*(Q(Y, \mathbf{W}, A) \mid \mathbf{x}, a) \cdot g_a(\mathbf{x}) dF_a(\mathbf{x}) < \infty, a = 1, \dots, S,$$

where $g_a(\mathbf{x})$ and $F_a(\mathbf{x})$ are defined in Theorem 1,

$$\begin{aligned} n^{-1} \sum_{i=1}^n R_i Q(Y_i, \mathbf{W}_i, A_i) + (1 - R_i) E_Z^*(Q(Y, \mathbf{W}, A) \mid Y_i, \mathbf{X}_i, A_i) \\ \xrightarrow{p} \sum_a \int E^*(Q(Y, \mathbf{W}, A) \mid \mathbf{X} = \mathbf{x}, A = a) \cdot g_a(\mathbf{x}) dF_a(\mathbf{x}). \end{aligned}$$

To derive the main asymptotic limiting results, we obtain the large sample limit of the second derivatives of $\ell^*(\boldsymbol{\Phi})$ in the following lemma.

2.3 Asymptotic theory

Lemma 3. Let $V_{YZ}^*(\cdot | \mathbf{X}, A)$ and $V_Z^*(\cdot | Y, \mathbf{X}, A)$ denote variance functions with respect to $P^*(Y, Z | \mathbf{X}, A)$ and $P^*(Z | Y, \mathbf{X}, A)$ respectively. With function $h(Y, W, A)$ as defined in Lemma 2, we show that

$$-\frac{1}{n} \frac{\partial^2 \ell^*(\Phi)}{\partial^2 \Phi} \xrightarrow{p} \sum_{a=1}^S \int V_{YZ}^* \left(\frac{\partial h}{\partial \Phi} | \mathbf{x}, a \right) g_a(\mathbf{x}) dF_a(\mathbf{x}) - M \equiv \zeta, \quad (2.6)$$

where $M = \sum_{y,a} \int (1 - \omega(y, \mathbf{x}, a)) V_Z^* \left(\frac{\partial h}{\partial \Phi} | Y = y, \mathbf{x}, a \right) P^*(Y = y | \mathbf{x}, s, a) g_a(\mathbf{x}) dF_a(\mathbf{x})$.

In the proof of Lemma 3, we show that the elements in M except for the submatrix corresponding to $(\beta^T, \theta^T)^T$ are equal to zero. If Z is fully observed, M becomes a zero matrix.

Built on the lemmas and theorems above, we establish the large sample properties of $(\hat{\beta}^T, \hat{\theta}^T)^T$ in Theorem 2.

Theorem 2. Under regularity conditions listed in the Supplementary Materials,

(i) The estimating equations $S^*(\Phi) = \sum_{i=1}^n \varphi(R_i, Y_i, \mathbf{X}_i, Z_i, A_i) = \mathbf{0}$ defined in (2.5) have a unique sequence of solutions $\{\hat{\Phi}_n\}_{n \geq 1}$. $\hat{\Phi}_n \xrightarrow{p} \Phi_0$, where $\Phi_0 \equiv (\beta_0^T, \theta_0^T)^T$ is the true value for $(\beta^T, \theta^T)^T$.

(ii) $\sqrt{n}(\hat{\Phi}_n - \Phi_0) \xrightarrow{d} N(\mathbf{0}, \Sigma)$, where $\Sigma = \zeta^{-1} - (\zeta^{-1})^T \Omega \zeta^{-1}$ with

$$\Omega = \sum_a \frac{n_{0a}}{n} \left(E((\partial \varphi / \partial \Phi) | Y = 0, A = a)^{\otimes 2} \right) + \frac{n_{1a}}{n} \left(E((\partial \varphi / \partial \Phi) | Y = 1, A = a)^{\otimes 2} \right).$$

2.4 Comparison with the PL method

By this theorem, the inverse of the observed “pseudo” information matrix $\zeta^{-1} |_{\Phi=\hat{\Phi}_n}$ is an asymptotically conservative estimator for $\text{Cov}(\hat{\Phi}_n)$. The bias is corrected by the term $(\zeta^{-1})^T \Omega \zeta^{-1}$. We show in the Supplementary Materials that $\Omega = \zeta(\sum_a \mathbf{K}_a^T \mathbf{K}_a) \zeta$ with

$$\mathbf{K}_a = \begin{pmatrix} \underbrace{0, \dots, 0}_{\text{length of } (\beta^T, \theta^T)} & , \underbrace{0, \dots, 0}_{a-1} & , \mu_a \left(\frac{n}{n_{0a}} + \frac{n}{n_{1a}} \right)^{1/2} & , \underbrace{0, \dots, 0}_{S-a} \end{pmatrix}.$$

Therefore, the elements in the correction term $(\zeta^{-1})^T \Omega \zeta^{-1}$ corresponding to $(\hat{\beta}^T, \hat{\theta}^T)$ are equal to zero. That is, the correction term does not influence the variance-covariance matrix of $(\hat{\beta}^T, \hat{\theta}^T)$, so that the variance-covariance matrix of $(\hat{\beta}^T, \hat{\theta}^T)$ can be obtained directly from the corresponding element of the inverse of the observed “pseudo” information matrix $\zeta^{-1} |_{\Phi=\hat{\Phi}_n}$. The variance of μ_a needs to subtract $\mu_a^2 (n/n_{0a} + n/n_{1a})$.

2.4 Comparison with the PL method

We will compare the efficiency of our maximum likelihood (“ML”) estimator to an existing pseudo-likelihood (“PL”) estimator (Chen et al., 2008) described below using our notations. The original PL required stratum-specific enumeration of cases and controls in the cohort for generating the stratified case-control sample to get an estimate of μ_a . Here we describe PL with μ_a fixed at the true value as this does not affect the estimation

2.4 Comparison with the PL method

of odds ratio parameters for $\mathbf{X}$ and Z . The PL estimator maximizes a pseudo two-phase prospective likelihood based on the prospective models (2.3) and the distribution function of Z in the stratified case-control sample, $P^*(Z | \mathbf{X}, A)$. But it did not specify a parametric model for $\Pr(Z | \mathbf{X}, A)$, and the nuisance distribution $P^*(Z | \mathbf{X}, A)$ was treated nonparametrically instead of using the form (2.4). To do this, Phase I data $\mathbf{X}$ was coarsened into discrete strata denoted as $K(\mathbf{X})$. The pseudo-likelihood function was written as

$$\begin{aligned} \ell_{PL} &= \sum_{i=1}^n \left[R_i \log \{P^*(Y_i | \mathbf{X}_i, Z_i, A_i)P^*(\mathbf{X}_i, Z_i, A_i | K(\mathbf{X}_i))\} \right. \\ &\quad \left. + (1 - R_i) \log \{P^*(Y_i | K(\mathbf{X}_i))\} \right] \\ &= \sum_{i=1}^n R_i \log \{P^*(Y_i | \mathbf{X}_i, Z_i, A_i)P^*(\mathbf{X}_i, Z_i, A_i | K(\mathbf{X}_i))\} \\ &\quad + (1 - R_i) \log \left\{ \int_{\mathbf{x}, z, a} P^*(Y_i | \mathbf{x}, z, a)P^*(\mathbf{X} = \mathbf{x}, Z = z, A = a | K(\mathbf{X}_i))d\mathbf{x}dzda \right\}, \end{aligned} \quad (2.7)$$

The estimation proceeds by maximizing this pseudo-likelihood function, where $P^*(\mathbf{X}_i, Z_i, A_i | K(\mathbf{X}_i))$ was profiled out as in the maximum likelihood method for the standard two-phase case-control data (Breslow and Cain, 1988; Scott and Wild, 1997). Note as revealed by $P^*(\mathbf{X}_i, Z_i, A_i | K(\mathbf{X}_i))$, the age-stratum A and any information in $\mathbf{X}$ that is not captured by $K(\mathbf{X})$ were also considered as incomplete data. Ignoring A from

---

Phase I avoids sparse cells resulting from stratifying on both $\mathbf{X}$ and A , allowing maximum stratification on $\mathbf{X}$ so as to increase efficiency for estimating odds ratio parameters for $\mathbf{X}$. The ML and PL methods differ in three important aspects. First, ML eliminated the need to estimate parameters corresponding to the matching strata in model (2.3) through rare disease approximation, while the PL method does estimate the intercept parameters α_a^* . Second, ML adopted a parametric model for Phase II variable Z . PL bypassed this need by enforcing coarsening of $\mathbf{X}$ into strata $K(\mathbf{X})$ to facilitate maximization of the pseudo-likelihood, which incurs loss of information on $\mathbf{X}$. Third, when our primary interest was the estimation of odds ratio parameters, compared with the complete data analysis using the Phase II data, ML method achieved efficiency gain by fully utilizing Phase I data. However, the extent of efficiency gain for PL estimates were tied to the construction of $K(\mathbf{X})$. The predictors in $\mathbf{X}$ that were not used to form $K(\mathbf{X})$ do not enjoy efficiency gain.

3. Real Data Analysis

To illustrate the proposed method, we developed a multivariate odds ratio function for predicting breast cancer risk using data from the Breast Cancer Detection and Demonstration Project (BCDDP). BCDDP started in 1973

---

and was originally designed to assess whether mammographic screening can reduce morbidity and mortality of breast cancer. The study recruited 243,221 white women from year 1973 to 1975 and followed each woman for at least 5 years. An age-stratified case-control sample was assembled within BCDDP in 1979 that included 2,808 cases and 3,119 controls. Data for age at first live birth (Ageflb; X_1), age at menarche (Agemen; X_2), number of previous breast biopsies (Nbiops; X_3) and number of first-degree relatives with breast cancer (Numrel; X_4), was analyzed in relation to the risk of breast cancer to develop an odds ratio model that was used in the NCI Breast Cancer Risk Assessment Tool (BCRAT; Gail et al. (1989)). Using the same study sample, it was shown later that incorporation of percent mammographic breast density (BD; Z) into the BCRAT as a new predictor could lead to improved discriminatory accuracy (Chen et al., 2006). BD was only available for a subset of 1,217 cases and 1,616 controls. The availability of BD was found to relate only to the case-control status and age. Therefore, the PL method was adopted for analysis, treating the data as if it arose from the two-phase age-stratified case-control sampling design. The stratified case-control data was treated as the Phase I sample, and the subset that had BD available was treated as the Phase II sample. BD was then considered as the Phase II covariate. The age strata (A^c) were coded as $1, 2, \dots, 9$

---

corresponding to age intervals $< 40, [40, 45), [45, 50), [50, 55), \dots, [70, 75)$ and ≥ 75 years. Here, we re-analyze the data using the proposed ML method, additionally characterizing the conditional distribution of BD. Readers are referred to Supplementary Tables S1 and S2 in Chen et al. (2008) for the descriptive statistics of the predictors. The variable weight, X_5 , which was available for all cases and controls, is included as the auxiliary information for BD because they are strongly correlated (Byrne et al., 1995). We followed the same convention as the BCRAT for discretizing $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5)^T$. We applied both PL and ML to analyze the data. In PL method, discrete strata $K(\mathbf{X})$ in (2.7) was constructed by the cross-classification of $1(X_1 > 0)$ and $1(X_2 > 1)$, resulting in 4 strata. Note that the pseudo-likelihood function (2.7) cannot be used for estimating $\boldsymbol{\theta}$.

We fitted the following logistic regression model to relate breast cancer status ($Y = 1$: case; $Y = 0$: control) to the predictors, where Z was categorized into 5 intervals, denoted as Z^c , following Chen et al. (2008):

$$\Pr(Y = 1 \mid \mathbf{X}, Z^c; A^c = a) = \frac{\exp(\alpha_a + \boldsymbol{\beta}_1^T \mathbf{X} + \beta_2 Z^c)}{1 + \exp(\alpha_a + \boldsymbol{\beta}_1^T \mathbf{X} + \beta_2 Z^c)}, \quad a = 1, \dots, 9. \quad (3.8)$$

Note that our method does not require categorization of Z . Z^c was used instead of continuous Z in (3.8) mainly to facilitate risk calculation. After some initial explorations (Supplementary Material), we adopted a mixture

---

of the Bernoulli distribution and beta regression model to parameterize the conditional distribution of continuous variable Z given $\mathbf{X}$ and continuous age A . Let f_{γ} denote the probability density function of the Beta distribution and F_{γ} the corresponding cumulative distribution function. Let κ denote the mean parameter and ϕ the precision parameter, which are allowed to vary with respect to $(\mathbf{X}, A)$ governed by logistic and exponential regression models with regression parameters γ_{mean} and $\gamma_{\text{precision}}$ respectively. Let γ denote $(\gamma_{\text{mean}}, \gamma_{\text{precision}})$. The conditional distribution function for Z is then specified as

$$p_{\theta}(Z = z | \mathbf{X}, A) = \begin{cases} \rho + (1 - \rho)F_{\gamma}(z_{\min} | \mathbf{X}, A) & \text{if } z = 0 \\ (1 - \rho)f_{\gamma}(Z = z | \mathbf{X}, A) & \text{if } z \in (z_{\min}, 1), \end{cases} \quad \theta = (\rho, \gamma^T)^T,$$

where $\rho \in (0, 1)$ is the parameter denoting the mixture proportion, and

$$f_{\gamma}(Z = z | \mathbf{X}, A) = \frac{\Gamma(\phi)}{\Gamma(\kappa\phi)\Gamma(\phi - \kappa\phi)} z^{\kappa\phi-1} (1 - z)^{\phi-\kappa\phi-1}, \quad (3.9)$$

$$\kappa = \text{logit}^{-1} \{(\gamma_{\text{mean}})^T(1, \mathbf{X}^T, A)^T\}, \quad (3.10)$$

$$\phi = \exp \{(\gamma_{\text{precision}})^T(1, \mathbf{X}^T, A)^T\}. \quad (3.11)$$

We first fitted a full model that included all the candidate predictors, age, Ageflb, Agemen, Nbiops, Numrel, Weight, BD and Age (data not shown)

in all models. It turned out that all predictors were significantly associated with breast cancer risk (3.8). In model $p_{\theta}(Z = z \mid \mathbf{X}, A)$, the mixture parameter ρ was significantly different from zero. Numrel was not significant in the model for κ , and Ageflb, Agemen, and Numrel were not significant in the model for ϕ . The results with these insignificant predictors eliminated from the corresponding models are presented in Table 1.

Table 1: Analysis of the BCDDP data using the ML and PL methods: The estimate (Est), asymptotic standard error (SE) and p values for the log odds ratio parameter estimates (β) and the parameters in the conditional distribution of Z , $(\gamma_{\text{mean}}^T, \gamma_{\text{precision}}^T, \rho)$.

		ML		PL	
		Est (SE)	p value	Est (SE)	p value
Log Odds Ratio	Ageflb (X_1)	0.145 (0.034)	< 0.001	0.120 (0.041)	0.004
	Agemen (X_2)	0.123 (0.042)	0.003	0.134 (0.047)	0.004
	Nbiops (X_3)	0.240 (0.049)	< 0.001	0.181 (0.069)	0.009
	Numrel (X_4)	0.641 (0.061)	< 0.001	0.652 (0.089)	< 0.001
Parameters β	Weight (X_5)	0.177 (0.030)	< 0.001	0.229 (0.043)	< 0.001
	Density (Z^c)	0.412 (0.039)	< 0.001	0.427 (0.044)	< 0.001
Mean Parameters	Intercept	1.551 (0.131)	< 0.001		
	Ageflb (X_1)	0.137 (0.023)	< 0.001		
	Agemen (X_2)	-0.106 (0.028)	< 0.001		
	Nbiops (X_3)	0.245 (0.028)	< 0.001		
	Weight (X_5)	-0.384 (0.023)	< 0.001		
γ_{mean}	Age (A)	-0.032 (0.002)	< 0.001		
	Intercept	1.180 (0.178)	< 0.001		
Precision Parameters	Nbiops (X_3)	0.129 (0.044)	0.003		
	Weight (X_5)	-0.090 (0.031)	0.003		
	Age (A)	0.008 (0.003)	0.008		
$\gamma_{\text{precision}}$					
Mixture Proportion ρ		0.109 (0.008)	< 0.001		

The ML and PL estimates of the log odds ratio association parameters were close, and the variance was 20% $\sim$ 53% smaller by the ML method. For Ageflb, Agemen and Nbioips, the ML method yielded smaller p values. To check the goodness of fit for the model of BD, we performed a Pearson's Chi-squared Test on the observed and estimated BD categories in controls. The p value when the expected counts were calculated using the ML estimates was 0.167. Therefore, there was no significant difference between the estimated and observed BD categories in controls, indicating the good fit of the zero-inflated Beta-regression model. To gain further insight into the relative efficiency of the ML method for estimating log OR parameters, we randomly deleted some BD data to increase the missingness probability. Table 2 shows that the variance for β was 23% $\sim$ 72% smaller by the ML method when the missingness probability was 70%, and 26% $\sim$ 91% smaller when the missingness probability was 90%. The efficiency gain was very apparent especially for Nbiops and Numrel, reflecting the information loss of PL method in the construction of the strata $K(\mathbf{X})$. Interestingly, the ML estimates were very close at different missingness proportions for BD, but the PL estimates varied quite noticeably. For example, the log odds ratio ML estimates for Numrel were 0.641, 0.641, and 0.641, respectively, but the PL estimates were 0.652, 0.809, and 1.144.

Table 2: Analysis of the BCDDP data using the ML and PL methods when the missingness probability was increased to 70% or 90%: The estimate (Est) and asymptotic standard error (SE) of the log odds ratio parameter estimates (β) and the parameters in the conditional distribution of Z , $(\gamma_{\text{mean}}^T, \gamma_{\text{precision}}^T, \rho)$.

		Missingness Probability 70%		Missingness Probability 90%	
		ML	PL	ML	PL
		Est (SE)	Est (SE)	Est (SE)	Est (SE)
Log Odds Ratio Parameters β	Ageflb (X_1)	0.148 (0.035)	0.123 (0.046)	0.123 (0.039)	0.137 (0.063)
	Agemen (X_2)	0.118 (0.042)	0.149 (0.050)	0.108 (0.046)	0.185 (0.064)
	Nbiops (X_3)	0.231 (0.050)	0.154 (0.091)	0.206 (0.056)	0.149 (0.158)
	Numrel (X_4)	0.641 (0.061)	0.809 (0.115)	0.641 (0.061)	1.144 (0.208)
	Weight (X_5)	0.175 (0.032)	0.230 (0.056)	0.177 (0.038)	0.280 (0.099)
	Density (Z^c)	0.417 (0.050)	0.454 (0.057)	0.462 (0.085)	0.459 (0.099)
Mean Parameters γ_{mean}	Intercept	1.590 (0.164)		1.106 (0.284)	
	Ageflb (X_1)	0.126 (0.029)		0.182 (0.048)	
	Agemen (X_2)	-0.088 (0.035)		-0.052 (0.061)	
	Nbiops (X_3)	0.272 (0.035)		0.316 (0.057)	
	Weight (X_5)	-0.375 (0.028)		-0.338 (0.046)	
	Age (A)	-0.033 (0.003)		-0.028 (0.004)	
Precision Parameters $\gamma_{\text{precision}}$	Intercept	1.221 (0.223)		0.774 (0.374)	
	Nbiops (X_3)	0.171 (0.055)		0.233 (0.098)	
	Weight (X_5)	-0.112 (0.038)		-0.065 (0.070)	
	Age (A)	0.008 (0.004)		0.015 (0.007)	
Mixture Proportion ρ		0.101 (0.011)		0.118 (0.019)	

4. Simulation Study

In this section, we conducted extensive simulation studies to compare the unbiasedness and statistical efficiency of the proposed ML estimator and the PL estimator. We also evaluated the robustness of our method with respect to violation of the rare disease assumption and mis-specification of the

---

model for Z . Firstly, we generated a large random sample of data for variable A , conventional predictors $\mathbf{X} = \{(X_1, X_2, X_3, X_4, X_5)^T\}$ and new predictor Z . A was generated from the Uniform distribution in the range of (30, 80). The matching variable A^c was obtained as a categorized version of A , taking values 1 to 5 corresponding to intervals $[30, 40)$, $[40, 50)$, $[50, 60)$, $[60, 70)$, and $[70, 80)$ respectively. Variables $\mathbf{X}$ were generated from similar distributions of the corresponding variables observed in the BCDDP controls as follows. X_1 and X_2 were generated independently from a multinomial distribution with the probabilities $(0.1, 0.35, 0.43, 0.12)$ and $(0.3, 0.55, 0.15)$, respectively. X_3 took values 0, 1, or 2 corresponding to values 0, 1 or ≥ 2 generated from the Poisson(0.5) distribution. X_4 took values 0 and 1 corresponding to values 0 or ≥ 1 generated from the Poisson(1) distribution. X_5 was generated by categorizing data from the truncated normal distribution $TN(140, 25^2; 80, 300)$, taking values 0, 1, 2, and 3 corresponding to intervals $(80, 125]$, $(125, 150]$, $(150, 175]$, and $(175, 300]$. Z was generated from the Beta distribution $f_{\gamma}(Z = z | \mathbf{X}, A)$ as specified in the data analysis above, with γ_{mean}^T and $\gamma_{\text{precision}}^T$ set to be $(1.5, 0.15, -0.1, 0.2, 0.05, -0.4, -0.03)$ and $(1.2, -0.08, -0.05, 0.1, 0.05, -0.1, 0.01)$, respectively. Z^c took values 0, 1, 2 and 3 corresponding to Z intervals $(0, 0.25)$, $[0.25, 0.50)$, $[0.50, 0.75)$ and $[0.75, 1)$, respectively. Lastly, we generated the outcome status from the lo-

---

gistic regression model (3.8) with $a = 1, 2, \dots, 5$, $\beta_1^T = (\beta_{1-X_1}, \beta_{1-X_2}, \beta_{1-X_3}, \beta_{1-X_4}, \beta_{1-X_5}) = (0.15, 0.1, 0.2, 0.65, 0.2)$, and $\beta_2 = \beta_{2-Z^c} = 0.4$. The stratum-specific intercept parameters $\alpha = (\alpha_1, \dots, \alpha_5)^T$ were chosen to have $\Pr(Y = 1) = 0.035$ and $\Pr(Y = 1 \mid A^c = a) = 1\%, 2\%, 3\%, 5\%, 6.5\%$ for $a = 1, \dots, 5$, respectively. We then selected via variable probability sampling 50, 100, 200, or 400 cases and the same number of controls from each of the five A^c matching strata, resulting in a total sample size of $n = 500, 1000, 2000$ or 4000 . To create a two-phase stratified case-control sample, we deleted data Z and Z^c from a proportion of cases and controls randomly, and the proportion ranged from $10\% \sim 90\%$. We applied ML and PL methods in different scenarios. However, PL cannot estimate parameters γ in the model for Z . Chen et al. (2007) directly fit $\Pr(Z \mid \mathbf{X}, A; \gamma)$ in controls to roughly resolve this problem because the breast cancer is rare. We view their analysis as an existing method to estimate γ and report the results for comparison. The simulation was repeated 1000 times.

Results summarized in Figure 1 and Table 3 demonstrate the consistency and high statistical efficiency of the proposed ML method. In the above simulation scenarios, the mean estimates were close to the true values and the averaged standard error estimates were close to the simulation standard error. The coverage was near the nominal value 95% for all esti-

---

mates. In Figure 1 where the missingness probability was 50%, the averaged estimates appeared to be close to their true values across all four sample sizes considered, and the variances of the ML estimates were all smaller than those of the PL. At sample size 1000, the relative efficiency (“RE”) of the ML method was 1.22~1.87 for estimating β and as high as above 2 for estimating γ . For some parameters such as the third component of $\gamma_{\text{precision}}$, the ML estimate can be even more efficient than the PL estimate obtained with doubled sample size. The RE for estimating β was larger than 1, and slowly decreased with increasing sample sizes. The RE for estimating γ was about 2 across all sample sizes, suggesting that the efficiency gain is mainly due to the incorporation of cases and the information from Phase I data. Table 3 shows the efficiency gain increase of the ML method with respect to increased missingness probability at sample size $n = 2,000$. When the missingness probability increased from 10% to 90%, the RE of the ML estimates increased from 1.039 to 2.041 for β_{X_1} and from 1.202 to 1.474 for β_Z . The RE for γ also increased but the increase was much smaller. These results indicate that the ML has superior performance when a large proportion of missingness occurs, but it can also have reasonable efficiency advantage even when the missingness probability is small. The results under sample sizes $n = 500$ and $n = 1,000$ were largely similar and

are presented in Tables S4 and S5 in Supplementary Materials.

Figure 1: The mean estimates and corresponding 95% confidence intervals for the parameters in the disease-risk model (β) and the parameters in the conditional distribution of Z ($\gamma_{\text{mean}}, \gamma_{\text{precision}}$). The relative efficiency (“RE”) of the ML estimates compared to PL estimates when sample size $n = 1,000$ is noted above each sub-figure. The prevalence was 0.035 and the missingness probability of Z in Phase II was 50%.

Table 3: Results of 1,000 repetitions as described in Section 4 when the prevalence was 0.035, the sample size $n = 2,000$ and missingness probability of Z was 10%, 30%, 50%, 70% and 90%.

Parameter	True	Missingness Probability=0.1			Missingness Probability=0.3			Missingness Probability=0.5			Missingness Probability=0.7			Missingness Probability=0.9		
		Mean	SE	SE	Mean	SE	SE	Mean	SE	SE	Mean	SE	SE	Mean	SE	SE
β_{1X_1}	0.15	0.153	0.057	0.057	0.153	0.057	0.057	0.151	0.059	0.058	0.148	0.059	0.059	0.152	0.064	0.066
β_{1X_2}	0.1	0.102	0.067	0.071	0.099	0.073	0.071	0.101	0.073	0.072	0.100	0.073	0.073	0.105	0.083	0.080
β_{1X_3}	0.2	0.203	0.062	0.061	0.207	0.061	0.061	0.204	0.062	0.062	0.206	0.064	0.063	0.207	0.070	0.071
β_{1X_4}	0.65	0.648	0.111	0.111	0.656	0.110	0.111	0.658	0.109	0.112	0.651	0.114	0.113	0.660	0.124	0.123
β_{1X_5}	0.2	0.205	0.053	0.054	0.203	0.054	0.055	0.204	0.056	0.057	0.207	0.062	0.061	0.205	0.078	0.078
β_{2Z^c}	0.4	0.393	0.052	0.054	0.392	0.060	0.061	0.391	0.070	0.072	0.396	0.068	0.093	0.399	0.155	0.159
$\gamma_{\text{mean},0}$	1.5	1.523	0.101	0.102	1.525	0.117	0.115	1.519	0.134	0.137	1.526	0.180	0.177	1.541	0.316	0.307
γ_{mean,X_1}	0.15	0.146	0.024	0.024	0.146	0.028	0.028	0.147	0.034	0.033	0.146	0.044	0.042	0.149	0.073	0.074
γ_{mean,X_2}	-0.1	-0.103	0.030	0.031	-0.100	0.034	0.035	-0.102	0.042	0.041	-0.100	0.054	0.053	-0.106	0.099	0.093
γ_{mean,X_3}	0.2	0.200	0.026	0.026	0.199	0.030	0.029	0.199	0.034	0.035	0.199	0.047	0.045	0.199	0.081	0.078
γ_{mean,X_4}	0.05	0.044	0.049	0.047	0.047	0.055	0.053	0.043	0.064	0.063	0.046	0.086	0.081	0.045	0.148	0.142
γ_{mean,X_5}	-0.4	-0.400	0.023	0.022	-0.400	0.026	0.025	-0.402	0.029	0.030	-0.400	0.039	0.039	-0.398	0.069	0.068
$\gamma_{\text{mean},A}$	-0.03	-0.030	0.001	0.001	-0.030	0.002	0.002	-0.030	0.002	0.002	-0.030	0.003	0.002	-0.031	0.005	0.004
$\gamma_{\text{precision},0}$	1.2	1.202	0.148	0.149	1.196	0.164	0.169	1.197	0.197	0.200	1.212	0.269	0.261	1.240	0.485	0.471
$\gamma_{\text{precision},X_1}$	-0.08	-0.080	0.036	0.036	-0.081	0.039	0.041	-0.081	0.050	0.049	-0.082	0.065	0.064	-0.086	0.121	0.116
$\gamma_{\text{precision},X_2}$	-0.05	-0.052	0.046	0.045	-0.055	0.051	0.051	-0.050	0.062	0.061	-0.048	0.080	0.080	-0.045	0.155	0.146
$\gamma_{\text{precision},X_3}$	0.1	0.101	0.039	0.039	0.102	0.043	0.044	0.101	0.050	0.052	0.102	0.068	0.068	0.101	0.128	0.123
$\gamma_{\text{precision},X_4}$	0.05	0.057	0.071	0.070	0.056	0.081	0.080	0.053	0.094	0.094	0.063	0.126	0.123	0.078	0.232	0.222
$\gamma_{\text{precision},X_5}$	-0.1	-0.096	0.033	0.033	-0.096	0.038	0.037	-0.098	0.044	0.044	-0.097	0.057	0.057	-0.103	0.107	0.104
$\gamma_{\text{precision},A}$	0.01	0.010	0.002	0.002	0.010	0.002	0.002	0.010	0.003	0.003	0.010	0.004	0.004	0.011	0.007	0.007

True, true value; Mean, mean estimates; SE, simulation standard error; $\widehat{\text{SE}}$, average asymptotic standard error estimates; RE, the relative efficiency of ML vs PL for estimating β , and the relative efficiency of ML method compared to the control-only analysis for estimating γ .

To examine robustness of the proposed ML method with respect to the outcome prevalence, we set the prevalence as $\Pr(Y = 1) = 0.1$ with $\Pr(Y = 1 \mid A^c = a) = 3\%, 6\%, 9\%, 14\%, 18\%$, $a = 1, \dots, 5$, or $\Pr(Y = 1) = 0.2$ with $\Pr(Y = 1 \mid A^c = a) = 6\%, 12\%, 18\%, 28\%, 36\%$, $a = 1, \dots, 5$. The estimation for β is presented in Table 4 where the sample size was $n = 1,000$ and missingness probability of Z was 50%. The biases appeared to be minimal and nearly remained unchanged with increasing outcome prevalence. The largest bias occurred for $\hat{\beta}_Z$, which was only about 4% with $\Pr(Y = 1) = 0.2$ and 13% with $\Pr(Y = 1) = 0.5$ (unreported result). The ML estimates had smaller averaged MSE than the PL estimates in all parameter settings, and they were nearly identical under different values for $\Pr(Y = 1)$. Both ML and PL methods had coverage probabilities close to the nominal level. The averaged asymptotic standard error remained close to the simulation standard error. The estimates for γ are presented in Supplementary Table S6. The percent biases ranged from 0% to 46% for the ML estimates and from 2% to 96% for the control-only analysis when $\Pr(Y = 1)$ was equal to 0.2.

To evaluate robustness of the proposed ML method for estimating β when the model for the conditional distribution of Z is misspecified, we fitted the same model (3.9) but with constant precision parameters $\gamma =$

$(\gamma_{\text{mean}}^T, \phi)^T$. The Pearson's Chi-squared goodness of fit test indicated that this model did not fit well in controls (p value = 0.03). The results are shown in Table 4 when the sample size was $n = 1,000$, the missingness probability of Z was 50%, and the prevalence was 0.035. The ML estimates remained to be close to the true values and had smaller MSEs. The largest bias occurred for β_Z which was only about 5%. The coverage probabilities were close to 95%, and the averaged asymptotic standard error remained close to the simulation standard error. Additional simulation to investigate the finite sample performance under other model for Z is presented in the Supplementary Materials.

5. Discussion

It is widely known that the intercept parameter is unidentifiable from the case-control data when the distribution of covariates is unrestricted (Prentice and Pyke, 1979). The stratum-specific intercept parameters $\alpha = (\alpha_1, \dots, \alpha_S)^T$ in model (2.1) are similarly unidentifiable. This is no longer the case when structural constraints are imposed onto the covariate distribution. For assessing gene-environment interactions under a logistic regression model with case-control data, the gene-environment independence

Table 4: *Simulation results under different prevalence for Y or when the conditional distribution of Z is modestly misspecified.*

		Mean		MSE×100		SE×100 ($\widehat{SE} \times 100$)		CP	
		PL	ML	PL	ML	PL	ML	PL	ML
Pr($Y = 1$) = 0.1	$\beta_{1,X_1} = 0.15$	0.149	0.149	0.9	0.7	9.3 (8.6)	8.3 (8.2)	92.4	94.4
	$\beta_{1,X_2} = 0.1$	0.102	0.101	1.4	1.1	11.7(10.8)	10.5(10.2)	93.3	94.6
	$\beta_{1,X_3} = 0.2$	0.210	0.209	1.3	0.8	11.4(10.7)	9.0 (8.8)	93.6	95.1
	$\beta_{1,X_4} = 0.65$	0.666	0.651	5.2	2.6	22.7(22.8)	16.1(16.1)	96.4	95.7
	$\beta_{1,X_5} = 0.2$	0.199	0.195	1.3	0.7	11.4(11.0)	8.4 (8.1)	95.0	93.4
	$\beta_{2,Z^c} = 0.4$	0.407	0.392	1.4	1.1	11.7(11.4)	10.5(10.3)	95.3	94.7
Pr($Y = 1$) = 0.2	$\beta_{1,X_1} = 0.15$	0.153	0.153	0.9	0.7	9.3 (8.6)	8.3 (8.2)	93.2	94.1
	$\beta_{1,X_2} = 0.1$	0.103	0.100	1.4	1.1	11.9(10.8)	10.6(10.2)	92.5	94.1
	$\beta_{1,X_3} = 0.2$	0.206	0.209	1.3	0.8	11.5(11.1)	8.9 (8.8)	93.8	94.4
	$\beta_{1,X_4} = 0.65$	0.680	0.665	6.0	2.6	24.2(23.4)	16.1(16.5)	94.2	96.0
	$\beta_{1,X_5} = 0.2$	0.211	0.197	1.3	0.7	11.1(11.0)	8.1 (8.1)	94.5	95.2
	$\beta_{2,Z^c} = 0.4$	0.411	0.383	1.2	1.0	11.1(11.4)	10.1(10.3)	95.8	94.3
Modest mis-specification for the conditional distribution of Z	$\beta_{1,X_1} = 0.15$	0.152	0.157	0.8	0.7	8.8 (8.3)	8.2 (8.1)	94.2	94.4
	$\beta_{1,X_2} = 0.1$	0.108	0.103	1.2	1.0	10.9(10.4)	10.2(10.1)	94.0	95.2
	$\beta_{1,X_3} = 0.2$	0.207	0.206	1.0	0.7	9.8 (9.6)	8.6 (8.7)	94.5	95.9
	$\beta_{1,X_4} = 0.65$	0.650	0.650	3.6	2.4	19.1(18.9)	15.6(15.8)	95.6	95.7
	$\beta_{1,X_5} = 0.2$	0.214	0.202	0.9	0.6	9.2 (9.2)	7.7 (7.8)	94.8	95.6
	$\beta_{2,Z^c} = 0.4$	0.410	0.381	1.0	0.8	10.1 (9.5)	8.7 (8.4)	94.0	93.7

Mean, the mean estimates; SE, simulation standard error; $\widehat{SE}$, average asymptotic standard error estimates; CP, 95% coverage probabilities; MSE was computed by averaging $(\beta - \hat{\beta})^2$ over 1000 repetitions.

constraint can be exploited to increase statistical efficiency (Chatterjee and Carroll, 2005). Interestingly, this constraint was found to result in identifiability of the intercept parameter. The information for estimating the intercept parameter turns out to be very sparse, as evidenced by much smaller efficiency gain for estimating gene-environment interaction odds ratio parameter (Chen and Chen, 2011). In this work, we considered the outcome variable to be rare to eliminate the need for estimating the intercept parameter. Simulation studies showed that our method is robust to the

---

rare outcome requirement, where the parameter estimates were reasonably close to the true values even when the outcome prevalence was 20%. Unreported simulation results showed that some parameter estimates, especially those for Phase II variables, were biased when the outcome prevalence was 50%. In practice, when there is concern on the rare outcome assumption, we suggest external information of $\Pr(Y = 1|A = a)$ be used to resolve the problem of estimating the intercept parameters.

For analyzing standard case-control data, the prospective logistic regression analysis that ignores retrospective sampling is equivalent to retrospective likelihood analysis (Prentice and Pyke, 1979) even in the presence of missing data (Roeder et al., 1996). When constraints are enforced on the covariate distribution, however, the prospective logistic regression analysis ignoring the information on the covariate distribution is valid but no longer semiparametric efficient. Under gene-environment independence assumption, Chatterjee and Carroll (2005) developed an efficient retrospective maximum likelihood approach by maximizing a “modified” prospective likelihood, which is different from the likelihood function for the prospective logistic regression analysis (Chatterjee et al., 2006). Such approach of solving “modified” prospective score functions to obtain ML estimates for odds ratio parameters has been extended to exploit different covariate

---

distribution constraints (Chen et al., 2007; Spinka et al., 2005; Mukherjee and Chatterjee, 2008). In this work, we established a key theoretical result that the stratified two-phase case-control data can be seen as arising from a novel “modified” prospective distribution under the rare disease assumption and a parametric model for Phase II variables, which leads to convenient statistical inference.

In real data analysis, we compared our proposed ML method with the PL method (Chen et al., 2008). Although more finer construction of strata $K(\mathbf{X})$ in (2.7) may improve the efficiency for the PL method, its application is limited by a practical trade-off exists between the number of strata and the number of cases and controls in each stratum. Additional exploration also showed ML estimates had smaller variance than PL estimates as long as the information of $\mathbf{X}$ was not fully captured in $K(\mathbf{X})$. We described our ML method considering a univariate Phase II variable, but our method is general and can readily accommodate multivariate Phase II variables if a parametric model can be specified for $\Pr(\mathbf{Z}|\mathbf{X}, A)$. Additional simulation results for bivariate Phase II variable (Z_1, Z_2) are presented in the Supplementary Materials. We demonstrated the robustness of our method through simulation studies under modest mis-specification of the conditional distribution for the Phase II variables. But the bias would become larger should

---

the mis-specification be severe. Therefore, caution needs to be exercised for deciding suitable distributions. Exploration of the model forms and check on the goodness-of-fit using controls' data as in our real data analysis can be highly useful in this regard. When Phase II variables are of low dimension, it is potentially feasible to adopt semiparametric or non-parametric model for joint distribution. In the absence of Phase I covariates $\mathbf{X}$, Song et al. (2009) developed a nonparametric maximum likelihood method where $P(\mathbf{Z})$ was estimated nonparametrically. With multivariate $\mathbf{Z}$ and $\mathbf{X}$, it is generally challenging to deal with $\Pr(\mathbf{Z}|\mathbf{X})$ likelihood-based methods. In a recent ML method for analyzing prospective two-phase data (Tao et al., 2017), a sieve approximation was adopted for $\Pr(\mathbf{Z}|\mathbf{X})$ which appeared to work well when $\mathbf{X}$ is of modest dimension. We will explore extensions of our method along this line in future work. We focused on increasing efficiency for parameter estimation in the current work. The estimated odds ratio function can be used as a risk score for risk assessment. Then it is of interest to estimate the corresponding predictive accuracy measures such as the area under the receiver operating characteristic curve (AUC), for which the estimated conditional distribution of the Phase II variables is needed. We will consider statistical inference procedures for estimating these measures using two-phase stratified case-control data in future work.

REFERENCES

Supplementary Materials

Supplementary material available online includes appendices and tables referenced in Section 2, 3 and 4.

Acknowledgments

Drs. Jinbo Chen, Lu Chen, and Yaqi Cao were supported by NIH grants R01-ES016626 and R01-CA236468-01A1. Drs. Ying Yang and Yaqi Cao were supported by National Natural Science Foundation of China, No. 11771241 and 11931001.

References

Breslow, N. E. and K. C. Cain (1988). Logistic regression for two-stage case-control data.

Biometrika 75(1), 11–20.

Breslow, N. E. and N. Chatterjee (1999). Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. *Applied Statistics* 48(4), 457–468.

Breslow, N. E. and R. Holubkov (1997a). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society, Series B* 59(2), 447–461.

Breslow, N. E. and R. Holubkov (1997b). Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Statistics in Medicine* 16(1), 103–116.

REFERENCES

- Byrne, C., C. Schairer, J. Wolfe, N. Parekh, M. Salane, L. A. Brinton, R. Hoover, and R. Haile (1995). Mammographic features and breast cancer risk: Effects with time, age, and menopause status. *Journal of the National Cancer Institute* 87(21), 1622–1629.
- Chatterjee, N. and R. J. Carroll (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* 92(2), 399–418.
- Chatterjee, N. and Y.-H. Chen (2007). 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, Series B* 69(2), 123–142.
- Chatterjee, N., C. Spinka, J. Chen, and R. J. Carroll (2006). Comment. *Journal of the American Statistical Association* 101(473), 108–111.
- Chen, H. Y. and J. Chen (2011). On information coded in gene-environment independence in case-control studies. *American Journal of Epidemiology* 174(6), 736–743.
- Chen, J., R. Ayyagari, N. Chatterjee, D. Pee, C. Schairer, C. Byrne, J. Benichou, and M. H. Gail (2008). Breast cancer relative hazard estimates from case-control and cohort designs with missing data on mammographic density. *Journal of the American Statistical Association* 103(483), 976–988.
- Chen, J., D. Pee, R. Ayyagari, B. Graubard, C. Schairer, C. Byrne, J. Benichou, and M. H. Gail (2006). Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. *Journal of the National Cancer Institute* 98(17), 1215–1226.

REFERENCES

- Chen, Y.-H., N. Chatterjee, and R. J. Carroll (2007). Retrospective analysis of haplotype-based case-control studies under a flexible model for gene-environment association. *Biostatistics* 9(1), 81–99.
- Fears, T. R. and C. C. Brown (1986). Logistic regression methods for retrospective case-control studies using complex sampling procedures. *Biometrics* 42(4), 955–960.
- Gail, M. H., L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, and J. J. Mulvihill (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute* 81(24), 1879–1886.
- Lawless, J. F., J. D. Kalbfleisch, and C. J. Wild (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society, Series B* 61(2), 413–438.
- Little, R. J. and D. B. Rubin (1987). *Statistical analysis with missing data*, Volume 793. John Wiley & Sons.
- Mukherjee, B. and N. Chatterjee (2008). Exploiting gene-environment independence for analysis of case-control studies: an empirical bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics* 64(3), 685–694.
- Neyman, J. (1938). Contribution to the theory of sampling from human populations. *Journal of the American Statistical Association* 33(201), 101–116.
- Prentice, R. L. and R. Pyke (1979). Logistic disease incidence models and case-control studies.

REFERENCES

- Biometrika* 66(3), 403–411.
- Roeder, K., R. J. Carroll, and B. G. Lindsay (1996). A semiparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association* 91(434), 722–732.
- Satten, G. and L. Kupper (1993). Inferences about exposure-disease associations using probability-of-exposure information. *Journal of the American Statistical Association* 88(421), 200–208.
- Schill, W., K.-H. Jockel, K. Drescher, and J. Timm (1993). Logistic analysis in case-control studies under validation sampling. *Biometrika* 80(2), 339–352.
- Scott, A. J. and C. Wild (2001). Maximum likelihood for generalized case-control studies. *Journal of Statistical Planning and Inference* 96(1), 3–27.
- Scott, A. J. and C. J. Wild (1991). Fitting logistic regression models in stratified case-control studies. *Biometrics* 47(2), 497–510.
- Scott, A. J. and C. J. Wild (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* 84(1), 57–71.
- Song, R., H. Zhou, and M. R. Kosorok (2009). A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome. *Biometrika* 96(1), 221–228.
- Spinka, C., R. J. Carroll, and N. Chatterjee (2005). Analysis of case-control studies of genetic

REFERENCES

---

and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genetic Epidemiology* 29(2), 108–127.

Tao, R., D. Zeng, and D.-Y. Lin (2017). Efficient semiparametric inference under two-phase sampling, with applications to genetic association studies. *Journal of the American Statistical Association* 112(520), 1468–1476.

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(1), 119–128.

Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, U.S.A.

Department of Mathematical Sciences, Tsinghua University, Beijing, P.R.China.

E-mail: Yaqi.Cao@pennmedicine.upenn.edu

Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, U.S.A.

E-mail: daisychen1009@gmail.com

Department of Mathematical Sciences, Tsinghua University, Beijing, P.R.China.

E-mail: yangying@tsinghua.edu.cn

Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, U.S.A.

E-mail: jinboche@pennmedicine.upenn.edu