Statistica Sinica **33** (2023), 2233-2256
doi:https://doi.org/10.5705/ss.202021.0214

SEMIPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATION WITH TWO-PHASE STRATIFIED CASE-CONTROL SAMPLING

Yaqi Cao[1*], Lu Chen[1*], Ying Yang[2] and Jinbo Chen[1]

[1] *University of Pennsylvania and* [2] *Tsinghua University*

Abstract: We develop 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. In addition, we characterize the distribution of incomplete covariates, conditional on fully observed ones. Here, we include all subjects in the analysis in order to achieve consistency in the parameter estimation and optimal statistical efficiency. We develop a semiparametric maximum likelihood approach under the rare disease assumption, where the parameter estimates are obtained using a novel reparametrized profile likelihood technique. We study the large-sample distribution theory for the proposed estimator, and use simulations to demonstrate that it performs well in finite samples and improves on the statistical efficiency of existing approaches. We apply the proposed method to analyze a stratified case-control study of breast cancer nested within the Breast Cancer Detection and Demonstration Project, where one breast cancer risk predictor, namely, percent mammographic density, was measured only for a subset of the women in the study.

Key words and phrases: Logistic regression model, profile likelihood, semiparametric maximum likelihood, stratified case-control study, two-phase sampling.

1. Introduction

In a stratified case-control study design, cases and controls are selected by stratified matching from population subgroups with or without the outcome of interest. The logistic regression model is most frequently adopted to relate the binary outcome status with the covariates of interest, where the stratified sampling is accounted for by adjusting matching strata in the model. Specialized methods have been studied when the covariates are fully observed and interest lies in the effect of the covariates involved in forming the matching strata

*Contributed equally to this work.

Corresponding author: Jinbo Chen, Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA. E-mail: jinboche@pennmedicine.upenn.edu.

(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 conform to a missing-at-random mechanism (Little and Rubin (1987)). We develop statistical methods to achieve two analytical goals, namely, to obtain consistent and efficient estimates of the odds ratio (OR) association parameters for the covariates collected in both phases, and to estimate the distribution of the incomplete covariates, conditional on the observed ones. The second goal is important when we need to quantify the added value of incompletely measured covariates, such as expensive biomarkers, for prediction.

The two-phase sampling design (Neyman (1938); White (1982)) is widely used as a cost-effective option for data collection. For binary outcomes, numerous statistical methods have been proposed for analyzing two-phase data when Phase I consists of a cross-sectional or unstratified case-control sample (Breslow and Chatterjee (1999); Breslow and Holubkov (1997b); Lawless, Kalbfleisch and Wild (1999); Chatterjee and Chen (2007)). These methods focus on improving the statistical efficiency for a consistent estimation of the OR 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. Although the data from a stratified case-control design are analyzed in essentially the same way as standard case-control data, available statistical methods for analyzing two-phase case-control data cannot accommodate the stratified matching in Phase I. When Phase II subjects are selected performed within the 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 analyses can be highly inefficient (Chen et al. (2008)), and may yield biased estimates when the Phase II sampling does not fully comply with the Phase I stratified matching.

We develop a novel semiparametric maximum likelihood (ML) approach when a parametric regression model is imposed for the conditional distribution of the incomplete covariates. Our method yields consistent and efficient estimates for all of the parameters simultaneously. In the same spirit as a prospective analysis of standard case-control data (Prentice and Pyke (1979)), our method does not impose structural constraints on the distribution of the fully observed covariates.

This nuisance distribution function is eliminated from the empirical retrospective likelihood function using 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 addresses the challenge when the selection of Phase II subjects is stratified on variables other than those used in the Phase I matching. The PL requires an enumeration of the 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 an estimate of the conditional distribution of the incomplete covariates. In Section 3, we demonstrate our method using data from the Breast Cancer Detection and Demonstration Project (BCDDP). Section 4 presents the results from extensive simulation studies used to evaluate the finite-sample performance of our method compared with that of the PL. Section 5 concludes the paper.

2. Methods

2.1. Notation and model assumptions

Let Y denote the case-control status ($Y = 1$: case; $Y = 0$: control) and $(\boldsymbol{X}, Z)$ denote the covariates of interest, where $\boldsymbol{X}$ is a $p \times 1$ vector, $\boldsymbol{X} = (X_1, \ldots, X_p)^T$, and Z is univariate. Let A denote matching strata taking the value a, for $a = 1, \ldots, S$, where A can be defined based on $\boldsymbol{X}$. In Phase I, n_{1a} cases and n_{0a} controls are sampled from the conditional distributions $\Pr(\boldsymbol{X} \mid Y = 1, A = a)$ and $\Pr(\boldsymbol{X} \mid Y = 0, A = a)$, respectively, for $a = 1 \ldots, S$, for which $\boldsymbol{X}$ is observed. The Phase I sample then consists of 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 by Bernoulli sampling, with the success probability $\omega(Y, \boldsymbol{X}, A)$ depending on the observed data $(Y, \boldsymbol{X}, A)$. Let R denote the selection decision ($R = 1$: yes; $R = 0$: no). The observed data then consist of $\mathcal{O} = \bigcup_{y=0,1} \bigcup_{a=1,\ldots,S} \mathbf{O}_{ya} \equiv \{(R_i, \boldsymbol{X}_i^T, Z_i R_i)^T, i = 1, \ldots, n_{ya}\}$, 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, that is, $\omega(Y, \boldsymbol{X}, A) \equiv \Pr(R \mid Y, \boldsymbol{X}, A) = \Pr(R \mid Y, \boldsymbol{X}, Z, A)$.

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

$$\Pr(Y = 1 \mid \boldsymbol{X}, Z; A = a) \equiv p_{\text{risk}}(Y = 1, A = a, \boldsymbol{X}, Z; \alpha_a, \boldsymbol{\beta})$$

$$= \frac{\exp(\alpha_a + \boldsymbol{\beta}_1^T \boldsymbol{X} + \beta_2 Z)}{1 + \exp(\alpha_a + \boldsymbol{\beta}_1^T \boldsymbol{X} + \beta_2 Z)}, \quad a = 1, \ldots, S, \quad (2.1)$$

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 $(\alpha_1, \ldots, \alpha_S)^T$. It is also of interest to describe the distribution of Z conditional on $(\boldsymbol{X}, A)$, $\Pr(Z \mid \boldsymbol{X}, A)$. Because $(\boldsymbol{X}, A)$ may have a large number of values, modeling $\Pr(Z \mid \boldsymbol{X}, A)$ nonparametrically is challenging in practice. Therefore, we adopt a parametric model for $\Pr(Z \mid \boldsymbol{X}, A)$, denoted as $p_{\boldsymbol{\theta}}(Z \mid \boldsymbol{X}, A)$, with $\boldsymbol{\theta}$ being the index parameter vector. We develop statistical methods to jointly estimate $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ using the observed data $\mathcal{O}$. In subsequent development, we do not impose a model structure on the distribution $\Pr(\boldsymbol{X} \mid A = a)$, because it is not of interest, and it is also infeasible to impose reasonable multivariate parametric models for this whole vector of covariates. We assume that the outcome is rare (Spinka, Carroll and Chatterjee (2005); Chen, Chatterjee and Carroll (2007)). Then, $\Pr(\boldsymbol{X} \mid A; Y = 0)$ and $\Pr(Z \mid \boldsymbol{X}, A; Y = 0)$ can be approximated by $\Pr(\boldsymbol{X} \mid A)$ and $p_{\boldsymbol{\theta}}(Z \mid \boldsymbol{X}, A)$, respectively.

2.2. Estimation of OR parameters and conditional distribution of Z

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

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

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 $\ell_0(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\delta})$. Using a result in Satten and Kupper (1993), $\Pr(\boldsymbol{X}, Z \mid A; Y = 1)$ is related to $\Pr(\boldsymbol{X}, Z \mid A; Y = 0)$, as follows:

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

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

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

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

Below, using the profile likelihood approach, we establish a key result that the 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 $\mu_a = \mu_a(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\delta}_a) = \exp(-\alpha_a)\cdot\{\Pr(Y=1\mid A=a)/\Pr(Y=0\mid A=a)\}$, for $a=1,\ldots,S$. Let $\boldsymbol{\mu}$ denote $(\mu_1,\ldots,\mu_S)$. Recall that the two-phase stratified case-control data arise from the distribution $\Pr(R,\boldsymbol{X},Z\mid Y=y,A=a)$, as defined in Section 2.1. We define a modified two-phase prospective study design in which the data conform to the distribution $P^*(R,Y,\boldsymbol{X},Z\mid A=a)$, as follows. Define a modified logistic regression model, as

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

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 $\boldsymbol{X}$ are collected from $P^*(\boldsymbol{X}=\boldsymbol{x}\mid A=a)$, which is unspecified. At Phase II, Z is collected from a modified conditional distribution $P^*(Z=z\mid \boldsymbol{x},a)$,

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

The modified and actual conditional distributions of Z turn out to be identical,

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

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

$$\begin{aligned}\ell^*_{\mathrm{ML}} = &\sum_{i=1}^{n} \log P^*(R_i \mid Y_i, \boldsymbol{X}_i, A_i) \\ &+ \sum_{i=1}^{n} R_i \log P^*(Y_i, Z_i \mid \boldsymbol{X}_i, A_i) + (1 - R_i) \log P^*(Y_i \mid \boldsymbol{X}_i, A_i).\end{aligned}$$

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

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

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 $\eta^a_{\boldsymbol{x}z} = \exp(\boldsymbol{\beta}_1^T \boldsymbol{x} + \beta_2 z)\ p_{\boldsymbol{\theta}}(Z = z \mid \boldsymbol{X} = \boldsymbol{x}, A = a)$. From (2.3) and (2.4), we obtain

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

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

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 $n_{\boldsymbol{x}z}$ denote the total number of cases and controls, with $\boldsymbol{X} = \boldsymbol{x}$ 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 the equations*

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

Although $\ell^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\mu})$ is not a genuine log-likelihood, we show below that the ML 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 ath stratum, for a in $1, \ldots, S$, we reduce the number of nuisance parameters, that is, the number of unique values of $\boldsymbol{X}$, to one. The form of $S^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\mu})$ is given in Lemma 2.

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

$$h(Y, \boldsymbol{W}, A) = Y \log \left(\frac{n_{1A} \cdot \eta^A_{\boldsymbol{X}Z}}{\mu_A} \right) + (1 - Y) \log(n_{0A}\ p_{\boldsymbol{\theta}}(Z \mid \boldsymbol{X}, A)).$$

Then the "pseudo" score function $S^*(\boldsymbol{\Phi})$ can be written as

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

where $E^*_Z(\cdot \mid Y, \boldsymbol{X}, A)$ and $E^*(\cdot \mid \boldsymbol{X}, A)$ denote the expectations taken with respect to $P^*(Z \mid Y, \boldsymbol{X}, A)$ and $P^*(Y, Z \mid \boldsymbol{X}, A)$, 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 $S^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\mu}) = 0$.

Lemmas 1 and 2 show that the MLE of $\boldsymbol{\beta}$ can be obtained by maximizing the modified prospective likelihood of the form $\Pr(Y = 1 | \boldsymbol{x}, z; A = a, R_{\text{scc}} = 1) \equiv P^*(Y = 1 | \boldsymbol{x}, z; A = a)$ in model (2.3), where 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). Here, $\Pr(R_{\text{scc}} = 1 | Y = y, A = a)$, 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. Proofs of the two lemmas are provided in the Supplementary Material.

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 proportion 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 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 $Q(R, Y, \boldsymbol{W}, A)$ that satisfies*

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

where $g_a(\boldsymbol{x}) = (n_{0a}/n) + (n_{1a}/n) \cdot (\sum_{z'} \eta^a_{\boldsymbol{x}z'}/\mu_a)$ and $F_a(\boldsymbol{x})$ is the distribution of $\boldsymbol{X}$ given $A = a$,

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

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 $S^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\mu})$. 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 the cases cancels out that by the controls.

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

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

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

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

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

Lemma 3. *Let $V^*_{YZ}(\cdot \mid \boldsymbol{X}, A)$ and $V^*_Z(\cdot \mid Y, \boldsymbol{X}, A)$ denote variance functions with respect to $P^*(Y, Z \mid \boldsymbol{X}, A)$ and $P^*(Z \mid Y, \boldsymbol{X}, A)$ respectively. With the function $h(Y, W, A)$, defined in Lemma 2, we show that*

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

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

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

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

Theorem 2. *Under the regularity conditions listed in the Supplementary Material, we have the following:*

(i) The estimating equations $S^*(\mathbf{\Phi}) = \sum_{i=1}^n \varphi(R_i, Y_i, \boldsymbol{X}_i, Z_i, A_i) = \mathbf{0}$ defined in (2.5) have a unique sequence of solutions $\{\hat{\mathbf{\Phi}}_n\}_{n \geqslant 1}$. $\hat{\mathbf{\Phi}}_n \xrightarrow{p} \mathbf{\Phi}_0$, where $\mathbf{\Phi}_0 \equiv (\boldsymbol{\beta}_0^T, \boldsymbol{\theta}_0^T)^T$ is the true value for $(\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$.

(ii) $\sqrt{n}(\hat{\mathbf{\Phi}}_n - \mathbf{\Phi}_0) \xrightarrow{d} N(\mathbf{0}, \Sigma)$, where $\Sigma = \zeta^{-1} - (\zeta^{-1})^T \Omega\, \zeta^{-1}$, with

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

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

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

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

2.4. Comparison with the PL method

Here, we will compare the efficiency of our ML estimator with that of an existing PL estimator (Chen et al. (2008)), described below using our notation. The original PL requires a stratum-specific enumeration of the cases and controls in the cohort used to generate the stratified case-control sample in order to get an estimate of μ_a. Here, we describe PL with μ_a fixed at the true value, because this does not affect the estimation of the OR parameters for $\boldsymbol{X}$ 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, $P^*(Z \mid \boldsymbol{X}, A)$. However, it does not specify a parametric model for $\Pr(Z \mid \boldsymbol{X}, A)$, and the nuisance distribution $P^*(Z \mid \boldsymbol{X}, A)$ is treated nonparametrically instead of using the form (2.4). To do this, the Phase I data $\boldsymbol{X}$ are coarsened into discrete strata, denoted as $K(\boldsymbol{X})$. The PL function is written as

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

The estimation proceeds by maximizing this PL function, where $P^*(\boldsymbol{X}_i, Z_i, A_i \mid K(\boldsymbol{X}_i))$ is profiled out, as in the ML for standard two-phase case-control data (Breslow and Cain (1988); Scott and Wild (1997)). Note that, as revealed by $P^*(\boldsymbol{X}_i, Z_i, A_i \mid K(\boldsymbol{X}_i))$, the age-stratum A and any information in $\boldsymbol{X}$ not captured by $K(\boldsymbol{X})$ are also considered to be incomplete data. Ignoring A from Phase I avoids sparse cells resulting from stratifying on both $\boldsymbol{X}$ and A, allowing

a maximum stratification on $\boldsymbol{X}$ that increases the efficiency of estimating the OR parameters for $\boldsymbol{X}$. The ML and PL methods differ in three important respects. First, the ML method eliminates the need to estimate parameters corresponding to the matching strata in model (2.3) using a rare disease approximation, whereas the PL method does estimate the intercept parameters α_a^*. Second, the ML method adopts a parametric model for the Phase II variable Z. The PL method bypasses this need by enforcing a coarsening of $\boldsymbol{X}$ into strata $K(\boldsymbol{X})$ to facilitate maximizing the PL, which incurs a loss of information on $\boldsymbol{X}$. Third, when our primary interest is to estimate the OR parameters, compared with the complete data analysis using the Phase II data, the ML method achieves an efficiency gain by fully using the Phase I data. However, the extent of the efficiency gain for the PL estimates is tied to the construction of $K(\boldsymbol{X})$. The predictors in $\boldsymbol{X}$ not used to form $K(\boldsymbol{X})$ do not enjoy an efficiency gain.

3. Real-Data Analysis

To demonstrate the proposed method, we develop a multivariate OR function for predicting breast cancer risk using data from BCDDP, which was started in 1973 to assess whether mammographic screening can reduce the morbidity and mortality of breast cancer. The study recruited 243,221 white women between 1973 and 1975, and followed each woman for at least five years. An age-stratified case-control sample was assembled within the 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) were analyzed in relation to the risk of breast cancer in order to develop an OR model for the NCI Breast Cancer Risk Assessment Tool (BCRAT; Gail et al. (1989)). Using the same study sample, it was later shown that incorporating the percent mammographic breast density (BD; Z) into the BCRAT as a new predictor may lead to improved discriminatory accuracy (Chen et al. (2006)). However, the BD data were 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 the analysis, treating the data as if it arose from a two-phase age-stratified case-control sampling design. The stratified case-control data were treated as the Phase I sample, and the subset that had BD data 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 $1, 2, \ldots, 9$, corresponding to the age intervals $< 40, [40, 45), [45, 50), [50, 55), \ldots, [70, 75)$, and $\geqslant 75$ years. Here, we

re-analyze the data using the proposed ML method, additionally characterizing the conditional distribution of BD; see the Supplementary Material, Tables S1 and S2, in Chen et al. (2008) for the descriptive statistics of the predictors. The variable weight, X_5, which is available for all cases and controls, is included as the auxiliary information for BD, because the two are strongly correlated (Byrne et al. (1995)). We followed the same convention as the BCRAT to discretize $\boldsymbol{X} = (X_1, X_2, X_3, X_4, X_5)^T$, and applied both the PL and the ML methods to analyze the data. For the PL method, discrete strata $K(\boldsymbol{X})$ in (2.7) were constructed by the cross-classification of $1(X_1 > 0)$ and $1(X_2 > 1)$, resulting in four strata. Note that the PL function (2.7) cannot be used to estimate $\boldsymbol{\theta}$.

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

$$\Pr(Y = 1 \mid \boldsymbol{X}, Z^c; A^c = a) = \frac{\exp(\alpha_a + \boldsymbol{\beta}_1^T \boldsymbol{X} + \beta_2 Z^c)}{1 + \exp(\alpha_a + \boldsymbol{\beta}_1^T \boldsymbol{X} + \beta_2 Z^c)}, \quad a = 1, \ldots, 9. \tag{3.1}$$

Note that our method does not require a categorization of Z. We use Z^c instead of continuous Z in (3.1) mainly to facilitate the risk calculation. After some initial explorations (see the Supplementary Material), we adopted a mixture of the Bernoulli distribution and beta regression model to parametrize the conditional distribution of the continuous variable Z, given $\boldsymbol{X}$ and continuous age A. Let $f_{\boldsymbol{\gamma}}$ denote the probability density function of the beta distribution, and let $F_{\boldsymbol{\gamma}}$ be the corresponding cumulative distribution function. Let κ denote the mean parameter and ϕ the precision parameter, which are allowed to vary with respect to $(\boldsymbol{X}, A)$, governed by logistic and exponential regression models with regression parameters $\boldsymbol{\gamma}_{\text{mean}}$ and $\boldsymbol{\gamma}_{\text{precision}}$, respectively. Let $\boldsymbol{\gamma}$ denote $(\boldsymbol{\gamma}_{\text{mean}}, \boldsymbol{\gamma}_{\text{precision}})$. The conditional distribution function for Z is then specified as

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

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

$$f_{\boldsymbol{\gamma}}(Z = z \mid \boldsymbol{X}, A) = \frac{\Gamma(\phi)}{\Gamma(\kappa\phi)\Gamma(\phi - \kappa\phi)} z^{\kappa\phi - 1} (1 - z)^{\phi - \kappa\phi - 1}, \tag{3.2}$$

$$\kappa = \text{logit}^{-1} \left\{ (\boldsymbol{\gamma}_{\text{mean}})^T (1, \boldsymbol{X}^T, A)^T \right\}, \tag{3.3}$$

$$\phi = \exp \left\{ (\boldsymbol{\gamma}_{\text{precision}})^T (1, \boldsymbol{X}^T, A)^T \right\}. \tag{3.4}$$

We first fitted a full model that included all of the candidate predictors, age,

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

		ML		PL	
		Est (SE)	p value	Est (SE)	p value
Log Odds Ratio Parameters $\boldsymbol{\beta}$	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
	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 $\boldsymbol{\gamma}_{\text{mean}}$	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		
	Age (A)	-0.032 (0.002)	< 0.001		
Precision Parameters $\boldsymbol{\gamma}_{\text{precision}}$	Intercept	1.180 (0.178)	< 0.001		
	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		
Mixture Proportion ρ		0.109 (0.008)	< 0.001		

Ageflb, Agemen, Nbiops, Numrel, Weight, BD, and Age (data not shown) in all models. The results showed that all predictors are significantly associated with breast cancer risk (3.1). In model $p_{\boldsymbol{\theta}}(Z = z \mid \boldsymbol{X}, A)$, the mixture parameter ρ is significantly different from zero. Numrel is nonsignificant in the model for κ, and Ageflb, Agemen, and Numrel are nonsignificant in the model for ϕ. The results with these nonsignificant predictors eliminated from the corresponding models are presented in Table 1.

The ML and PL estimates of the log OR association parameters are close, with the variance of the ML method being between 20% and 53% smaller. For Ageflb, Agemen, and Nbioips, the ML method yields smaller p-values. To check the goodness of fit for the BD model, we performed a Pearson's chi-squared test on the observed and estimated BD categories in the controls. The p-value when the expected counts are calculated using the ML estimates is 0.167. Therefore, there is no statistically significant difference between the estimated and observed BD categories in the controls, indicating the good fit of the zero-inflated beta regression model. To gain further insight into the relative efficiency of the ML method when estimating the log OR parameters, we randomly deleted some BD data to increase the missingness probability. Table 2 shows that the variance of

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

		Missingness Probability 70%		Missingness Probability 90%	
		ML	PL	ML	PL
		Est (SE)	Est (SE)	Est (SE)	Est (SE)
Log Odds Ratio Parameters $\boldsymbol{\beta}$	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 $\boldsymbol{\gamma}_{\text{mean}}$	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 $\boldsymbol{\gamma}_{\text{precision}}$	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)	

the ML method for $\boldsymbol{\beta}$ is between 23% and 72% smaller when the missingness probability is 70%, and between 26% and 91% smaller when the missingness probability is 90%. The efficiency gain is very apparent, especially for Nbiops and Numrel, reflecting the information loss of the PL method when constructing of the strata $K(\boldsymbol{X})$. Interestingly, the ML estimates are very close at different missingness proportions for BD, whereas the PL estimates vary quite noticeably. For example, the log OR ML estimates for Numrel are 0.641, 0.641, and 0.641, but the PL estimates are 0.652, 0.809, and 1.144.

4. Simulation Study

In this section, we report on extensive simulation studies used to compare the unbiasedness and statistical efficiency of the proposed ML estimator and the PL estimator. We also evaluate the robustness of our method with respect to a violation of the rare disease assumption and a misspecification of the model for Z. First, we generated a large random sample of data for the variable A, conventional predictors $\boldsymbol{X} = \{(X_1, X_2, X_3, X_4, X_5)^T\}$, and new predictor Z. A is generated from a uniform distribution in the range (30, 80). The match-

ing variable A^c is obtained as a categorized version of A, taking values one to five corresponding to the intervals [30, 40), [40, 50), [50, 60), [60, 70), and [70, 80), respectively. The variables $\boldsymbol{X}$ were generated from similar distributions of the corresponding variables observed in the BCDDP controls, as follows: X_1 and X_2 are 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 takes the value zero, one, or two, corresponding to the values 0, 1, or ≥ 2 generated from the Poisson(0.5) distribution; X_4 takes a value of zero or one, corresponding to the values 0 or ≥ 1 generated from the Poisson(1) distribution; X_5 is generated by categorizing data from the truncated normal distribution $\mathrm{TN}(140, 25^2; 80, 300)$, taking the value zero, one, two, or three, corresponding to the intervals (80, 125], (125, 150], (150,175], and (175, 300], respectively; Z is generated from the beta distribution $f_{\boldsymbol{\gamma}}(Z = z \mid \boldsymbol{X}, A)$, as specified in the data analysis above, with $\boldsymbol{\gamma}^T_{\text{mean}}$ and $\boldsymbol{\gamma}^T_{\text{precision}}$ set as $(1.5, 0.15, -0.1, 0.2, 0.05, -0.4, -0.03)$ and $(1.2, -0.08, -0.05, 0.1, 0.05, -0.1, 0.01)$, respectively; and Z^c takes the value zero, one, two, or three, corresponding to the 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 logistic regression model (3.1), with $a = 1, 2, \ldots, 5$, $\boldsymbol{\beta}^T_1 = (\beta_{1\_X_1}, \beta_{1\_X_2}$, $\beta_{1\_X_3}$, $\beta_{1\_X_4}, \beta_{1\_X_5}) = (0.15, 0.1, 0.2, 0.65, 0.2)$, and $\beta_2 = \beta_{2\_Z^c} = 0.4$. The stratum-specific intercept parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_5)^T$ were chosen to have $\Pr(Y = 1) = 0.035$ and $\Pr(Y = 1 \mid A^c = a) = 1\%, 2\%, 3\%, 5\%, 6.5\%$, for $a = 1, \ldots, 5$, respectively. We then used variable probability sampling to select 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% to 90%. We applied the ML and PL methods in different scenarios. However, the PL method cannot estimate the parameters $\boldsymbol{\gamma}$ in the model for Z. Chen, Chatterjee and Carroll (2007) directly fit $\Pr(Z|\boldsymbol{X}, A; \boldsymbol{\gamma})$ in the controls to roughly resolve this problem, because the breast cancer is rare. We view their analysis as an existing method of estimating $\boldsymbol{\gamma}$, and report the results for comparison. The simulation was repeated 1,000 times.

The 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 are close to the true values, and the averaged standard error estimates are close to the simulation standard error. The coverage is near the nominal value 95% for all estimates. In Figure 1, where the missingness probability was 50%, the averaged estimates appear to be close to their true

values across all four sample sizes considered, and the variances of the ML estimates are all smaller than those of the PL method. At a sample size of 1,000, the relative efficiency ("RE") of the ML method is 1.22∼1.87 for estimating $\boldsymbol{\beta}$, and as high as above two for estimating $\boldsymbol{\gamma}$. For some parameters, such as the third component of $\boldsymbol{\gamma}_{\text{precision}}$, the ML estimate can be even more efficient than the PL estimate obtained from double the sample size. The RE for estimating $\boldsymbol{\beta}$ is larger than one, and decreases slowly with increasing sample sizes. The RE for estimating $\boldsymbol{\gamma}$ is about two across all sample sizes, suggesting that the efficiency gain is mainly from incorporating the cases and the information from the Phase I data. Table 3 shows the efficiency gain increase of the ML method with respect to an increased missingness probability at a sample size of $n = 2000$. When the missingness probability increases from 10% to 90%, the RE of the ML estimates increases from 1.039 to 2.041 for β_{X_1}, and from 1.202 to 1.474 for β_Z. The RE for $\boldsymbol{\gamma}$ also increases but the increase is much smaller. These results indicate that the ML method exhibits superior performance when a large proportion of missingness occurs, but that it can also have a reasonable efficiency advantage even when the missingness probability is small. The results under the sample sizes $n = 500$ and $n = 1000$ are largely similar, and are presented in Tables S4 and S5 in Supplementary Material.

To examine the robustness of the proposed ML method with respect to the outcome prevalence, we set the prevalence as $\Pr(Y = 1) = 0.1$ with $\Pr(Y = 1 \mid A^c = a) = 3\%, 6\%, 9\%, 14\%, 18\%, a = 1, \ldots, 5$, or $\Pr(Y = 1) = 0.2$ with $\Pr(Y = 1 \mid A^c = a) = 6\%, 12\%, 18\%, 28\%, 36\%, a = 1, \ldots, 5$. The estimation for $\boldsymbol{\beta}$ is presented in Table 4, where the sample size is $n = 1000$ and the missingness probability of Z is 50%. The biases appear to be minimal, and remain nearly unchanged with increasing outcome prevalence. The largest bias occurs for $\hat{\beta}_Z$, but is only about 4% with $\Pr(Y = 1) = 0.2$ and 13% with $\Pr(Y = 1) = 0.5$ (unreported result). The ML estimates have smaller averaged MSEs than those of the PL estimates in all parameter settings, and are nearly identical under different values for $\Pr(Y = 1)$. Both methods have coverage probabilities close to the nominal level. The averaged asymptotic standard error remains close to the simulation standard error. The estimates for $\boldsymbol{\gamma}$ are presented in the Supplementary Material, 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 the robustness of the proposed ML method for estimating $\boldsymbol{\beta}$ when the model for the conditional distribution of Z is misspecified, we fitted the same model (3.2), but with constant precision parameters $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{\text{mean}}^T, \phi)^T$.

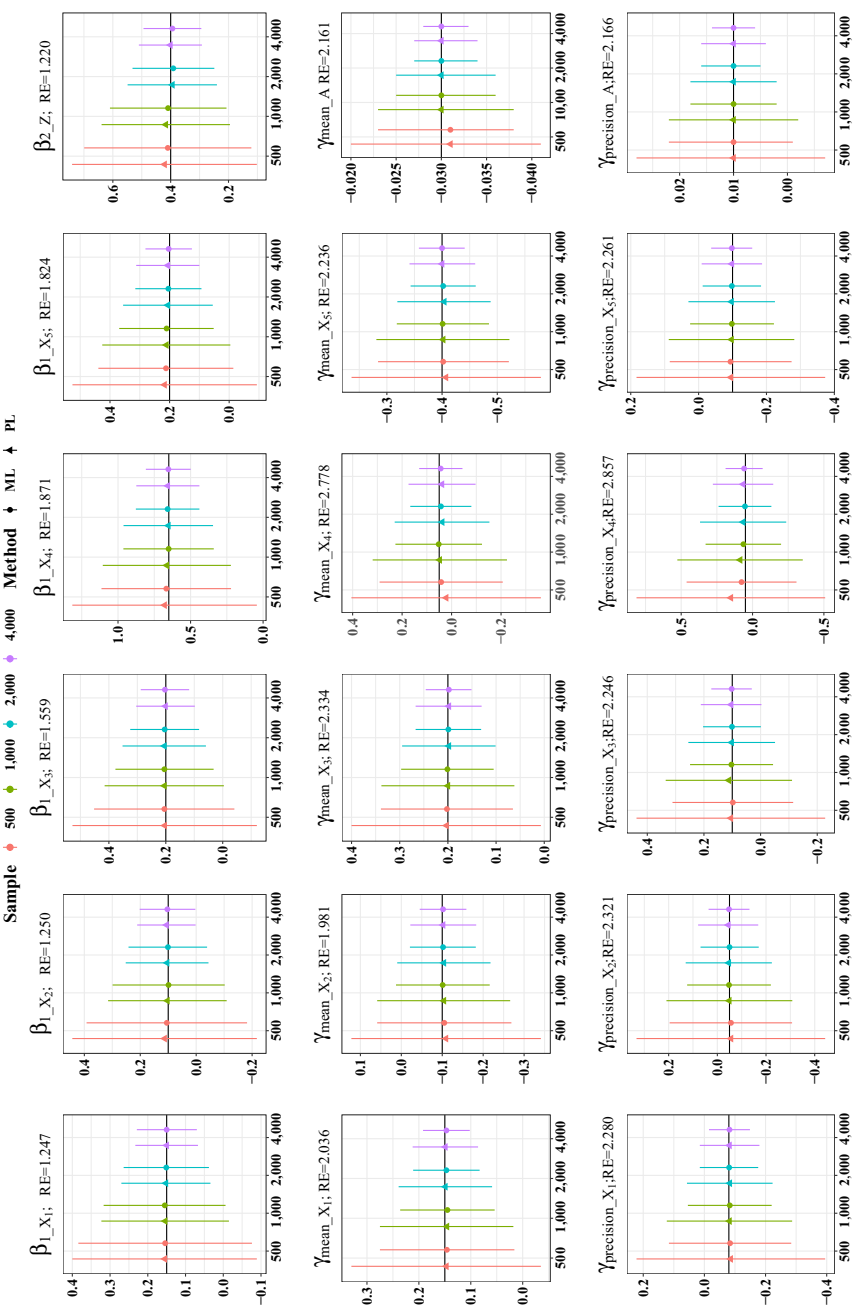

Figure 1. The mean estimates and corresponding 95% confidence intervals for the parameters in the disease-risk model ($\boldsymbol{\beta}$) and the parameters in the conditional distribution of Z ($\boldsymbol{\gamma}_{\text{mean}}, \boldsymbol{\gamma}_{\text{precision}}$). The relative efficiency ("RE") of the ML estimates compared with that of the PL estimates when the sample size is $n = 1000$ is noted above each sub-figure. The prevalence is 0.035 and the missingness probability of Z in Phase II is 50%.

Table 3. Results of 1,000 repetitions, as described in Section 4, when the prevalence is 0.035, the sample size $n = 2000$, and the missingness probability of Z is 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	$\widehat{\text{SE}}$	RE	Mean	SE	$\widehat{\text{SE}}$	RE	Mean	SE	$\widehat{\text{SE}}$	RE	Mean	SE	$\widehat{\text{SE}}$	RE	Mean	SE	$\widehat{\text{SE}}$	RE
$\beta_{1\_X_1}$	0.15	0.153	0.057	0.057	1.039	0.153	0.057	0.057	1.102	0.151	0.059	0.058	1.209	0.148	0.059	0.059	1.361	0.152	0.064	0.066	2.041
$\beta_{1\_X_2}$	0.1	0.102	0.067	0.071	1.063	0.099	0.073	0.071	1.119	0.101	0.073	0.072	1.175	0.100	0.073	0.073	1.431	0.105	0.083	0.080	2.525
$\beta_{1\_X_3}$	0.2	0.203	0.062	0.061	1.031	0.207	0.061	0.061	1.245	0.204	0.062	0.062	1.594	0.206	0.064	0.063	2.072	0.207	0.070	0.071	5.529
$\beta_{1\_X_4}$	0.65	0.648	0.111	0.111	1.129	0.656	0.110	0.111	1.465	0.658	0.109	0.112	2.048	0.651	0.114	0.113	3.336	0.660	0.124	0.123	10.065
$\beta_{1\_X_5}$	0.2	0.205	0.053	0.054	1.135	0.203	0.054	0.055	1.410	0.204	0.056	0.057	1.785	0.207	0.062	0.061	2.682	0.205	0.078	0.078	5.314
$\beta_{2\_Z^c}$	0.4	0.393	0.052	0.054	1.202	0.392	0.060	0.061	1.213	0.391	0.070	0.072	1.243	0.396	0.098	0.093	1.198	0.399	0.155	0.159	1.474
$\gamma_{\text{mean\_0}}$	1.5	1.523	0.101	0.102	2.010	1.525	0.117	0.115	1.998	1.519	0.134	0.137	1.988	1.526	0.180	0.177	1.931	1.541	0.316	0.307	2.207
$\gamma_{\text{mean\_}X_1}$	0.15	0.146	0.024	0.024	2.063	0.146	0.028	0.028	2.079	0.147	0.034	0.033	2.016	0.146	0.044	0.042	1.968	0.149	0.073	0.074	2.413
$\gamma_{\text{mean\_}X_2}$	-0.1	-0.103	0.030	0.031	2.008	-0.100	0.034	0.035	2.127	-0.102	0.042	0.041	2.011	-0.100	0.054	0.053	2.135	-0.106	0.099	0.093	2.201
$\gamma_{\text{mean\_}X_3}$	0.2	0.200	0.026	0.026	1.943	0.199	0.030	0.029	2.126	0.199	0.034	0.035	2.196	0.199	0.047	0.045	2.176	0.199	0.081	0.078	2.488
$\gamma_{\text{mean\_}X_4}$	0.05	0.044	0.049	0.047	2.195	0.047	0.055	0.053	2.499	0.043	0.064	0.063	2.600	0.046	0.086	0.081	2.538	0.045	0.148	0.142	2.826
$\gamma_{\text{mean\_}X_5}$	-0.4	-0.400	0.023	0.022	2.139	-0.400	0.026	0.025	2.023	-0.402	0.029	0.030	2.060	-0.400	0.039	0.039	2.156	-0.398	0.069	0.068	2.498
$\gamma_{\text{mean\_}A}$	-0.03	-0.030	0.001	0.001	2.232	-0.030	0.002	0.002	1.992	-0.030	0.002	0.002	2.041	-0.030	0.003	0.002	2.061	-0.031	0.005	0.004	2.194
$\gamma_{\text{precision\_0}}$	1.2	1.202	0.148	0.149	2.014	1.196	0.164	0.169	2.175	1.197	0.197	0.200	2.151	1.212	0.269	0.261	2.120	1.240	0.485	0.471	2.451
$\gamma_{\text{precision\_}X_1}$	-0.08	-0.080	0.036	0.036	2.180	-0.081	0.039	0.041	2.284	-0.081	0.050	0.049	2.039	-0.082	0.065	0.064	2.101	-0.086	0.121	0.116	2.654
$\gamma_{\text{precision\_}X_2}$	-0.05	-0.052	0.046	0.045	2.189	-0.055	0.051	0.051	2.333	-0.050	0.062	0.061	2.161	-0.048	0.080	0.080	2.329	-0.045	0.155	0.146	2.624
$\gamma_{\text{precision\_}X_3}$	0.1	0.101	0.039	0.039	2.217	0.102	0.043	0.044	2.197	0.101	0.050	0.052	2.541	0.102	0.068	0.068	2.318	0.101	0.128	0.123	2.759
$\gamma_{\text{precision\_}X_4}$	0.05	0.057	0.071	0.070	2.689	0.056	0.081	0.080	2.815	0.053	0.094	0.094	2.610	0.063	0.126	0.123	3.106	0.078	0.232	0.222	3.468
$\gamma_{\text{precision\_}X_5}$	-0.1	-0.096	0.033	0.033	2.170	-0.096	0.038	0.037	1.998	-0.098	0.044	0.044	2.366	-0.097	0.057	0.057	2.213	-0.103	0.107	0.104	2.729
$\gamma_{\text{precision\_}A}$	0.01	0.010	0.002	0.002	2.154	0.010	0.002	0.002	2.218	0.010	0.003	0.003	2.262	0.010	0.004	0.004	2.095	0.011	0.007	0.007	2.484

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 $\boldsymbol{\beta}$, and the relative efficiency of the ML method compared with the control-only analysis for estimating $\boldsymbol{\gamma}$.

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

True		Mean		MSE×100		SE×100 ($\widehat{\text{SE}} \times 100$)		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, mean estimates; SE, simulation standard error; $\widehat{\text{SE}}$, average asymptotic standard error estimates; CP, 95% coverage probabilities; MSE is computed by averaging $(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^2$ over 1,000 repetitions.

The Pearson's chi-squared goodness of fit test indicated that this model did not fit well in the controls (p-value = 0.03). The results are shown in Table 4, where the sample size is $n = 1000$, the missingness probability of Z is 50%, and the prevalence is 0.035. The ML estimates remain close to the true values and have smaller MSEs. The largest bias occurred for β_Z, and is only about 5%. The coverage probabilities are close to 95%, and the averaged asymptotic standard error remains close to the simulation standard error. An additional simulation that investigates the finite-sample performance under a different model for Z is presented in the Supplementary Material.

5. Discussion

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

case-control data, the gene-environment independence constraint can be exploited to increase statistical efficiency (Chatterjee and Carroll (2005)). Interestingly, this constraint was found to result in the identifiability of the intercept parameter. The information used to estimate the intercept parameter turns out to be very sparse, as evidenced by a much smaller efficiency gain when estimating the gene-environment interaction OR parameter (Chen and Chen (2011)). In this work, we considered the outcome variable to be rare to eliminate the need to estimate 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, are biased when the outcome prevalence is 50%. In practice, when there is concern about the rare outcome assumption, we suggest using the external information of $\Pr(Y = 1|A = a)$ to resolve the problem of estimating the intercept parameters.

When analyzing standard case-control data, a prospective logistic regression analysis that ignores retrospective sampling is equivalent to a retrospective likelihood analysis (Prentice and Pyke (1979)), even in the presence of missing data (Roeder, Carroll and Lindsay (1996)). However, when constraints are enforced on the covariate distribution, the prospective logistic regression analysis that ignores the information on the covariate distribution is valid, but no longer semiparametric efficient. Under a gene-environment independence assumption, Chatterjee and Carroll (2005) developed an efficient retrospective ML approach by maximizing a "modified" prospective likelihood, which differs from the likelihood function for the prospective logistic regression analysis (Chatterjee et al. (2006)). This approach of solving "modified" prospective score functions to obtain ML estimates for OR parameters has been extended to exploit different covariate distribution constraints (Chen, Chatterjee and Carroll (2007); Spinka, Carroll and Chatterjee (2005); Mukherjee and Chatterjee (2008)). In this work, we have established a key theoretical result that 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 a real-data analysis, we compared our proposed ML method with the PL method (Chen et al. (2008)). Although a finer construction of strata $K(\boldsymbol{X})$ in (2.7) may improve the efficiency of the PL method, its application is limited by a practical trade-off between the number of strata and the number of cases and controls in each stratum. Additional results showed that ML estimates have a

smaller variance than PL estimates do, as long as the information of $\boldsymbol{X}$ is not fully captured in $K(\boldsymbol{X})$. We have described our ML method by 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 $\Pr(\boldsymbol{Z}|\boldsymbol{X}, A)$. Additional simulation results for the bivariate Phase II variable (Z_1, Z_2) are presented in the Supplementary Material. We have demonstrated the robustness of our method using simulation studies under a modest misspecification of the conditional distribution for the Phase II variables. However, the bias becomes larger if the misspecification becomes severe. Therefore, caution needs to be exercised when deciding on suitable distributions. Here, exploring the model forms and checking the goodness of fit using control data, as in our real-data analysis, can be highly useful. When Phase II variables are of low dimension, it is potentially feasible to adopt a semiparametric or nonparametric model for the joint distribution. In the absence of Phase I covariates $\boldsymbol{X}$, Song, Zhou and Kosorok (2009) developed a nonparametric ML method in which $P(\boldsymbol{Z})$ is estimated nonparametrically. With multivariate $\boldsymbol{Z}$ and $\boldsymbol{X}$, it is generally challenging to deal with $\Pr(\boldsymbol{Z}|\boldsymbol{X})$ likelihood-based methods. In a recent ML method for analyzing prospective two-phase data (Tao, Zeng and Lin (2017)), a sieve approximation was adopted for $\Pr(\boldsymbol{Z}|\boldsymbol{X})$, which appears to work well when $\boldsymbol{X}$ is of modest dimension. We will explore extensions of our method along this line in future work. We have focused on increasing the efficiency of the parameter estimation in this work. The estimated OR 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, 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.

Supplementary Material

The online Supplementary material includes appendices and the tables referenced in Sections 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 the National Natural Science Foundation of China, No. 11771241 and 11931001.

References

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

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

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

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

Byrne, C., Schairer, C., Wolfe, J., Parekh, N., Salane, M., Brinton, L. A. et al. (1995). Mammographic features and breast cancer risk: Effects with time, age, and menopause status. *Journal of the National Cancer Institute* **87**, 1622–1629.

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. (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 (Statistical Methodology)* **69**, 123–142.

Chatterjee, N., Spinka, C., Chen, J. and Carroll, R. J. (2006). Comment. *Journal of the American Statistical Association* **101**, 108–111.

Chen, H. Y. and Chen, J. (2011). On information coded in gene-environment independence in case-control studies. *American Journal of Epidemiology* **174**, 736–743.

Chen, J., Ayyagari, R., Chatterjee, N., Pee, D., Schairer, C., Byrne, C. et al. (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**, 976–988.

Chen, J., Pee, D., Ayyagari, R., Graubard, B., Schairer, C., Byrne, C. et al. (2006). Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. *Journal of the National Cancer Institute* **98**, 1215–1226.

Chen, Y.-H., Chatterjee, N. and Carroll, R. J. (2007). Retrospective analysis of haplotype-based case-control studies under a flexible model for gene-environment association. *Biostatistics* **9**, 81–99.

Fears, T. R. and Brown, C. C. (1986). Logistic regression methods for retrospective case-control studies using complex sampling procedures. *Biometrics* **42**, 955–960.

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C. et al. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute* **81**, 1879–1886.

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, Series B (Statistical Methodology)* **61**, 413–438.

Little, R. J. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Mukherjee, B. and Chatterjee, N. (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**, 685–694.

Neyman, J. (1938). Contribution to the theory of sampling from human populations. *Journal of the American Statistical Association* **33**, 101–116.

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

Roeder, K., Carroll, R. J. and Lindsay, B. G. (1996). A semiparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association* **91**, 722–732.

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

Schill, W., Jockel, K.-H., Drescher, K. and Timm, J. (1993). Logistic analysis in case-control studies under validation sampling. *Biometrika* **80**, 339–352.

Scott, A. J. and Wild, C. (2001). Maximum likelihood for generalized case-control studies. *Journal of Statistical Planning and Inference* **96**, 3–27.

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.

Song, R., Zhou, H. and Kosorok, M. R. (2009). A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome. *Biometrika* **96**, 221–228.

Spinka, C., Carroll, R. J. and Chatterjee, N. (2005). Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genetic Epidemiology* **29**, 108–127.

Tao, R., Zeng, D. and Lin, D.-Y. (2017). Efficient semiparametric inference under two-phase sampling, with applications to genetic association studies. *Journal of the American Statistical Association* **112**, 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**, 119–128.

Yaqi Cao

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

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

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

Lu Chen

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

E-mail: daisychen1009@gmail.com

Ying Yang

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

E-mail: yangying@tsinghua.edu.cn

Jinbo Chen
Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA.
E-mail: jinboche@pennmedicine.upenn.edu

(Received June 2021; accepted January 2022)