STATISTICAL METHODS FOR OUTCOME-DEPENDENT SAMPLING DESIGNS

Le Wang

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2018

Supervisor of Dissertation

_____

Jinbo Chen, Professor of Biostatistics

Graduate Group Chairperson

_____

Nandita Mitra, Professor of Biostatistics

Dissertation Committee

Pamela Shaw, Associate Professor of Biostatistics

Qi Long, Professor of Biostatistics

Mary Putt, Professor of Biostatistics

Scott Damrauer, Assistant Professor of Surgery

STATISTICAL METHODS FOR OUTCOME-DEPENDENT SAMPLING DESIGNS

ACKNOWLEDGEMENT

I have been working at Penn for almost five years, and I am grateful for all the people who have advised, inspired, encouraged, supported, and helped me along the way. First of all, I would like to express my sincere gratitude to my advisor, Prof. Jinbo Chen, for advising me with great patience, supporting me generously, and encouraging me every time I face difficulties with her warm heart. Her enthusiasm and vision for research as a scientist and her diligent work ethics as a researcher have inspired me to further pursue my career in academia.

Second, I would like to thank my dissertation committee: Prof. Pamela Shaw, Prof. Scott Damrauer, Prof. Qi Long, and Prof. Mary Putt for being the most supportive committee I could have ever asked for, providing insightful and valuable comments and discussions on my research, and offering great advice on my career development. Next, I want to thank Prof. Yong Chen for participating in the first project of my dissertation work, generously sharing his own learning and working experience in academia with me, and being supportive all along. Moreover, I want to thank my former RA advisor, Prof. Benjamin French, who was the first person welcoming me to the department. Ben advised me to work on different collaborative projects with doctors and clinicians at Penn, and fully supported my job application in the academic field. I would also like to thank my consulting II project advisor, Prof. Rebecca Hubbard, for discussing research with me patiently, helping me improve coding techniques, and sharing her own experience and providing career advice selflessly. She is another female role model of mine. Besides, I would like to thank my former academic advisor Prof. Mingyao Li, professors in the Biostatistics Department especially those whom I took classes with, supportive administrative staff Cathy and Marissa, and my fellow students at Penn.

Last but not least, I would like to thank my family and friends. Words cannot express my gratitude to my parents enough. Thank you for always having faith in me, being my first and best mentors, supporting and loving me unconditionally. You are my greatest inspiration in life and I couldn't have done any of this without you two. I want to thank my husband, Dong, for being my partner in every aspect of life, from work and research to movie and music to traveling and cooking. Thank you for alway being there for me and for taking such good care of our home to let me stay focused in the past five years. I look forward to the new adventure ahead for us together!

ABSTRACT

STATISTICAL METHODS FOR OUTCOME-DEPENDENT SAMPLING DESIGNS

Le Wang

Jinbo Chen

My dissertation work focuses on the development of novel outcome-dependent sampling designs and statistical methods of analysis. In a biomedical cohort study for assessing association between a binary outcome variable and a set of covariates, it is common that some covariates can only be measured on a subgroup of study subjects. An important design question is which subjects to select into the subgroup towards increased statistical efficiency. Existing designs can achieve improved efficiency for estimating odds ratio parameters for the completely observed covariates. Our goal is to improve efficiency for the incomplete covariates, which is of great importance in studies where the covariates of interest cannot be fully collected. In the first two projects, we proposed a novel sampling design in a common scenario where an external model is available relating the outcome and complete covariates. Our design oversampled cases and controls whose probabilities of having their own outcome were low as predicted by the external model and at the same time matched cases and controls on complete covariates. We developed a pseudo-likelihood method for estimating odds ratio parameters. Through simulation studies and a real cohort study, we showed that our design led to reduced asymptotic variances of the odds ratio parameter estimates for both incomplete and complete covariates. In the third project, we developed a family-supplemented inverse-probability-weighted empirical likelihood approach to correcting for a type of outcome-dependent selection bias in case-control genetic association studies, where genotype data were incomplete for reasons that were related to the genotype itself. Genetic association analysis would be biased if such non-ignorable missingness were naively ignored. Our method exploited genetic data from family members to help infer missing genotype data. It jointly estimated odds ratio parameters for genetic association and missingness, where a logistic regression model was used to relate missingness with genotype and other covariates. In the estimating equation for genetic association parameters, we weighted the empirical likelihood score function based on subjects who had genotype data by the inversed probabilities that their genotype data were available. We studied large and finite sample performance of our method and applied it to a case-control study of breast cancer.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

My dissertation work focuses on the development of novel outcome-dependent sampling designs and statistical methods of analysis. Outcome-dependent sampling strategies have been widely applied in biomedical studies. The most well-known sampling scheme is the case-control design for studying a rare binary outcome, that improves statistical efficiency by increasing the proportion of cases in the sample compared with that in the full population. To elucidate whether an observed association between a binary outcome and an exposure variable is confounded, it is cost-effective to collect data for covariates only on a subset of cases and controls. Then for estimating the odds ratio (OR) parameter with respect to the exposure, an equal number of cases and controls matched on the exposure is more informative than the same number of unmatched cases and controls. This matched sampling, referred to as the "balanced design" (Breslow and Cain, 1988), is generally more efficient particularly for a rare exposure, heuristically because the exposed subjects are oversampled. The key to the improved efficiency of an oversampling scheme is to identify informative subjects. In these well-known designs, subjects in rare groups are generally considered more informative. Appropriate statistical methods are required to account for the oversampling scheme in order to obtain unbiased association parameter estimates.

It is common that some covariates can only be collected for a subgroup of subjects in an association study for a binary outcome. The resultant incomplete data structure is usually described using a two phase sampling scheme (Neyman, 1938; White, 1982), where the outcome and the completely observed covariates are collected on all subjects at phase I and the remaining covariates are collected on a subgroup at phase II. The oversampling schemes are usually implemented in two-phase designs for the purpose of greater efficiency. Existing oversampling designs mainly focus on improving efficiency for estimating association parameters with respect to phase I variables. Phase II variables, however, are of great importance in studies where covariates of interest cannot be fully collected. Therefore, in the first two projects of this dissertation work, we aim to propose novel outcome-dependent sampling designs to improve efficiency for phase II covariates.

Efficient sampling designs have been implemented in "big data" literature in order to increase com-

putational efficiency. Although outcome and covariates are available for all subjects, data reduction is desirable to reduce computation burden. An innovative sampling scheme, the "local case-control design" (LCC) Fithian and Hastie (2014) , was proposed recently for studying binary outcomes in this context. It oversamples subjects who have lower predicted probability of having their true case or control status based on a preliminary model and achieved greater efficiency compared with case-control sampling. We find that the LCC design sheds new light into two-phase sampling.

In the first two projects, we mainly focus on the efficiency aspect of outcome-dependent sampling designs. Another important issue under the outcome-dependent sampling framework is to adjust for participation bias, an outcome-dependent selection bias. In case-control genetic association studies, it is common that many individuals' genotype data are missing for reasons related to the disease under study. This type of selection bias, referred to as the "participation bias", results in a non-ignorable missing structure and leads to biased inference if the genetic variants of interest are related to the disease and responsible for missingness. Participation bias widely exists in biomedical studies (Anderson et al., 2011; Chard, 1991; Falcone et al., 2013; Horsfall, Nazareth, and Petersen, 2012; Liew et al., 2015), but not many options regarding the sampling designs or statistical methods to adjust for the bias are available (Aschengrau and Seage, 2013; Haneuse and Chen, 2011; Haneuse et al., 2016). Chen, Weinberg, and Chen (2016) developed a family-supplemented design (FSD) that adjusts for participation bias using a maximum likelihood approach where they substitute first-degree family members' genotype data for deceased individuals. They proposed a valid estimator for the association parameter and showed greatly increased statistical power. In this study we aim to extend FSD to more general case-control settings where one is allowed to incorporate covariates and interaction effects into the association model.

In Chapter 2, we extend LCC to the two-phase settings in a common scenario where an external model is available to relate the outcome variable and phase I covariates. We propose the Goodness-of-fit based sampling scheme ("GOF") that oversamples cases and controls with worse goodness-of-fit based on the external model. We develop a pseudo-likelihood method for estimating OR parameters and find that GOF has a unique advantage of increasing efficiency for estimating OR parameters for the incomplete phase II covariates and consistently outperforms case-control and balanced sampling in both simulated and real data settings in this aspect. GOF provides a new perspective to define informative subjects in oversampling designs. Those who lack goodness-of-fit

based on the external model only including phase I covariates indicate the necessity to incorporate phase II covariates into the model for a better fit and thus are considered more informative with respect to the phase II covariates.

The balanced design, however, gains great efficiency for estimating the matching variables at phase I. Comparing the efficiency of GOF and the balanced sampling motivates us to propose a new hybrid two-phase sampling scheme in Chapter 3, the balanced goodness-of-fit based sampling design, which performs GOF sampling first and then further matches the GOF subsample on complete covariates similarly to the balanced design. We propose a pseudo-likelihood method for estimating OR parameters and develop its asymptotic properties. Through simulation studies and explorations in a real cohort study, we find that BGOF generally leads to reduced asymptotic variances of the OR estimates and the reduction for the matching covariates is comparable to that of the balanced design.

In Chapter 4, we develop an estimating equation approach, the family-supplemented weighted empirical likelihood method, to correcting for participation bias under non-ignorable missing structure, a type of outcome-dependent selection bias, in case-control genetic association studies. The novelty of the proposed method is to use first-degree family's genetic information as a proxy for an individual's missing genotype. We apply a logistic regression model to relate missingness with genotype and covariates, and use the expectation of the corresponding logistic regression score function conditional on all the observed data, including family's genotype data, as the estimating equation for the missingness odds ratio parameters. We develop an empirical likelihood for the genetic association parameters and weight the empirical likelihood among individuals with complete data by the inverse of their probabilities of genotype data availability as the estimating equation for the association parameters. We estimate the nuisance parameters, i.e., the covariate distribution conditional on genotype, nonparametrically using data of controls inversely-weighted by their probabilities that complete data have been collected. Finally, we obtain the estimators for association and missingness by jointly solving these estimating equations. We develop the asymptotic properties of this method and evaluate the finite and large sample properties in simulation studies and a family-based case-control genetic association study of young-onset breast cancer.

CHAPTER 2

A NOVEL GOODNESS-OF-FIT BASED SAMPLING DESIGN FOR STUDYING BINARY OUTCOMES

2.1. Introduction

In biomedical studies, it is common to oversample informative subjects to increase statistical efficiency. The best known example is the case-control design for studying rare outcomes, where cases are oversampled such that the proportion of cases is much larger than that in the population. To elucidate whether an observed association between a binary outcome and an exposure variable is confounded, it is cost-effective to collect data for covariates only on a subset of cases and controls. Then for estimating the odds ratio (OR) parameter with respect to the exposure, an equal number of cases and controls matched on the exposure is more informative than the same number of unmatched cases and controls. This matched sampling, referred to as the "balanced design" (Breslow and Cain, 1988), is generally more efficient particularly for a rare exposure, heuristically because the exposed subjects are oversampled. The oversampling can be accounted for by applying suitable statistical methods, and the resultant inference is comparable to that if the sampling were unbiased. The key to the improved efficiency of an oversampling scheme is to identify informative subjects.

In general, for studying the association between a binary outcome and a set of covariates, it is common that data for some covariates can only be made available for a subset of subjects. The resultant incomplete data structure is usually described using a two phase sampling scheme (Neyman, 1938; White, 1982), where the outcome and the completely observed covariates are collected on all subjects at phase I and the remaining covariates are collected on a subset at phase II. When it is at investigators' disposal to decide whose covariates to measure at phase II, the strategy for subset selection affects the efficiency for estimating association parameters. Stratification on phase I variables has been commonly adopted in this regard, where strata and sampling proportion within each stratum have to be determined *a priori*. Besides the "balanced" design, the "optimal" sampling, where the sampling proportion within each phase I stratum is chosen to minimize the asymptotic variances of the estimated OR parameters, has been derived in various scenarios (Holcroft and

Spiegelman, 1999; McIsaac and Cook, 2014; McNamee, 2005; Reilly, 1996; Wild et al., 2008). To implement these designs, phase I variables used for stratification have to be discretized, and this data coarsening for continuous variables incurs information loss. To ensure a sufficient number of subjects within each phase I stratum for phase II sampling, the number of phase I sampling strata has to be carefully chosen. These decisions become very challenging with multiple phase I covariates. The efficiency of a sampling strategy necessarily depends on the statistical method for analysis. Consequently, the optimal design derived under one analysis method is usually not optimal under alternative methods. Implementation of the optimal design usually requires true parameter values, and the design efficiency is compromised when assumed values deviate from the truth.

Efficient sampling has been frequently implemented in the "big data" literature, mainly to increase computational efficiency. For binary outcomes, data points in one category (cases) may be far fewer than those in the other category (controls), resulting in high "imbalance". A large number of control data points are redundant in terms of statistical efficiency (Breslow and Day, 1980), and they incur great computation burden. Data reduction is then desirable even if both the outcome and covariates are fully observed for all subjects. Case-control subsampling is an obvious choice for this purpose. Recently, an innovative sampling strategy was proposed, which oversamples subjects who have lower predicted probability of having their true case or control status (Fithian and Hastie, 2014). Because each data point has its own probability of being sampled, this sampling scheme was termed "local" case-control sampling (LCC). By applying a prospective logistic regression model with an appropriate offset term that accounts for subsampling (Breslow and Cain, 1988), this strategy has greater statistical efficiency and yields the same inference as if the full data were used even under model mis-specification.

In this chapter, we investigated the efficiency of LCC when extended to the setting of two-phase sampling. We call this extension goodness-of-fit based sampling (GOF) for reasons that we will explain in section 2.2. We found that GOF has a unique advantage of increasing efficiency for estimating OR parameters for the incomplete phase II covariates and consistently outperforms case-control and balanced sampling in both simulated and real data settings in this aspect. This advantage is highly valued in studies where covariates of interest cannot be fully collected for economic reasons.

The rest of this chapter is organized as follows. In Section 2.2, we describe the sampling and estimation procedures of GOF, and compare its efficiency with case-control (CC) and balanced two-phase sampling designs (BD) for estimating OR parameters. We provide insight into the improved efficiency of GOF using a simulated data example in Section 2.4. In Section 2.5, we further evaluate GOF using data from an ongoing biomarker study of gestational diabetes. We make final remarks in Section 2.6.

2.2. Goodness-of-Fit Based Design for Two-Phase Sampling

In this section, we describe sampling and inference procedures of GOF. Let Y denote the binary outcome variable with $Y = 1$ indicating cases and $Y = 0$ controls. Let $\boldsymbol{X}$ denote phase I covariates that are available for all subjects, and $\boldsymbol{Z}$ denote phase II covariates that can only be measured on a subset of subjects. A logistic regression model is used to describe the relationship between Y and covariates $\boldsymbol{X}$ and $\boldsymbol{Z}$,

$$\text{logit } P(Y = 1|\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{x}\boldsymbol{\beta}_1 + \boldsymbol{z}\boldsymbol{\beta}_2, \tag{2.1}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are the OR parameters of interest. Note here to simplify notation, a variable with value equal to one is implicitly included in $\boldsymbol{X}$, with the corresponding regression coefficient in $\boldsymbol{\beta}_1$ being the intercept parameter. Let $\boldsymbol{\beta}$ denote the vector of all parameters $(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$. The outcome and phase I variables $(Y, \boldsymbol{X})$ are collected from a cross-sectional sample of N subjects. We wish to select a subset of $m\,(m < N)$ subjects for the measurement of $\boldsymbol{Z}$. The main difference among the existing and our proposed sampling designs is the probability of selecting subjects into phase II, which is a function of completely collected data $(Y, \boldsymbol{X})$. In a balanced design, discrete sampling strata are defined based on Y and $\boldsymbol{X}$.

2.2.1. A Brief Review of the LCC Sampling Design

In the LCC design (Fithian and Hastie, 2014), the regression variables Y, $\boldsymbol{X}$, and $\boldsymbol{Z}$ are fully observed on all N subjects. A preliminary model that is in the same form as equation (2.1) was assumed to exist, which can be derived from external sources or from a small subset of the data put aside specifically for deriving this model. Denote this model as $P^{le}(Y = 1|\boldsymbol{x}, \boldsymbol{z})$, where superscripts "l" and "e" represent "local" and "external" to the study dataset, respectively. This design selects

subjects with probability

$$|y - P^{le}(Y = 1|\boldsymbol{x}, \boldsymbol{z})|.$$

That is, a Bernoulli experiment is conducted for each of the N subjects, with the success probability for subject i equal to $|y_i - P^{le}(Y = 1|\boldsymbol{x}_i, \boldsymbol{z}_i)|$, $i = 1, 2, \ldots N$. A subject is selected if the experiment yields a success. Therefore, cases who have a smaller predicted probability of being a case and controls who have a smaller predicted probability of being a control according to model P^{le} have greater probabilities of being selected. To achieve a bigger or smaller desirable sample size, the sampling probabilities can be multiplied by a suitable constant number.

2.2.2. Extension of LCC to the Setting of Two-phase Sampling

While in LCC, the outcome variable Y and all covariates are available for all subjects, the two-phase sampling design measures $\boldsymbol{Z}$ on a subset of subjects selected based only on $(Y, \boldsymbol{X})$. In our extension, we assume that an external model exists for relating Y to phase I covariates $\boldsymbol{X}$, i.e.,

$$\text{logit } P^e(Y = 1|\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\eta}_1,$$

where the parameters $\boldsymbol{\eta}_1$ are known. We note that such preliminary models often exist. For example, the relationship between the risk of breast cancer and reproductive risk factors has been well established. Either the complete set or a subset of $\boldsymbol{X}$ can be involved depending on the value of $\boldsymbol{\eta}_1$. We similarly define a quantity $S(y, \boldsymbol{x})$ that measures the difference between Y and $P^e(Y = 1|\boldsymbol{X})$, i.e.,

$$S(y, \boldsymbol{x}) = |y - P^e(Y = 1|\boldsymbol{x})|.$$

Let R denote whether a phase I subject is selected into phase II, with $R = 1$ indicating selection and $R = 0$ non-selection, and $V = \{i : R_i = 1, i = 1, \ldots N\}$ denote the selected subset. We aim to select m subjects, that is $\sum_{i=1}^{N} R_i = m$. When N is large, sampling based on S may lead to a phase II sample size greater than m as in LCC. Therefore, we propose to select phase II subjects based on $S(y, \boldsymbol{x})$ as below. Suppose that it is desirable to achieve a pre-specified case-control ratio within m phase II subjects as commonly done in epidemiological studies. We propose the sampling probability $P(R = 1|y, \boldsymbol{x})$ to be $S(y, \boldsymbol{x})$ multiplied by a constant $c_1 (c_1 > 0)$ for cases and $c_0 (c_0 > 0)$

for controls, i.e.,

$$P(R=1|Y=1,\boldsymbol{x}) = \min\{1, c_1 S(1,x_1)\},$$
$$P(R=1|Y=0,\boldsymbol{x}) = \min\{1, c_0 S(0,x_1)\}.$$

Because function $S(y,\boldsymbol{x})$ informs goodness-of-fit of the external model $P^e(Y=1|\boldsymbol{x})$, we term our extension to LCC as the goodness-of-fit based sampling, with the notion of "goodness-of-fit" further explored in Section 2.4.

When implementing GOF in practice, it may be necessary to try multiple values for c_1 and c_0 to achieve the final sample size m. Compared with CC and BD, GOF selects phase II subjects based on both Y and multiple phase I covariates without having to form discrete sampling strata. Notably, the sampling probability $P(R=1|y,\boldsymbol{x})$ used for GOF is a deterministic function. One can flexibly manipulate values of c_1 and c_0 according to practical needs. For example, when the cohort is sufficiently large and outcome prevalence is not low, one may use the same value c $(0<c<1)$ for c_1 and c_0 as in LCC, that is,

$$P(R=1|y,\boldsymbol{x}) = \min\{1, cS(y,\boldsymbol{x})\}.$$

This is equivalent to sampling in two steps. First a subset V is selected using sampling probabilities $S(y,\boldsymbol{x})$. A proportion c is then further randomly selected into the final sample at phase II, where the value of c is selected to achieve the predetermined total sample size m. When it is desirable to include all the cases in phase II, the sampling probability for cases can be set as 1, and c_0 is selected to achieve the targeted number of controls:

$$P(R=1|Y=1,\boldsymbol{x}) = 1,$$
$$P(R=1|Y=0,\boldsymbol{x}) = \min\{1, c_0 S(0,\boldsymbol{x})\}.$$

2.2.3. Statistical Inference for the GOF Sampling Design

Because GOF selects phase II subjects based on both Y and $\boldsymbol{X}$, fitting model (2.1) to phase II data naively ignoring the sampling scheme can lead to biased estimates of parameters $\boldsymbol{\beta}$. Note that the

probability of observing a case given covariates $(\boldsymbol{X}, \boldsymbol{Z})$ in phase II is

$$P(Y=1|\boldsymbol{x},\boldsymbol{z};R=1)=\frac{P(Y=1|\boldsymbol{x},\boldsymbol{z})P(R=1|Y=1,\boldsymbol{x})}{P(R=1|\boldsymbol{x},\boldsymbol{z})}. \tag{2.2}$$

Therefore, the logit of this probability takes the same form as model (2.1) but with an offset term $o(\boldsymbol{x}) \equiv \log\min\{1, c_1 S(1,\boldsymbol{x})\} - \log\min\{1, c_0 S(0,\boldsymbol{x})\}$, i.e.,

$$\text{logit } P(Y=1|\boldsymbol{x},\boldsymbol{z};R=1)=\boldsymbol{x}\boldsymbol{\beta}_1+\boldsymbol{z}\boldsymbol{\beta}_2+o(\boldsymbol{x}). \tag{2.3}$$

In the special case where $P(R=1|y,\boldsymbol{x}) = cS(y,\boldsymbol{x})$, the offset $o(\boldsymbol{x})$ equals $-\boldsymbol{x}\boldsymbol{\eta}_1$. With $c_1 = 1$ as described above, the offset $o(\boldsymbol{x})$ equals $-\log\min(1, c_0 S(0,\boldsymbol{x}))$. Similar to Fithian and Hastie (2014), we propose to maximize a pseudo-likelihood based upon $P(Y=1|\boldsymbol{x},\boldsymbol{z};R=1)$ for estimating parameters $\boldsymbol{\beta}$ in model (2.1), the logarithm of which takes the following form

$$\ell(\boldsymbol{\beta})=\sum_{i=1}^{m}\left[y_i\{\boldsymbol{w}_i\boldsymbol{\beta}+o(\boldsymbol{x}_i)\}-\log\left\{1+e^{\{\boldsymbol{w}_i\boldsymbol{\beta}+o(\boldsymbol{x}_i)\}}\right\}\right],$$

where $\boldsymbol{w}_i = (\boldsymbol{x}_i, \boldsymbol{z}_i)$. We note that this is a pseudo-likelihood rather than a maximum likelihood approach, because $\ell(\boldsymbol{\beta})$ is not based on the two-phase data likelihood (e.g., Lawless, Kalbfleisch, and Wild, 1999). It is essentially the same as the conditional likelihood approach for two-phase data under variable probability sampling (Breslow and Cain, 1988), except that the offset term $o(\boldsymbol{x})$ is a specified function. Parameters $\boldsymbol{\beta}$ can be solved from the pseudo-likelihood score equation

$$U(\boldsymbol{\beta}) \equiv \partial\ell(\boldsymbol{\beta})/\partial\boldsymbol{\beta}=\sum_{i=1}^{m}(y_i-\mu_i^g)\boldsymbol{w}_i^T=0, \tag{2.4}$$

with $\mu_i^g = \text{expit}\{\boldsymbol{w}_i\boldsymbol{\beta}+o(\boldsymbol{x}_i)\}$. Note that $E\{U(\boldsymbol{\beta})\}=0$ because $\text{E}(Y|\boldsymbol{w}_i, R_i=1)=\mu_i^g$. Therefore, by standard Z-estimation theory, the estimated coefficients $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normally distributed, i.e.,

$$\sqrt{m}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\to \text{N}\left\{\boldsymbol{0},(\boldsymbol{W}^T V \boldsymbol{W})^{-1}\right\}$$

with $V = \text{diag}\{\mu_i^g(1-\mu_i^g)\}$ and $\boldsymbol{W}$ being the covariate matrix. Note that the form of the asymptotic variance is exactly the same as if the data were prospectively generated under the pseudo-model $P(Y=1|\boldsymbol{x},\boldsymbol{z};R=1)$. This suggests that $\hat{\boldsymbol{\beta}}$ and its asymptotic variance estimates can be directly obtained from standard statistical software by fitting logistic regression model (2.1) to phase II data

with the offset term $o(\boldsymbol{x})$. This is the same case for LCC.

2.3. Simulation Studies

We conducted simulation studies to evaluate the efficiency of GOF for estimating OR parameters in order to understand its merits relative to CC and BD. We considered different phase I cohort sizes, outcome prevalences, effect sizes of the covariate of interest, correlation among phase I variables and correlation between phase I and phase II covariates. We included three phase I covariates X_1, X_2 and X_3, that is, $\boldsymbol{X} = (X_1, X_2, X_3)$. X_1 was a uniform variable in the range of 0 to 1, X_2 was a standard normal variable, and X_3 was a binary variable with the success probability 0.3. Phase II included a single covariate Z which was also a standard normal variable. Specifically, we generated the four covariates jointly as follows. We first generated a multivariate normal random variable (T_1, T_2, T_3, T_4)

$$\begin{bmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho & \rho & 0.6 \\ \rho & 1 & \rho & 0.5 \\ \rho & \rho & 1 & 0.3 \\ 0.6 & 0.5 & 0.3 & 1 \end{bmatrix} \sigma^2 \right).$$

Then we generated X_1 by applying the cumulative distribution function for T_1, and dichotomized T_3 at its 70th percentile to create X_3. X_2 and Z were set to equal T_2 and T_4, respectively. We set σ^2 equal to 1 and ρ to 0 or 0.3 corresponding to weak versus moderate correlations among phase I covariates. We then generated the binary outcome variable Y for a cohort of 3000 individuals from the logistic regression model

$$\text{logit } P\left(Y = 1 | \boldsymbol{x}, z\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 z.$$

The intercept parameter β_0 was selected to achieve the prevalence $P(Y = 1)$ of 0.05 or 0.10. The log OR parameters $(\beta_1, \beta_2, \beta_3)$ were set to be $(0.5, 0.6, -0.7)$, and different values for β_4 were considered in the range of 0.5–0.9 corresponding to weak to strong effects of Z. We similarly generated a separate cohort, which served as the "external data" to fit a logistic regression model

for Y given only $\boldsymbol{X}$, i.e.,

$$\text{logit } P^e\left(Y=1|\boldsymbol{x}\right)=\eta_0+\eta_1 x_1+\eta_2 x_2+\eta_3 x_3.$$

The fitted model $P^e\left(Y=1|\boldsymbol{x}\right)$ was then treated as the known external model and was used in all simulation iterations. $S(y,\boldsymbol{x})$ was calculated for each of the 3000 subjects. We included all cases in phase II and sampled controls with probability $\min\{1, cS(0,\boldsymbol{x})\}$. The constant c was chosen such that 2.5 times as many controls as the cases on average across all iterations were initially selected. Then a subset of controls were further randomly selected to ensure an equal number of cases and controls in the final phase II sample. This two step selection was adopted to guarantee the pre-specified case-control ratio in every iteration. For CC sampling, each phase II sample included all the cases and an equal number of controls. For BD sampling, we selected all the cases and an equal number of controls in each of the two strata formed by X_1 that was dichotomized at its median. We chose X_1 to perform stratification because it had the highest correlation with phase II covariate Z. For each set of parameter combinations, we repeated the above steps 1,000 times. We compared the three sampling designs with respect to the empirical mean biases, empirical variances, and mean asymptotic variances of $\hat{\boldsymbol{\beta}}=(\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3,\hat{\beta}_4)$.

The results on estimation of β_4 are presented in Table 2.1. Under GOF, CC and BD sampling designs, the averaged estimates were close to the true values, the empirical and mean asymptotic standard errors were similar, and the 95% coverage probabilities were nearly identical to the nominal level. Similar results for three association parameters of phase I covariates are presented in Tables C.1, C.2, and C.3 in APPENDIX C. Figure 2.1 presents the mean asymptotic variance of $\hat{\beta}_4$ (Panels 2.1a and 2.1b), the estimated log OR of phase II covariate Z, and $\hat{\beta}_1$ (Panels 2.1c and 2.1d), the estimated log OR of the stratifying covariate in BD when the outcome prevalence was 0.05. The variance of $\hat{\beta}_4$ under GOF appeared to be smaller, with percent reduction ranging from 20–30% compared with CC and 15–27% compared with BD across all simulation scenarios. The variance of $\hat{\beta}_1$ was smaller in GOF than in BD when all phase I variables were uncorrelated (Panel 2.1c). But BD achieved smaller variance when the phase I covariates were moderately correlated (Panel 2.1d). Variances of phase I variables X_2 and X_3 were smaller under GOF than under CC or BD. The performance of these sampling designs were similar when the outcome prevalence was higher, i.e., 0.10. In general, GOF achieved smaller variance relative to CC and BD for estimat-

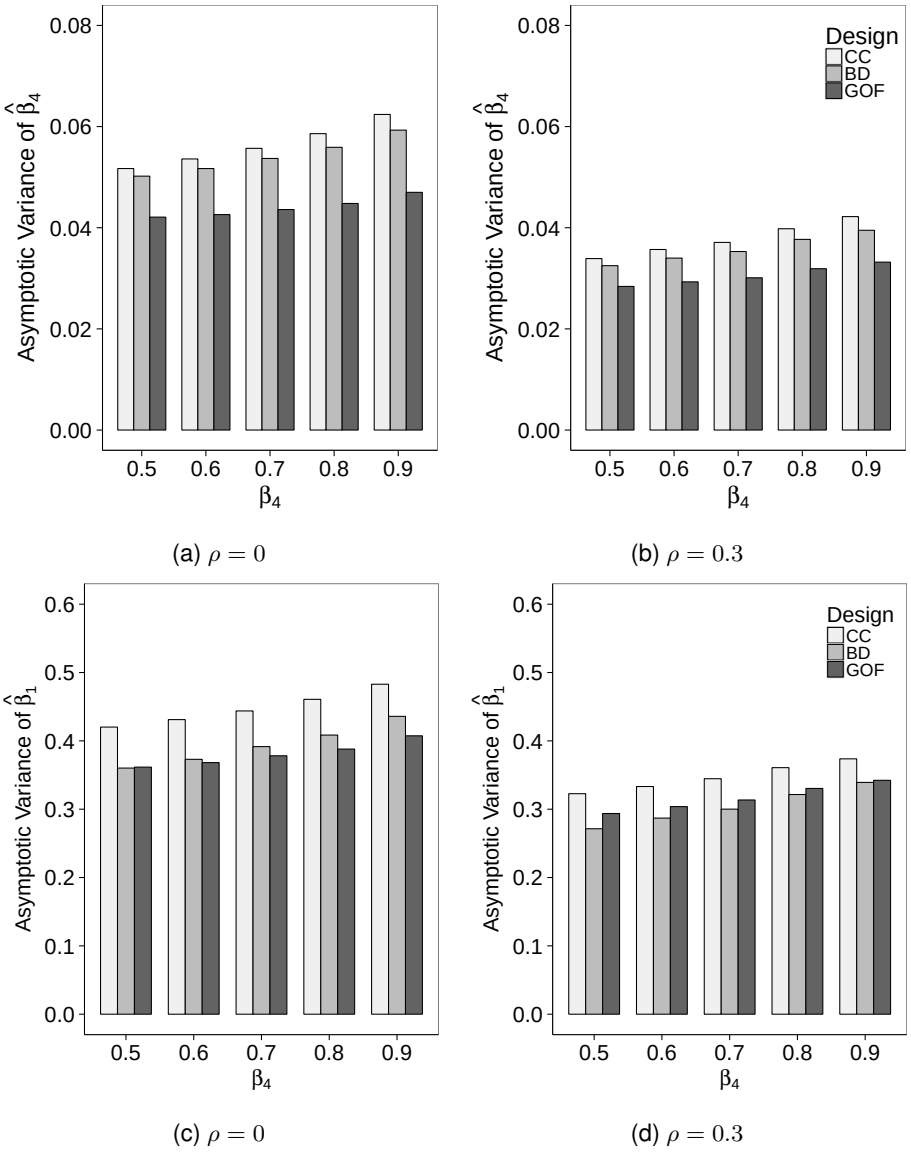
Table 2.1: The estimated log OR of phase II covariate ($\hat{\beta}_4$) under the goodness-of-fit based design (GOF), the case-control design (CC), and the balanced design (BD). The phase I cohort size was 3000, the prevalence was 0.05 and 0.10, the correlation parameter ρ for phase I variables was 0 and 0.3, and the true value of β_4 was 0.5, 0.7, and 0.9. The mean asymptotic standard error ("asym"), empirical standard error ("emp"), and coverage probability ("coverage") of $\hat{\beta}_4$ were calculated based on 1000 simulations.

$P(Y=1)$	ρ	β_4	GOF (asym/emp)	coverage	CC (asym)	BD (asym)
0.05	0	0.5	0.52 (0.21/0.21)	0.951	0.52 (0.23)	0.53 (0.22)
		0.7	0.72 (0.21/0.22)	0.948	0.71 (0.24)	0.72 (0.23)
		0.9	0.93 (0.22/0.22)	0.947	0.92 (0.25)	0.93 (0.24)
	0.3	0.5	0.52 (0.17/0.17)	0.948	0.51 (0.18)	0.52 (0.18)
		0.7	0.72 (0.17/0.18)	0.941	0.71 (0.19)	0.72 (0.19)
		0.9	0.93 (0.18/0.19)	0.935	0.93 (0.21)	0.93 (0.20)
0.10	0	0.5	0.51 (0.14/0.14)	0.951	0.51 (0.17)	0.51 (0.16)
		0.7	0.71 (0.15/0.15)	0.939	0.70 (0.16)	0.71 (0.16)
		0.9	0.91 (0.15/0.15)	0.953	0.91 (0.17)	0.91 (0.17)
	0.3	0.5	0.50 (0.12/0.12)	0.943	0.50 (0.13)	0.50 (0.13)
		0.7	0.71 (0.12/0.12)	0.951	0.71 (0.13)	0.71 (0.13)
		0.9	0.91 (0.13/0.13)	0.944	0.92 (0.14)	0.91 (0.14)

ing phase II covariates and most phase I covariates at lower outcome prevalence, lower correlation among phase I variables and larger true effect size of the phase II covariate.

2.4. Insight into the Efficiency of GOF

Figure 2.1 showed that GOF had superior efficiency for phase II covariates and phase I covariates except for that used for stratification (X_1) in BD, and for the latter BD may have higher efficiency. To gain insight into the efficiency advantage of GOF, we performed additional simulation studies using the same settings as in Section 2.3 except that the cohort size was 2×10^4. This large sample size admittedly is not typical in epidemiological studies, but it ensures that sampling based directly on $S(y, \boldsymbol{x})$ will yield sufficient subjects for inclusion in phase II. We therefore considered this setting to inform the full advantage of sampling based on "goodness-of-fit". For GOF, we first generated a subset using $S(y, \boldsymbol{x})$ as the sampling probability, and then further selected a random sample of $m = 600$ subjects from the subset as the phase II sample. The case-control sampling included 300 cases and 300 controls. In the balanced sampling, we randomly sampled 150 subjects from each of the four strata formed by the dichotomized X_1 and Y. The full results in this big cohort scenario are presented in Tables C.7 - C.11 in APPENDIX C.



(a) $\rho = 0$ (b) $\rho = 0.3$

(c) $\rho = 0$ (d) $\rho = 0.3$

Figure 2.1: Mean asymptotic variance of the estimated log OR for phase II covariate ($\hat{\beta}_4$, panels a and b) and phase I stratifying covariate ($\hat{\beta}_1$, panels c and d) under the case-control sampling (CC), the balanced sampling (BD), and the goodness-of-fit based sampling (GOF). The cohort size was 3000 with $P(Y = 1) = 0.05$, and the true value of β_4 was between 0.5–0.9. Phase I variables were uncorrelated in panels a and c and modestly correlated in panels b and d.

Because $S(y, \boldsymbol{x})$ indicated the goodness-of-fit of the external model $P^e(Y = 1|\boldsymbol{x})$, we exploited the standardized Pearson residuals to investigate the nature of the efficiency gain for GOF. The standardized Pearson residuals are commonly used to assess the goodness-of-fit of logistic regression models (Agresti and Kateri, 2011). Let P_i^e denote $P^e(Y = 1|\boldsymbol{x}_i)$. If the external model $P^e(Y = 1|\boldsymbol{x})$ were the true model, the standardized Pearson residual for the ith subject is defined as

$$d_i = \frac{y_i - \hat{p}_i^e}{\sqrt{\hat{p}_i^e(1 - \hat{p}_i^e)(1 - \hat{h}_i)}}.$$

Note that the denominator in this expression is the estimated standard error of the numerator with $\hat{h}_i$ being the ith subject's estimated leverage. Larger absolute values of d indicates worse goodness-of-fit of $P^e(Y = 1|\boldsymbol{x})$.

For illustration purpose, we present in Figure 2.2 results in the scenario where GOF achieved the highest efficiency relative to other designs for estimating β_4, the log OR of phase II covariate Z. The outcome prevalence was 0.05, the correlation among phase I variables was 0, and β_4 equaled 0.9. Panels 2.2a and 2.2b demonstrate the relationship between $|d|$ and the sampling probability $P(R = 1|y, \boldsymbol{x})$ in GOF. As $P(R = 1|y, \boldsymbol{x})$ increases, the magnitude of $|d|$ also increases. In other words, subjects who have worse goodness-of-fit are more likely to be selected into phase II under GOF. Panels 2.2c and 2.2d display the distribution of the empirical mean $|d|$ in the phase II sample under simple random sampling (RS), CC, BD, and GOF separately for cases and controls, which was estimated based on 1000 simulated datasets. While the mean of $|d|$ for cases was comparable under BD and GOF, the mean for controls under GOF was significantly larger than that under RS, CC, or BD. GOF improved efficiency for estimating β_4 by 27% compared with BD, and by 33% compared with CC. Our reasoning for the improved efficiency of GOF therefore is as follows. Because the full model (2.1) was the *true* model, lack of fit based on P^e would be suggestive of the necessity to include $\boldsymbol{Z}$ in the model in order to achieve better goodness-of-fit. The subjects who have larger contributions to lack-of-fit are therefore more supportive that phase II covariates should be included in the model. GOF also improved efficiency by 20% for X_1, 30% for X_2 and 31% for X_3 compared with CC. The efficiency of GOF relative to BD was 0.86 for X_1, 1.23 for X_2 and 1.25 for X_3.

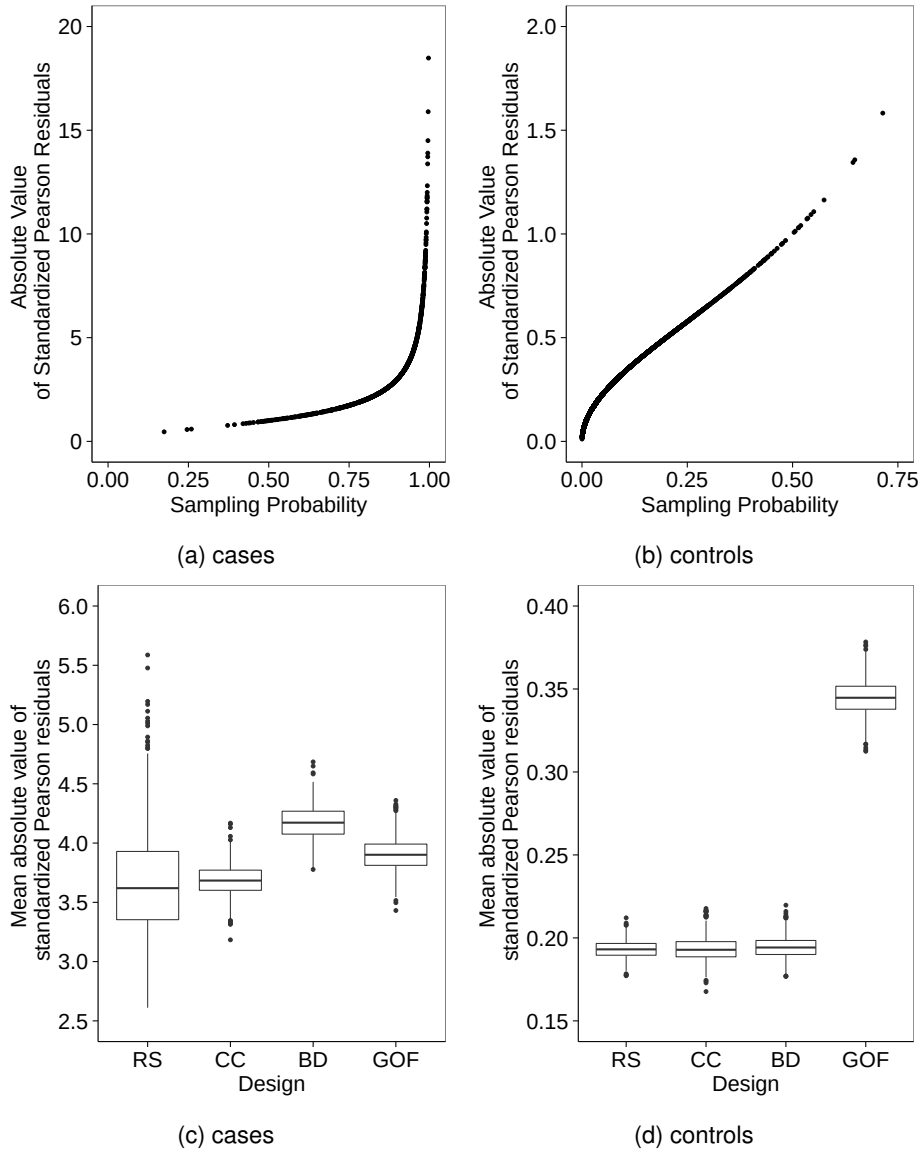


Figure 2.2: Insights into the relative efficiency of GOF. Panels a and b show the relationship between $|d|$ and $P(R = 1|y, \boldsymbol{x})$ separately for cases and controls. Panels c and d show the distribution of the mean $|d|$ from 1000 simulated datasets in the phase II sample among cases and controls. The phase I cohort size was 2×10^4 with $P(Y = 1) = 0.05$, and the log OR of the phase II variable Z was 0.9.

2.5. Illustration of GOF in a Real Study Setting

In this section, we used data from a biomarker study of gestational diabetes (Zhu et al., 2016) to illustrate the implementation and performance of GOF in a real study setting. This study consisted of 2,701 pregnant women among whom 100 developed gestational diabetes mellitus. The median age was 28 years, and the median BMI was 24.2 with an interquartile range of 21.6–28.1. Approximately equal proportions of the participants were non-Hispanic White (27$\%$), non-Hispanic black (28$\%$), and Hispanic (29$\%$), with a smaller proportion (16$\%$) being Asian. Among these participants, 22$\%$ had a family history of gestational diabetes. To compare different designs, we treated age, race, and BMI as phase I covariates $\boldsymbol{X}$ and family history as phase II covariate Z. Let $I()$ denote the indicator function. We bootstrapped $(\boldsymbol{X}, Z)$ for all 2,701 subjects and generated gestational diabetes mellitus, denoted by Y, from model

$$\begin{aligned}\text{logit } P(Y=1|\text{age, race, BMI, family history}) &= -8.38 + 0.073 \times \text{age} - \\ &0.54 \times I(\text{non-Hispanic Black}) + 0.47 \times I(\text{Hispanic}) + \\ &0.71 \times I(\text{Asian}) + 0.10 \times \text{BMI} + 0.57 \times I(\text{family history}),\end{aligned}$$

which was developed from the original dataset. We considered three external logistic regression models for sampling, each using a different subset of phase I covariates as predictors: race only (model e1), race and age (model e2), and race, age and BMI (model e3). This allowed us to evaluate the efficiency of GOF when an increasing amount of phase I information becomes available. The regression parameters in the external model were obtained from the original dataset. We sampled all the 100 cases and twice as many controls into phase II. For GOF, in order to ensure that 200 controls can be sampled, we aimed to select 600 into the subset $\{i : R_i = 1, i = 1, 2, \ldots, N\}$. Therefore, we estimated the constant c_0 roughly as the ratio of 600 over the empirical mean number of controls, obtained as the sum of the sampling probability p^e for all the controls in the full cohort of 2,701 individuals. We repeated the simulation 1,000 times. In BD sampling, we stratified on covariate race because race was most strongly correlated with phase II covariate family history. We also considered race and BMI as phase II covariates and the matching variables used in BD are race and age, respectively.

Table 2.2 presents the point estimate and mean asymptotic standard error of the coefficient for

Table 2.2: The point estimate of the log OR for phase II covariate and its mean asymptotic standard error (SE) under the goodness-of-fit based design (GOF), the case-control design (CC) and the balanced design (BD). The point estimate from the full cohort for family history was 0.57, for BMI was 0.10, and for race was: Black -0.54, Hispanic 0.47, Asian 0.71. Relative efficiency was calculated as the asymptotic variance under CC or BD over that of GOF.

Phase II	Sampling		Estimate	Relative Efficiency	
Variable	Design	External Model	(SE)	vs. CC	vs. BD
Family	GOF	e1: race	0.582 (0.29)	1.02	0.97
History		e2: race+age	0.577 (0.29)	1.07	1.02
		e3: race+age+BMI	0.575 (0.27)	1.19	1.12
	CC		0.579 (0.30)		
	BD		0.580 (0.29)		
BMI	GOF	e1: race	0.104 (0.03)	0.99	0.97
		e2: race+age	0.104 (0.02)	1.02	1.00
		e3: race+age+fh	0.102 (0.02)	1.04	1.02
	CC		0.103 (0.02)		
	BD		0.102 (0.02)		
Race	GOF	e1: age	-0.584 (0.43)	1.02	1.00
			0.468 (0.34)	1.06	1.01
			0.719 (0.39)	1.08	1.01
		e2: age+fh	-0.587 (0.43)	1.04	1.01
			0.466 (0.34)	1.07	1.01
			0.713 (0.39)	1.09	1.02
		e3: age+fh+BMI	-0.544 (0.41)	1.15	1.12
			0.473 (0.33)	1.18	1.11
			0.744 (0.39)	1.07	0.99
	CC		-0.570 (0.44)		
			0.474 (0.36)		
			0.707 (0.41)		
	BD		-0.586 (0.43)		
			0.480 (0.35)		
			0.715 (0.39)		

phase II covariate and the relative efficiency with respect to CC and BD, when family history, BMI and race were considered as phase II covariate, respectively. For example, when family history was considered as phase II covariate, among the set of nested external models e1, e2 and e3, the efficiency of GOF ranged from 1.02 to 1.19 relative to CC and from 0.97 to 1.12 compared with BD. The results can be intuitively explained as follows. As richer phase I covariates are included in the external model, it becomes more likely that the lack of goodness-of-fit is caused by phase II covariates that were absent from the external model. Therefore, subjects who are better indicative of lack-of-fit contributed more to the improvement on statistical efficiency for assessing phase II covariates. Results were similar when we considered BMI or race as phase II covariate.

2.6. Discussion

The LCC sampling design (Fithian and Hastie, 2014) was originally proposed in the big data literature to increase computational efficiency. We found that it sheds new light on two-phase sampling. Through GOF, we offer a new perspective to define "informative" subjects for efficient sampling. When it is necessary to select a subset of individuals for measuring expensive variables, our proposed GOF is a powerful sampling method to improve efficiency compared with case-control and balanced designs. A unique feature of GOF is that it is highly efficient for estimating ORs for phase II covariates. The balanced design was originally proposed to improve efficiency of assessing a rare exposure by oversampling the "exposed" subjects while adjusting for expensive confounding variables measured at phase II. When phase II covariates are of interest, BD may only have marginal efficiency advantage compared with CC depending on the relationship between the exposure and phase II covariates. The efficiency advantage of GOF is derived from oversampling more informative subjects who have worse goodness-of-fit based on a preliminary model using only phase I covariates as predictors.

GOF and BD improve statistical efficiency for estimating association parameters via different mechanisms, that explains their respective advantages of greater efficiency for phase II covariates and the phase I matching covariates. When only phase II covariates are of interest, GOF is highly preferred. Ideally, an efficient design without sacrificing the efficiency of the matching variables may have a wider range of application. Therefore, in chapter 3, we will propose a hybrid design that inherits the advantages of both GOF and BD, that is able to achieve high efficiency for both phase

I and phase II covariates.

The advantage of GOF lies in the fact that it makes full use of all phase I covariate data to determine the sampling probability without coarsening. This very fact is also an important practical advantage of GOF: it relieves data analysts from the necessary step of making decisions on creating phase I sampling strata in BD. In practice, it may be necessary in GOF to experiment with sampling probabilities to achieve pre-specified numbers of phase II cases and controls as we considered in both simulation studies and the real data example. We suggest that the sample size of the subset V in the first step of GOF be on average two to three times larger than the targeted phase II sample size. This was found to work well in our numerical studies, because those who had a poorer goodness-of-fit were still able to maintain a greater probability of being sampled. The high efficiency of GOF coupled with the fact that data from GOF can be analyzed using standard software and existing R packages makes GOF an attractive sampling method for reducing study cost while maintaining statistical efficiency.

CHAPTER 3

THE BALANCED GOODNESS-OF-FIT BASE SAMPLING DESIGN

3.1. Introduction

Outcome-dependent sampling strategies have been widely applied in biomedical studies. The most well-known sampling schemes are the case-control design for studying a rare binary outcome and the "balanced design" (Breslow and Cain, 1988) where cases and controls are further matched on several discrete exposure variables. The case-control design improves statistical efficiency by increasing the proportion of cases in the sample compared with that in the full population. The balanced design is more efficient at estimating association between an outcome and a rare exposure variable, heuristically because it oversamples the "exposed" subjects. The key to the improved efficiency of an oversampling scheme is to identify informative subjects. In these well-known designs, subjects in rare groups are generally considered more informative. Appropriate statistical methods are required to account for the oversampling scheme in order to obtain unbiased association parameter estimates.

It is economical to collect some covariates only for a subgroup of subjects in an association study for a binary outcome. The resultant incomplete data structure is usually described using a two phase sampling scheme (Neyman, 1938; White, 1982), where the outcome and the completely observed covariates are collected on all subjects at phase I and the remaining covariates are collected on a subgroup at phase II. Different strategies for subgroup selection affect the efficiency for estimating association parameters. The oversampling scheme can be implemented in two-phase designs for the purpose of greater efficiency. Existing oversampling designs mainly focus on improving efficiency for estimating association parameters with respect to phase I covariates. However, phase II covariates are of great importance in studies where covariates of interest cannot be fully collected for economic reasons. An ideal sampling design, of course, is able to increase efficiency for estimating association for phase II covariates without sacrificing that for phase I covariates.

In the "big data" literature, efficient sampling designs have been frequently implemented in order to increase computational efficiency. Although outcome and covariates are available for all subjects, data reduction is desirable to reduce computation burden. An innovative sampling scheme, the

"local case-control design" (LCC) Fithian and Hastie (2014) , was proposed recently for studying binary outcomes in this context. It oversamples subjects who have lower predicted probability of having their true case or control status based on a preliminary model and achieved greater efficiency compared with case-control sampling. In Chapter 2, we extended LCC to the two-phase sampling setting, and developed a goodness-of-fit based sampling design (GOF). In a common scenario where an external model is available to relate the outcome variable and phase I covariates, GOF oversamples cases and controls who have worse goodness-of-fit based on the external model. GOF provides a new perspective to define informative subjects in oversampling designs. Those who lack goodness-of-fit based on the external model only including phase I covariates indicate the necessity to incorporate phase II covariates into the model for a better fit and thus are considered more informative with respect to the phase II covariates. It has been shown that GOF has a great advantage of increasing efficiency for estimating odds ratio (OR) parameters for the incomplete phase II covariates, but the balanced design gains great efficiency for estimating the matching variable. Comparing the efficiency of GOF and the balanced sampling motivated us to propose a new two-phase sampling scheme, balanced goodness-of-fit based sampling (BGOF) in this chapter, which performs GOF sampling first and then balanced sampling on the GOF subsample. BGOF is easy to implement, and we found that it consistently outperforms case-control, balanced, and GOF sampling in both simulated and real data settings.

The rest of this chapter is organized as follows. In section 3.2, we describe the sampling and estimation procedures of BGOF, and compare its efficiency with GOF, case-control (CC) and balanced designs (BD) for estimating OR parameters. In section 3.4, we further evaluate BGOF using data from an ongoing biomarker study of gestational diabetes. We discuss future extensions in Section 3.5.

3.2. Balanced Goodness-of-Fit Based Design

In this section, we describe sampling and inferential procedures of BGOF and we use the same notation as for GOF in Chapter 2. Let Y denote the binary outcome variable with $Y = 1$ indicating cases and $Y = 0$ controls. Let $\boldsymbol{X}$ denote phase I covariates that are available for all subjects, and $\boldsymbol{Z}$ denote phase II covariates that can only be measured on a subset of subjects. A logistic regression

model is used to describe the relationship between Y and covariates $\boldsymbol{X}$ and $\boldsymbol{Z}$,

$$\text{logit } P(Y=1|\boldsymbol{x},\boldsymbol{z}) = \boldsymbol{x}\boldsymbol{\beta}_1 + \boldsymbol{z}\boldsymbol{\beta}_2,$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are the OR parameters of interest. Note here to simplify notation, a variable with value equal to one is implicitly included in $\boldsymbol{X}$, with the corresponding regression coefficient in $\boldsymbol{\beta}_1$ being the intercept parameter. Let $\boldsymbol{\beta}$ denote the vector of all parameters $(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$. The outcome and phase I variables $(Y, \boldsymbol{X})$ are collected from a cross-sectional sample of N subjects. We wish to select a subset of $m\,(m < N)$ subjects for the measurement of $\boldsymbol{Z}$.

3.2.1. Sampling Procedure

Motivated by the high efficiency of BD for estimating ORs of the phase I stratifying covariates and the high efficiency of GOF for estimating ORs of all the remaining covariates that was illustrated in Chapter 2, we propose a novel hybrid design that unifies the advantages of both sampling methods. This new design performs phase II sampling in two steps, GOF followed by BD. Therefore we term this new design the balanced goodness-of-fit based sampling. At the first step, a subset $V = \{i : R_i = 1, i = 1, \ldots N\}$ is generated using GOF, where R indicates selection ($R = 1$: yes; $R = 0$: no) and M is the number of individuals in subset V, i.e., $M \equiv \sum_{i=1}^{N} R_i$. At the second step, BD sampling is further performed on V. Discrete strata L are formed based on phase I variables $\boldsymbol{X}$, and subjects in V are cross-classified according to Y and L. Let M_{yl} denote the total number of subjects in stratum defined by $Y = y$ and $L = l$, i.e., $M_{yl} \equiv \sum_{i=1}^{M} I(Y_i = y, L = l)$. The balanced sampling randomly selects m_{yl} ($m_{yl} < M_{yl}$) subjects from cell $Y = y$ and $L = l$. With multiple phase I covariates, a decision needs to be made about how to form strata L, for example, which variables to stratify on, or how to categorize a continuous variable. We suggest that variables of which statistical efficiency are of great interest, e.g., a rare exposure, have priority. The number of strata should be chosen to guarantee that M_{yl} is not too small.

3.2.2. Statistical Inference

Let π_{yl} denote the sampling probability in cell $Y = y$ and $L = l$, $\pi_{yl} \equiv m_{yl}/M_{yl}$. To make statistical inference under BGOF, huristically, BGOF can be seen as a balanced design in cohort V which is randomly generated from the distribution $P(Y = 1|\boldsymbol{x}, \boldsymbol{z}, R = 1)$ as defined in equation (2.3).

Therefore, motivated by methods of statistical inference for standard two-phase designs (Breslow and Cain, 1988; Scott and Wild, 1991), we propose to estimate OR parameters $\boldsymbol{\beta}$ by solving an estimating equation that takes the same form as (2.4) but with μ_i^g replaced by μ_i^l:

$$\mu_i^l = \text{expit}\left\{\boldsymbol{w}_i\boldsymbol{\beta} + o(\boldsymbol{x}_i) + \log\frac{\pi_{1l}}{\pi_{0l}}\right\},$$

where the ith subject is in stratum l. Similar to estimation under GOF, the corresponding estimator $\hat{\boldsymbol{\beta}}$ can be obtained using standard software for fitting logistic regression models with offset term $o(\boldsymbol{x}_i) + \log(\pi_{1l}/\pi_{0l})$ for the ith observation in stratum l. We present in the corollary below the large sample properties of $\hat{\boldsymbol{\beta}}$ and present the proof in APPENDIX A. Let E_{yl}^* and E_l^* denote expectations with respect to the distribution of $\boldsymbol{W}$ in cell (y,l) and stratum l, respectively. Let ψ_{yl} denote the large sample limit of M_{yl}/N as $N\to\infty$, γ_{yl} that of m_{yl}/m as $m\to\infty$, and θ that of m/N as $N\to\infty$.

Corollary *Suppose that the true value $\boldsymbol{\beta}$ lies inside a compact space, and that component of $(\boldsymbol{X},\boldsymbol{Z})$ are bounded. Define matries H and G as*

$$\begin{aligned} H &\equiv \sum_{y=0}^{1}\sum_{l=1}^{L}\mathrm{E}_{yl}^*\{\boldsymbol{w}_i^T\boldsymbol{w}_i\mu_i^l(1-\mu_i^l)\} \\ G &\equiv \sum_{y=0}^{1}\sum_{l=1}^{L}\left(\gamma_{yl}^{-1}-\theta\psi_{yl}^{-1}\right)\mathrm{E}_l^*\left\{\boldsymbol{w}_i^T\mu_i^l\left(1-\mu_i^l\right)\right\}\mathrm{E}_l^*\left\{\boldsymbol{w}_i\mu_i^l\left(1-\mu_i^l\right)\right\}. \end{aligned}$$

We show that $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normally distributed with

$$\sqrt{m}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{D} N\left\{0, H^{-1}(H-G)H^{-1}\right\}.$$

The asymptotic variance-covariance matrix $H^{-1}(H-G)H^{-1}$ can be consistently estimated empirically with

$$\begin{aligned} \hat{H} &= \frac{1}{m}\sum_{y=0}^{1}\sum_{l=1}^{L}\sum_{i=1}^{m_{yl}}\boldsymbol{w}_i^T\boldsymbol{w}_i\mu_i^l\left(1-\mu_i^l\right), \\ \hat{G} &= \frac{1}{m}\sum_{y=0}^{1}\sum_{l=1}^{L}\left[\left(m_{yl}^{-1}-M_{yl}^{-1}\right)\sum_{y=0}^{1}\sum_{i=1}^{m_{yl}}\left\{\boldsymbol{w}_i^T\mu_i^l\left(1-\mu_i^l\right)\right\}\sum_{y=0}^{1}\sum_{i=1}^{m_{yl}}\left\{\boldsymbol{w}_i\mu_i^l\left(1-\mu_i^l\right)\right\}\right]. \end{aligned}$$

The asymptotic variance of $\hat{\boldsymbol{\beta}}$ differed from that in Proposition 1 of Breslow and Cain (1988) only by

an extra term that represents variance reduction owing to the case-control sampling in phase I in the latter. The same difference was observed for standard prospective and retrospective two-phase designs which differ only in whether phase I is a simple random or retrospective sample (Scott and Wild, 1991). These connections imply that the R package "osDesign" (Haneuse, Saegusa, and Lumley, 2011) for the standard two-phase prospective design (option "cohort=TRUE") can be used directly for obtaining variance estimates for $\hat{\boldsymbol{\beta}}$, with the GOF subsample treated as phase I sample. In other words, although GOF generates an outcome-dependent subsample because the sampling probability depends on Y, in terms of variance estimation, it can be safely treated as a prospective sample. As explained in APPENDIX A, the key to this interesting result is that the pseudo-model for the GOF sub-sample, $P(Y = 1|\boldsymbol{x}, \boldsymbol{z}; R = 1)$, is a genuine probability function.

3.3. Simulation Studies

We conducted simulation studies to compare the efficiency of BGOF, GOF, BD, and CC using the same setup as for GOF in Chapter 2. The results on estimation of β_4 under BGOF are presented in Table 3.1. Across all the simulation scenarios, the averaged estimates were close to the true values, the empirical and mean asymptotic standard errors were similar, and the coverage probabilities were nearly at the 95% nominal level. Similar results for three OR parameters of phase I covariates are presented in Tables C.1, C.2, and C.3 in APPENDIX C.

Table 3.2 presents the mean asymptotic variance of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$ under BGOF and its efficiency relative to the CC, BD, and GOF sampling designs, calculated as the ratio between the mean asymptotic variance under each of the other three designs and that under BGOF. BGOF and GOF achieved comparable efficiency for estimating $\hat{\beta}_2$, $\hat{\beta}_3$, and $\hat{\beta}_4$. Notably, BGOF had the highest efficiency for estimating $\hat{\beta}_1$. Across all parameter settings, BGOF improved efficiency over CC by 27–34$\%$ for $\hat{\beta}_1$, 20–34$\%$ for $\hat{\beta}_2$, 14–34$\%$ for $\hat{\beta}_3$, and 18–34$\%$ for $\hat{\beta}_4$. Compared with BD, BGOF improved efficiency by 8–21$\%$ for $\hat{\beta}_1$, 16–24$\%$ for $\hat{\beta}_2$, 10–24$\%$ for $\hat{\beta}_3$, and 15–28$\%$ for $\hat{\beta}_4$. In general, BGOF achieved higher efficiency relative to CC and BD at lower outcome prevalence, lower correlation among phase I variables and larger true effect size of the phase II covariate. For comparison purpose, we also used the binary variable X_3 as the stratifying variable in BGOF and BD and the results were largely similar and are presented in Table C.4.

Table 3.2 showed that the correlation among phase I variables might affect the efficiency of BGOF

Table 3.1: The estimated log OR of phase II covariate ($\hat{\beta}_4$) under balanced goodness-of-fit based design (BGOF). The phase I cohort size was 3000, the prevalence was 0.05 and 0.10, the correlation parameter ρ for phase I variables was 0 and 0.3, and the true value of β_4 was 0.5, 0.7, and 0.9. The mean asymptotic standard error ("asym"), empirical standard error ("emp"), and coverage probability ("coverage") of $\hat{\beta}_4$ were calculated based on 1000 simulations.

$P(Y=1)$	ρ	β_4	BGOF (asym/emp)	coverage
0.05	0	0.5	0.52 (0.20/0.21)	0.949
		0.7	0.70 (0.21/0.21)	0.949
		0.9	0.92 (0.22/0.21)	0.946
	0.3	0.5	0.52 (0.17/0.17)	0.955
		0.7	0.72 (0.17/0.18)	0.947
		0.9	0.92 (0.18/0.19)	0.945
0.10	0	0.5	0.51 (0.14/0.14)	0.950
		0.7	0.71 (0.15/0.14)	0.943
		0.9	0.91 (0.15/0.15)	0.954
	0.3	0.5	0.50 (0.12/0.12)	0.948
		0.7	0.70 (0.12/0.13)	0.938
		0.9	0.91 (0.13/0.13)	0.944

relative to BD, that is, the relative efficiency decreased with the correlation. Then a question arose about whether the correlation between phase I and phase II variables may affect the relative efficiency of BGOF versus BD. Therefore, we compared the four designs in a more extreme scenario where phase II covariate Z was independent of all phase I covariates $\boldsymbol{X}$ with the remaining setup identical to that above. For estimating the coefficient of the binary variable X_3, BGOF lost 3–7$\%$ efficiency compared with CC and BD when the three phase I covariates were uncorrelated, but BGOF was slightly more efficient when they were moderately correlated, i.e., ρ = 0.3 (Table C.5 in APPENDIX C). For estimating ORs of X_2 and Z, BGOF remained the most efficient in all parameter settings, although the gain became smaller compared with scenarios where phase I and phase II covariates were correlated. BGOF and BD had similar efficiency for the stratifying variable X_1. We next investigated whether BGOF using X_3 as the stratifying variable could boost efficiency of BGOF for estimating its OR relative to BD, but it did not help (Table C.6). We also evaluated the performance of BGOF in all the settings described above except with a large phase I cohort of size 2×10^4 rather than 3000 and phase II sample size 600. The results are largely similar and presented in Table C.7 - C.14.

Lastly, because GOF and BGOF required an existing external model, we examined the robustness of their efficiency with respect to the "accuracy" of the external model, that is, how closely this

Table 3.2: Asymptotic variance of $\hat{\boldsymbol{\beta}}$ under the balanced goodness-of-fit based design (BGOF) and its efficiency relative to the goodness-of-fit based design (GOF), the balanced design (BD), and the case-control design (CC). The phase I cohort size was 3000. The prevalence was 0.05 and 0.10, the correlation parameter ρ for phase I variables was 0 and 0.3, and the true value of β_4 was 0.5, 0.7, and 0.9. The correlation between phase II variable X_4 and phase I variables X_1, X_2, and X_3 was 0.6, 0.5, and 0.3, respectively.

$P(Y=1)$	ρ	β_4	Var($\hat{\beta}_4$)	RE[a] of BGOF vs.			Var($\hat{\beta}_1$)	RE of BGOF vs.		
				GOF	BD	CC		GOF	BD	CC
0.05	0	0.5	0.042	1.00	1.20	1.23	0.32	1.14	1.13	1.32
		0.7	0.043	1.01	1.24	1.29	0.33	1.14	1.18	1.34
		0.9	0.046	1.01	1.28	1.34	0.36	1.13	1.21	1.34
	0.3	0.5	0.028	1.00	1.15	1.20	0.25	1.17	1.08	1.29
		0.7	0.030	1.01	1.18	1.24	0.27	1.16	1.11	1.28
		0.9	0.033	1.01	1.20	1.28	0.30	1.16	1.15	1.26
0.10	0	0.5	0.020	1.00	1.18	1.21	0.15	1.14	1.12	1.31
		0.7	0.021	1.01	1.20	1.25	0.16	1.13	1.15	1.32
		0.9	0.023	1.01	1.24	1.29	0.17	1.13	1.18	1.33
	0.3	0.5	0.014	1.00	1.14	1.18	0.12	1.17	1.09	1.29
		0.7	0.015	1.01	1.16	1.22	0.13	1.16	1.11	1.28
		0.9	0.016	1.01	1.17	1.24	0.14	1.15	1.13	1.27
			Var($\hat{\beta}_2$)	RE of BGOF vs.			Var($\hat{\beta}_3$)	RE of BGOF vs.		
				GOF	BD	CC		GOF	BD	CC
0.05	0	0.5	0.026	1.00	1.22	1.26	0.094	1.00	1.10	1.13
		0.7	0.026	1.01	1.24	1.30	0.090	1.01	1.18	1.22
		0.9	0.027	1.02	1.27	1.34	0.090	1.01	1.24	1.30
	0.3	0.5	0.021	1.00	1.19	1.24	0.077	1.00	1.12	1.21
		0.7	0.021	1.01	1.22	1.29	0.076	1.00	1.17	1.28
		0.9	0.022	1.01	1.22	1.31	0.078	1.01	1.19	1.34
0.10	0	0.5	0.013	1.02	1.18	1.21	0.044	1.00	1.11	1.14
		0.7	0.013	1.03	1.19	1.24	0.043	1.00	1.16	1.20
		0.9	0.014	1.03	1.20	1.27	0.044	1.01	1.21	1.27
	0.3	0.5	0.010	1.01	1.16	1.20	0.038	1.01	1.11	1.19
		0.7	0.011	1.02	1.16	1.22	0.038	1.01	1.14	1.25
		0.9	0.011	1.02	1.16	1.24	0.039	1.02	1.16	1.30

[a] Relative Efficiency (RE): calculated as the asymptotic variance under each design over that of BGOF

external model reflected the true relationship between Y and $\boldsymbol{X}$ in the data. To this end, we deliberately changed $\boldsymbol{\eta}_1$ by 10% from the "true" values to mimic the scenario where the external information deviated from the truth. BGOF still maintained the highest efficiency for both phase I and phase II covariates (Figure 3.1). GOF and BGOF were relatively robust, and the efficiency advantage of BGOF may decrease as the external model becomes far less accurate.

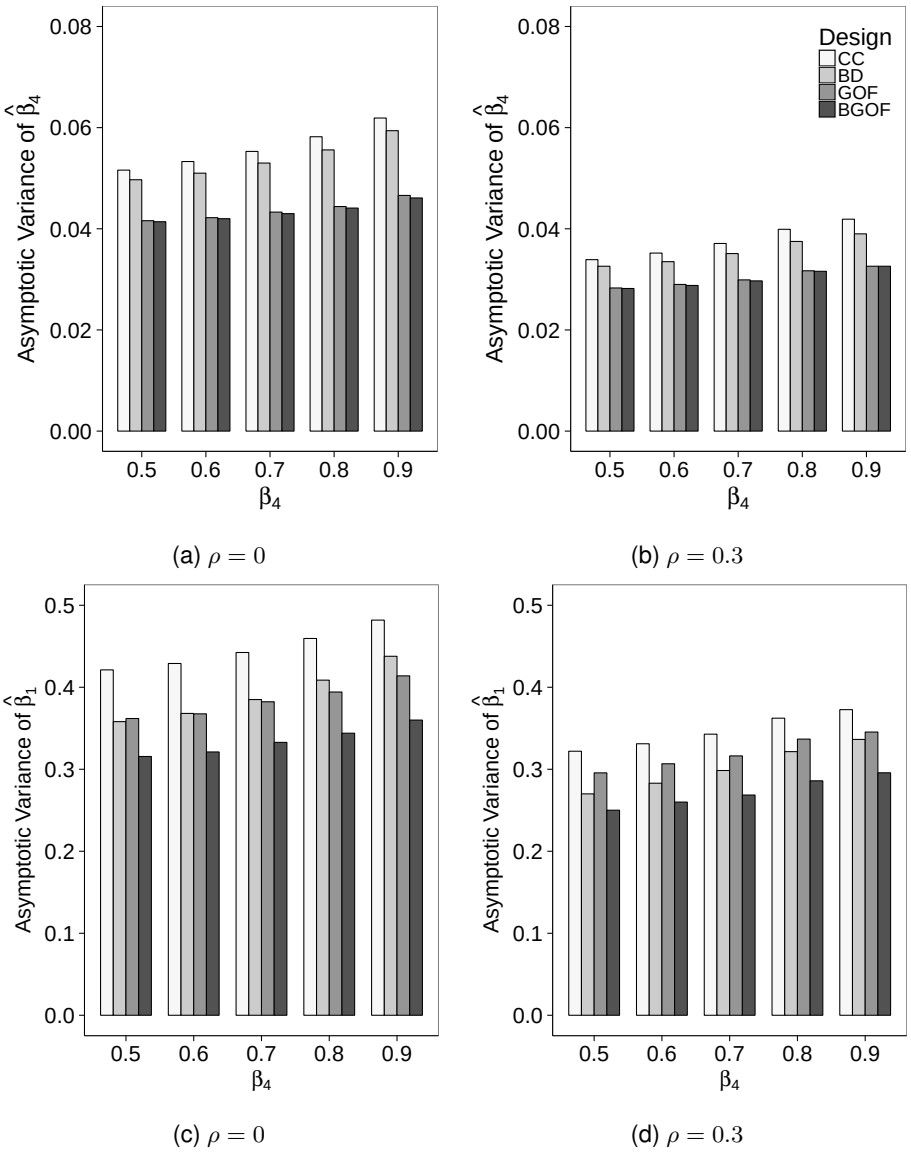
3.4. Application of BGOF in a Biomarker Study of Genstational Diabetes

In this section, we considered the same biomarker study of gestational diabetes (Zhu et al., 2016) as in Chapter 2 to illustrate the implementation and performance of BGOF. This study consisted of 2,701 pregnant women among whom 100 developed gestational diabetes mellitus. We included four standard risk factors for gestational diabetes, i.e., age, BMI, race and family history of gestational diabetes. We used the same settings as in Chapter 2, section 2.5. To perform BGOF, in order to ensure that 200 controls can be sampled, we aimed to select 600 into the subset $\{i : R_i = 1, i = 1, 2, \ldots, N\}$ at the GOF step. Therefore, we estimated the constant c_0 roughly as the ratio of 600 over the empirical mean number of controls, obtained as the sum of the sampling probability p^e for all the controls in the full cohort.

Table 3.3 presents the point estimate and mean asymptotic standard error of the coefficient for phase II covariate and the relative efficiency compared with CC and BD. When family history was considered the phase II covariate, among the set of nested external models e1, e2 and e3, the efficiency of BGOF ranged from 1.05 to 1.23 relative to CC and from 1.00 to 1.16 relative to BD. BGOF achieved slightly higher efficiency compared with GOF (Table 2.2). At the GOF step, as richer phase I covariates are included in the external model, the lack of goodness-of-fit is more likely to be caused by phase II covariates absent from the external model. Therefore, subjects who are better indicative of lack-of-fit are more informative with respect to phase II covariates and have a greater probability of being selected into phase II. Results were similar when we considered BMI or race as phase II covariate.

3.5. Discussion

The key working mechanism of an oversampling scheme is to identify informative subjects and to increase their proportion in the sample. In existing designs, subjects in rare groups are generally



(a) $\rho = 0$ (b) $\rho = 0.3$

(c) $\rho = 0$ (d) $\rho = 0.3$

Figure 3.1: Mean asymptotic variance of the estimated log OR for phase II covariate ($\hat{\beta}_4$, panels a and b) and phase I stratifying covariate ($\hat{\beta}_1$, panels c and d) under the case-control sampling (CC), the balanced sampling (BD), the goodness-of-fit based sampling (GOF) and the balanced goodness-of-fit based sampling (BGOF). The cohort size was 3000 with $P(Y = 1) = 0.05$, and the true value of β_4 was between 0.5–0.9. Phase I variables were uncorrelated in panels a and c and modestly correlated in panels b and d. In the external model, $\boldsymbol{\eta}_1$ was deliberately increased by 10%.

Table 3.3: The point estimate of the log OR for phase II covariate and its mean asymptotic standard error (SE) under the balanced goodness-of-fit based design (BGOF), the case-control design (CC) and the balanced design (BD). The point estimate from the full cohort for family history was 0.57, for BMI was 0.10 and for race was: Black -0.54, Hispanic 0.47, Asian 0.71. Relative efficiency was calculated as the asymptotic variance under CC or BD over that of BGOF or GOF.

Phase II Variable	Sampling Design	External Model	Estimate (SE)	Relative Efficiency vs. CC	Relative Efficiency vs. BD
Family History	BGOF	e1: race	0.584 (0.29)	1.05	1.00
		e2: race+age	0.579 (0.28)	1.11	1.05
		e3: race+age+BMI	0.575 (0.27)	1.23	1.16
	CC		0.579 (0.30)		
	BD		0.580 (0.29)		
BMI	BGOF	e1: race	0.103 (0.02)	1.01	0.99
		e2: race+age	0.103 (0.02)	1.05	1.03
		e3: race+age+fh	0.103 (0.02)	1.07	1.04
	CC		0.103 (0.02)		
	BD		0.102 (0.02)		
Race	BGOF	e1: age	-0.600 (0.43)	1.03	1.00
			0.467 (0.34)	1.06	1.01
			0.709 (0.39)	1.08	1.01
		e2: age+fh	-0.586 (0.43)	1.04	1.01
			0.469 (0.34)	1.08	1.02
			0.713 (0.39)	1.09	1.02
		e3: age+fh+BMI	-0.555 (0.41)	1.15	1.12
			0.473 (0.33)	1.18	1.12
			0.732 (0.39)	1.07	0.99
	CC		-0.570 (0.44)		
			0.474 (0.36)		
			0.707 (0.41)		
	BD		-0.586 (0.43)		
			0.480 (0.35)		
			0.715 (0.39)		

considered more informative. Through GOF and BGOF, we provide a new perspective to define "informative" subjects for efficient sampling. When it is necessary to select a subset of individuals for measuring expensive variables, our proposed BGOF is a powerful sampling method to improve efficiency compared with case-control and balanced designs. The efficiency advantage of both GOF and BGOF is derived from oversampling more informative subjects who have worse goodness-of-fit based on a preliminary model using only phase I covariates as predictors. In our numerical studies, BGOF was slightly less efficient than BD only for binary phase I variables in unrealistic situations where all covariates were uncorrelated, or when phase II covariates are not associated with the outcome and all phase I covariates are uncorrelated (data now shown). Intuitively, binary phase I covariates minimally contribute to assessment of model fit, and therefore are least capable of deriving efficiency gain from GOF sampling.

BGOF further improves GOF on estimating phase I covariates. Compared with BD, BGOF had notably higher efficiency for assessing phase I covariates. Instead of four sampling cells that we considered in the numerical studies, BD strata can also be based on composite categories created based on multiple phase I variables. BGOF remained more efficient in unreported numerical studies. The advantage of GOF and BGOF lies in the fact that it makes full use of all phase I covariate data to determine the sampling probability without coarsening. In practice, it may be necessary in BGOF to experiment with sampling probabilities to achieve pre-specified numbers of phase II cases and controls as we considered in the real data example. To realize the full efficiency advantage of BGOF, the sample size achieved in the GOF step needs to be balanced against the targeted phase II numbers. In one extreme scenario, where every subject is included in the GOF sample, BGOF will be the same as BD and lose efficiency advantage over BD. In another extreme scenario, where the GOF step directly selects the targeted number of phase II samples, BGOF will be the same as GOF and thus not able to further improve the efficiency for assessing the stratifying covariates of interest. We suggest that the sample size at the GOF step be on average two to three times larger than the targeted phase II sample size. It worked well in our numerical analysis, because those who had a poorer goodness-of-fit were still more likely to be sampled. BGOF is an attractive sampling method for reducing study cost while maintaining high statistical efficiency for both phase I and phase II covariates. Moreover, data from BGOF can be analyzed in use of existing packages in standard software, that allows a highly practical application of this new design.

We used similar pseudo-likelihood methods for analyzing data collected under BGOF and BD. Semiparametric maximum likelihood approaches are available for analyzing data from BD (Breslow and Holubkov, 1997; Lawless, Kalbfleisch, and Wild, 1999; Qin and Lawless, 1994; Scott and Wild, 1997). These methods require that phase I covariates be discrete. Consequently, these methods are not applicable to BGOF. We will compare the efficiency of BGOF when BD is analyzed using maximum likelihood methods. But neither pseudo-likelihood nor maximum likelihood methods had meaningfully improved efficiency for phase II covariates (e.g., Breslow and Chatterjee, 1999). This also points to an alternative method that may have improved efficiency for analyzing BGOF: a semiparametric maximum likelihood type of method that treats the subsample obtained at the GOF step as the cohort while appropriately accounting for the GOF sampling may have improved efficiency. We will investigate this method in future work.

BGOF relies on an existing preliminary model that relates the outcome variable with phase I covariates. In the absence of such external information, it is plausible to use internal phase I data to derive such a model. Extension of BGOF based on the resultant model needs to deal with the non-trivial complication that sample selection and subsequent statistical inference are based on data from the same individuals. We will consider this extension in our future work. BGOF assumes that phase I data are a cross-sectional sample. Two-phase case-control sampling, where phase I consists of a case-control sample, has also been widely studied in the literature (Breslow and Cain, 1988; Breslow and Holubkov, 1997). We are interested in extending BGOF to this setting as well.

CHAPTER 4

ADJUSTING FOR PARTICIPATION BIAS IN CASE-CONTROL GENETIC ASSOCIATION STUDIES WITH GENOTYPE DATA SUPPLEMENTED FROM FAMILY MEMBERS: AN EMPIRICAL LIKELIHOOD BASED ESTIMATING EQUATION APPROACH

4.1. Introduction

In case-control genetic association studies, it is common that many individuals' genotype data are missing for reasons related to the disease under study. This type of selection bias, referred to as the "participation bias", results in a non-ignorable missing structure and leads to biased inference if the genetic variants of interest are related to the disease and responsible for missingness. For highly lethal diseases, subjects, especially diseased cases, may be selectively excluded from the study due to mortality related to the disease or not be able to participate in genotype data collection because of their advanced disease status. Moreover, in pregnancy studies, chromosomal abnormalities are responsible for a great proportion of early pregnancy loss and thus conditions that can only be examined after birth will be missing (Chard, 1991). Participation bias will be induced if the pregnancy loss is related to the condition under study (Liew et al., 2015). Participation bias has also been reported in genetic association studies of age-related macular degeneration (Chiu et al., 2011) , ovarian cancer (Lacour et al., 2011), coronary heart disease (Williams, Pendyala, and Superko, 2011), and other cardiovascular diseases (Anderson et al., 2011; Falcone et al., 2013; Horsfall, Nazareth, and Petersen, 2012).

Participation bias is not restricted to scenarios where individuals are severely diseased. For example, it widely exists in various electronic health record based studies and often leads to non-ignorable missing structure (Haneuse et al., 2016). Moreover, healthy controls may cause participation bias because they are less motivated to contribute genetic information to a study. We are facing this issue in the Two Sister Study (O'Brien et al., 2016), a family-based case-control genetic association study of breast cancer where genetic information on the single-nucleotide polymorphisms (SNPs) near the gene TOX3 on chromosome 16 were missing for 40% of the 924 controls. Standard statistical methods for solving missing data issues, such as multiple imputation (Allison, 2000; Rubin, 1996) and inverse-probability weighting (Robins, Rotnitzky, and Zhao, 1994), are not

applicable when the missing at random (MAR) assumption does not hold: MAR requires that the missingness does not depend on the missing genetic information. However, the driving force behind missingness remains unknown and thus an investigation on the missing mechanism and an appropriate statistical method to adjust for potential participation bias is required.

Despite the widespread of participation bias in genetic association studies and the critical issues in statistical inference caused by participation bias, not many options regarding the sampling designs or statistical methods to resolve this problem are available, probably because participation bias is usually difficult to estimate or adjust for (Aschengrau and Seage, 2013; Haneuse and Chen, 2011; Haneuse et al., 2016). Chen, Weinberg, and Chen (2016) developed a family-supplemented design (FSD) that adjusts for participation bias in use of a maximum likelihood approach where they substitute first-degree family members' genotype data for deceased individuals. They proposed a valid estimator for the association parameter and showed greatly increased statistical power. In this study, we aim to develop a method to analyze data collected from the family supplemented design, which allows adjustments for covariates and interactive effects. The novelty of the proposed method lies in the unique advantage of genetic information that transmits along family members. Thus, family's genotype, such as parents' or spouse and children's, are highly informative proxy and can be utilized to infer an individual's missing genotype. It is challenging because of possible correlations between genotype and covariates besides the non-ignorable missing structure. We apply a logistic regression model to relate missingness with genotype and covariates, and use the expectation of the corresponding logistic regression score function conditional on all the observed data, including family's genotype data, as the estimating equation for the missingness odds ratio parameters (OR). We develop an empirical likelihood for the genetic association parameters and weight the empirical likelihood among individuals with complete data by the inverse of their probabilities of genotype data availability as the estimating equation for the association parameters. We estimate the nuisance parameters, i.e., the covariate distribution conditional on genotype, nonparametrically using data of controls inversely-weighted by their probabilities that complete data have been collected. We will obtain the estimators for association and missingness by jointly solving these estimating equations. We term this new method the family-supplemented weighted empirical likelihood method (FS-WEL).

The rest of this paper is organized as below. In section 4.2, we developed the estimation proce-

dure and the large sample inference for FS-WEL. In section 4.3, we evaluated the finite sample performance of FS-WEL via simulation studies across various scenarios. In section 4.4, we further evaluated the proposed method by applying it to a young-onset breast cancer genetic association study. We make final remarks in Section 4.5.

4.2. Methods

4.2.1. Framework and Notation

We use a logistic regression model to describe the relationship between the binary phenotype variable Y with genotype G and a vector of covariates X,

$$\text{logit}P(Y = 1|X, G) = \beta_0 + f_\beta(X, G), \tag{4.1}$$

with $Y = 1$ denoting a case and $Y = 0$ a control. We will refer to this model as the "association model". The function $f_\beta(G, X)$ is a pre-specified log odds function of genotype G and covariate X. In a log-additive model for G,

$$f_\beta(X, G) = \beta_1 X + \beta_2 G,$$

where G is coded as the minor allele count, 0, 1, or 2, for a di-allelic SNP A/a. Let θ denote the minor allele frequency (MAF). Under assumptions of Hardy-Weinberg equilibrium (HWE), random mating and Mendelian inheritance, the frequencies of G is determined by θ as

$$P_\theta(g) = \theta^g(1 - \theta)^{2-g} \cdot \{1 + I(G = 1)\}, \tag{4.2}$$

where $I(\cdot)$ is the indicator function. Define $\boldsymbol{\beta} = (\beta_1, \beta_2)$ and $\boldsymbol{\eta} = (\beta_1, \beta_2, \theta)$. Data for (Y, X) were collected for N_1 cases and N_0 controls, and let N denote the total number of subjects $N_0 + N_1$. The genotype data G were made available for a subset of n_1 ($n_1 < N_1$) cases and n_0 ($n_0 < N_0$) controls. We refer to the subset of data with both X and G available as the "complete observations", and denote the total number $n_1 + n_0$ as n. Let J denote the total number of unique values taken by covariate X in the data.

Let R denote whether a subject has genotype data available or not ($R = 1$: yes; $R = 0$: no). We

use a logistic regression model to describe the missing mechanism,

$$\text{logit}P_\alpha(R=1|Y,X,G) = \alpha_0 + \alpha_1 Y + \alpha_2 X + \alpha_3 G + \alpha_4 YX + \alpha_5 YG. \tag{4.3}$$

Henceforth we refer to model (4.3) as the "missingness model". Let $\boldsymbol{\alpha}$ denote all the parameters $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)^T$ in the missingness model. This model recognizes that the missingness depends on genotype itself. This non-ignorable missingness structure renders the imputation-based methods for estimating association parameters infeasible, and fitting the association model restricted to complete observations will also result in biased estimates for OR parameters $\boldsymbol{\beta}$. Methods based on inverse probability weighting are standard for consistent estimation. However, G is not available for individuals whose outcome R is 0, which makes it infeasible to estimate the probability of missingness by standard logistic regression analyses. Hence, we aim to develop a novel method that allows one to apply weighting technique under the non-ignorable missingness. Genes transmit within families, so that genotype data from family members can be exploited to help infer an individual's missing genotype. This very fact motivated a family-based supplementary design (Chen, Weinberg, and Chen, 2016), where genotype data from family members are collected to help infer missing genotypes. Suppose that genotype data of a subject's spouse, denoted as G^s and, of a child, denoted by G^c, can be made available for every subject who do not have data on G. Let G^f denote the collection of all available familial genotype data (G^s, G^c). The key step is to incorporate G^f into the estimating equation for the missingness model in terms of the conditional probability $P_\theta(G|G^f)$, a function of MAF, and to compensate for the missingness of G by inferring its distribution based on G^f. Let G_i denote the ith subject's genotype, G_i^f the ith subject's family's genotype, and X_i the ith subject's covariate and $i = 1, 2, ...N$. Define $\delta_{xg} = P(X = x|G = g, Y = 0)$ and $\boldsymbol{\delta}_{xg} = (\delta_{xg}; x = x_1, x_2, ..., x_J$ and $g = 0, 1, 2)$.

4.2.2. Empirical Likelihood of the Association Model

Had genotype data been obtained for all N subjects in the case-control sample, the likelihood could have been explicitly written out as a function of the OR parameter $\boldsymbol{\beta}$ in the association model, minor allele frequency θ, and the nuisance conditional probability $\boldsymbol{\delta}_{xg}$, i.e.,

$$L(\boldsymbol{\eta}, \boldsymbol{\delta}_{xg}) = \prod_{i=1}^{N_0} P(x_i, g_i|Y=0) \prod_{i=N_0+1}^{N} P(x_i, g_i|Y=1)$$

Using a result from Satten and Kupper (1993), the joint distribution of X and G for cases is related to that for controls as

$$P(x,g|Y=1)=\frac{e^{f(x,g)}P(x,g|Y=0)}{\sum_x\sum_g e^{f(x,g)}P(x,g|Y=0)}.$$

Then the above likelihood can be written as

$$\prod_{i=1}^{N_0}P(g_i|Y=0)P(x_i|g_i,Y=0)\prod_{i=N_0+1}^{N}\frac{e^{f_\beta(x_i,g_i)}P(g_i|Y=0)P(x_i|g_i,Y=0)}{\sum_x\sum_g e^{f_\beta(x,g)}P(g|Y=0)P(x|g,Y=0)}.$$

Note that the intercept parameter β_0 falls out of the likelihood function in this formulation. Therefore, we have avoided estimating probability $P(Y|x,g)$, which is not estimable using a case-control sample, in the likelihood function. We approximate the distribution of genotype in controls by that in the underlying population, i.e.,

$$P(g|Y=0)\approxeq P_\theta(g).$$

We obtain the empirical likelihood by replacing $P(X|G,Y=0)$ with the point mass δ_{xg},

$$\prod_{i=1}^{N}P_\theta(g_i)\delta_{x_ig_i}\left\{\frac{e^{f_\beta(x_i,g_i)}}{\sum_x\sum_g e^{f_\beta(x,g)}P_\theta(g)\delta_{xg}}\right\}^{y_i}.$$

Here, instead of treating $P(g|Y=0)$, the genotype frequencies, as probabilities for a multinomial variable, we model it as a parametric function of the minor allele frequency θ as governed by population genetic theory. This modeling is not for the purpose of improving statistical efficiency. But rather, we model it to be consistent with a later step in our approach where the conditional probability $P_\theta(g|g^f)$ needs to be estimated as a function of θ as required for inferring an individual's missing genotype. This point will be described in section 4.2.4, equation 4.7.

Maximizing the above empirical likelihood jointly with respect to all parameters, (β_1,β_2), θ, and $\boldsymbol{\delta}_{xg}$, leads to a more efficient estimator of (β_1,β_2) than that from a standard logistic regression analysis due to the parametric modeling of $P(G|Y=0)$. In standard logistic regression analysis where $P(G|Y=0)$ is treated as multinomial probabilities, the closed-form profile likelihood function for (β_1,β_2), obtained by maximizing the above empirical likelihood with respect to $\boldsymbol{\delta}_{xg}$ with (β_1,β_2) kept fixed, can greatly facilitate computation. We found that it is infeasible to derive such closed-form

profile-likelihood when $P(G|Y=0)$ is modeled parametrically. Therefore, we consider the following procedure to reduce computational cost at the expense of losing some efficiency due to ignoring information contained in cases.

For each combination of (x, g), we estimate the empirical distribution δ_{xg} nonparametrically by the ratio of corresponding cell counts only among controls, i.e.,

$$\hat{\delta}_{xg} = \frac{\sum_{i=1}^{N_0} I(X_i = x, G_i = g)}{\sum_{i=1}^{N_0} I(G_i = g)}, \tag{4.4}$$

where $g = 0, 1, 2$ and $x = x_1, x_2, ..., x_{J-1}$, and

$$\hat{\delta}_{(x_J)g} = 1 - \sum_{x=1}^{x_{J-1}} \hat{\delta}_{xg}.$$

Define $\hat{\boldsymbol{\delta}}_{xg} = (\hat{\delta}_{xg}; x = x_1, x_2, ..., x_J$ and $g = 0, 1, 2)$.

4.2.3. Inverse-Probability-Weighted Empirical Likelihood

The non-ignorable missingness structure in the case-control sample brings new challenges in estimation. First, the nuisance parameter involves the genotype variable G and we are restricted to controls with genotype data available for estimating $\boldsymbol{\delta}_{xg}$. Because G is associated with its missingness, in order to construct an unbiased estimator for $\boldsymbol{\delta}_{xg}$, we modify $\hat{\delta}_{xg}$ in equation 4.4 by weighting complete controls by the inverse of the probability of genotype availability:

$$\hat{\delta}_{xg}(\alpha) = \frac{\sum_{i=1}^{N_0} \frac{I(R_i=1, X_i=x, G_i=g)}{P_\alpha(R=1|Y=0, X=x_i, G=g_i)}}{\sum_{i=1}^{N_0} \frac{I(R_i=1, G_i=g)}{P_\alpha(R=1|Y=0, X=x_i, G=g_i)}}, \tag{4.5}$$

where $g = 0, 1, 2$ and $x = x_1, x_2, ..., x_{J-1}$. Second, because genotype is not missing completely at random (MCAR), naively applying the empirical likelihood method only in use of the complete observations will lead to biased estimates. In order to form an unbiased estimating equation for the association model, we weight a complete observation's empirical score function $U_i(\boldsymbol{\eta}, \hat{\boldsymbol{\delta}}_{xg}(\boldsymbol{\alpha}))$ by

the inverse of the probability that one's genotype is available

$$\sum_{i=1}^{N} \frac{R_i}{\pi_i(\boldsymbol{\alpha})} U_i(\boldsymbol{\eta}, \hat{\boldsymbol{\delta}}_{xg}(\boldsymbol{\alpha})) = \mathbf{0}, \tag{4.6}$$

where $\pi_i(\boldsymbol{\alpha}) = P_\alpha(R = 1|y_i, x_i, g_i)$ and

$$U_i(\boldsymbol{\eta}, \hat{\boldsymbol{\delta}}_{xg}(\boldsymbol{\alpha})) = \frac{d \log L_i(\boldsymbol{\eta}, \boldsymbol{\delta}_{xg})}{d\boldsymbol{\eta}} = \begin{bmatrix} y_i\{x_i - \frac{\sum_x \sum_g TT \cdot x}{\sum_x \sum_g TT}\} \\ y_i\{g_i - \frac{\sum_x \sum_g TT \cdot g}{\sum_x \sum_g TT}\} \\ \left(\frac{g_i}{\theta} - \frac{2-g_i}{1-\theta}\right) - y_i \left\{ \frac{\sum_x \sum_g TT \cdot (\frac{g}{\theta} - \frac{2-g}{1-\theta})}{\sum_x \sum_g TT} \right\} \end{bmatrix},$$

where $TT = \hat{\delta}_{xg}(\boldsymbol{\alpha}) \theta^g (1-\theta)^{2-g} \{1 + I(G = 1)\} e^{\beta_1 x + \beta_2 g}$.

4.2.4. Estimating Equation for the Missingness Model

The major challenge resulted from the non-ignorable missingness structure is for estimating the OR parameter $\boldsymbol{\alpha}$ in the missingness model, which is required when calculating weights in estimating equations 4.5 and 4.6. Although the case-control sample can be treated prospectively with respect to estimation of $\boldsymbol{\alpha}$ since the outcome Y is only a covariate in the missingness model, one cannot directly fit a logistic regression model defined by equation 4.3 to estimate $\boldsymbol{\alpha}$ because G is not available for every individual with $R = 0$. Therefore, we propose an estimating equation that sums up the expectation of each observation's score function $S_i(\boldsymbol{\alpha})$ of the missingness model conditional on one's observed data $\boldsymbol{ob}_i = (R_i = 0, Y_i, X_i, G_i^f)$ or $(R_i = 1, Y_i, X_i, G_i)$, i.e.,

$$\begin{aligned} U^m(\boldsymbol{\alpha}, \boldsymbol{\eta}, \hat{\boldsymbol{\delta}}_{xg}(\boldsymbol{\alpha})) &\equiv \sum_{i=1}^{N} E\big(S_i(\boldsymbol{\alpha})|\boldsymbol{ob}_i\big) \\ &= \sum_{i=1}^{N} \Big\{ R_i S_i(\boldsymbol{\alpha}) + (1 - R_i) E\big(S_i(\boldsymbol{\alpha})|\boldsymbol{ob}_i\big) \Big\} = 0, \end{aligned}$$

where $S_i(\boldsymbol{\alpha})$ is in the form of a standard score function for a logistic regression model, i.e.,

$$S_i(\boldsymbol{\alpha}) = \boldsymbol{d}_i \big[r_i - P_\alpha(R = 1|y_i, x_i, g_i) \big],$$

and $\boldsymbol{d}_i = (1, y_i, x_i, g_i, y_i x_i, y_i g_i)^T$. It is an unbiased estimating equation because its expectation reduces to the expectation of a regular score function $S(\boldsymbol{\alpha})$ and thus equals 0. Moreover, by condi-

tioning on the fully observed data, familial genotype data G^f is introduced into the estimation procedure. Then we derive the estimating equation for the missingness model as follows. First, similar to the estimating equation for the association model, we relate the joint distribution of (X, G) among cases to that among controls (Satten and Kupper, 1993) in order to avoid estimating $P(Y|x, g)$ under retrospective sampling. Second, we model the distribution of genotype as a parametric function of θ, which renders an expression of the estimating equation in terms of $P_\theta(g|g^f)$ feasible. Finally, we explicitly wrote out the estimating equation for the missingness model as

$$\sum_{i=1}^{N} \Bigg\{ R_i \boldsymbol{d}_i [1 - \pi_i(\boldsymbol{\alpha})] - (1 - R_i) \sum_g \frac{\boldsymbol{d}_i P_\alpha(R=1|y_i, x_i, g) P_\alpha(R=0|y_i, x_i, g) e^{y f_\beta(x_i, g)} \hat{\delta}_{x_i g}(\boldsymbol{\alpha}) P_\theta(g_i^f) P_\theta(g|g_i^f)}{\sum_g P_\alpha(R=0|y_i, x_i, g) e^{y f_\beta(x_i, g)} \hat{\delta}_{x_i g}(\boldsymbol{\alpha}) P_\theta(g_i^f) P_\theta(g|g_i^f)} \Bigg\} = \mathbf{0}. \tag{4.7}$$

This expression reformulated the challenging task of estimating $S_i(\boldsymbol{\alpha})$ for an individual without genotype data into the key step of estimating $P_\theta(g|g^f)$, which informs a missing genotype using familial genetic information. Consistent with the estimating equation for the association model, nuisance parameter $\boldsymbol{\delta}_{xg}$ are estimated nonparametrically only among controls by equation 4.5. Finally, taking expectation with respect to the missing variable G integrates it out of the estimating equation. Detailed derivation under the rare disease assumption is shown in APPENDIX A. Note that conditional on outcome Y, covariate X, and genotype G, the probability of genotype availability does not depend on family member's genotype G^f, i.e. $P(r_i|y_i, x_i, g, g_i^f) = P(r_i|y_i, x_i, g)$. We referred to Supplementary Table 1 of Chen, Weinberg, and Chen (2016) for the joint distribution of (g, g^f). We obtain consistent point estimates of $\boldsymbol{\eta}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\delta}_{xg}$ by jointly solving unbiased estimating equations 4.5, 4.6, and 4.7.

4.2.5. Asymptotic Properties

Following standard Z-estimation theory, we derived the asymptotic properties of the FS-WEL method. We define $A_i = \frac{R_i}{\pi_i(\boldsymbol{\alpha})} U_i(\boldsymbol{\eta}, \boldsymbol{\delta}_{xg})$ and $B_i = U_i^s(\boldsymbol{\alpha}, \boldsymbol{\delta}_{xg}, \boldsymbol{\eta})$. We re-express $\hat{\delta}_{xg}(\boldsymbol{\alpha})$ as

$$\hat{\delta}_{xg}(\boldsymbol{\alpha}) - \delta_{xg} = \frac{1}{N_0} \sum_{i=1}^{N_0} \left\{ \frac{I(R_i = 1, X_i = x, G_i = g)}{C_i} - \delta_{xg} \right\} =: \frac{1}{N_0} \sum_{i=1}^{N_0} f_i^{xg}$$

where $C_i = P(R = 1|Y = 0, x_i, g_i)P(G = g|Y = 0)$. Let f_i denote the vector of $\{f_i^{xg}, x = 1, 2, ..., J-1$ and $g = 0, 1, 2\}$. Then we define the following terms

$$
\begin{aligned}
c_1 &= E_{\{y,x,g\}}\left\{\frac{\partial U_i(\boldsymbol{\eta}, \boldsymbol{\delta}_{xg})}{\partial \boldsymbol{\eta}}\right\}, \\
c_2 &= E_{\{y,x,g\}}\left\{\frac{\partial U_i(\boldsymbol{\eta}, \boldsymbol{\delta}_{xg})}{\partial \boldsymbol{\delta}_{xg}}\right\}, \\
c_3 &= E_{\{y,x,g\}}\left\{\frac{\partial U_i(\boldsymbol{\eta}, \boldsymbol{\delta}_{xg})}{\partial \boldsymbol{\delta}_{xg}}\frac{\partial \hat{\boldsymbol{\delta}}_{xg}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right\} - E_{\{y,x,g\}}\left\{U(\boldsymbol{\eta}, \boldsymbol{\delta}_{xg})\frac{\partial \mathsf{log}\pi_i(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right\}, \\
d_1 &= E_{\{r,y,x,g\}}\left\{\frac{\partial U_i^s(\boldsymbol{\alpha}, \boldsymbol{\delta}_{xg}, \boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\right\}, \\
d_2 &= E_{\{r,y,x,g\}}\left\{\frac{\partial U_i^s(\boldsymbol{\alpha}, \boldsymbol{\delta}_{xg}, \boldsymbol{\eta})}{\partial \boldsymbol{\delta}_{xg}}\right\}, \\
d_3 &= E_{\{r,y,x,g\}}\left\{\frac{\partial U_i^s(\boldsymbol{\alpha}, \boldsymbol{\delta}_{xg}, \boldsymbol{\eta})}{\partial \boldsymbol{\alpha}}\right\} + E_{\{r,y,x,g\}}\left\{\frac{\partial U_i^s(\boldsymbol{\alpha}, \boldsymbol{\delta}_{xg}, \boldsymbol{\eta})}{\partial \boldsymbol{\delta}_{xg}}\frac{\partial \hat{\boldsymbol{\delta}}_{xg}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right\}.
\end{aligned}
$$

Expectations $\mathsf{E}_{\{y,x,g\}}(\cdot)$ and $\mathsf{E}_{\{r,y,x,g\}}(\cdot)$ are taken with respect to joint distributions $P(Y, X, G)$ and $P(R, Y, X, G, G^f)$ in the sample, respectively. We showed in APPENDIX B that $(\hat{\boldsymbol{\eta}}^T, \hat{\boldsymbol{\alpha}}^T)^T$ is consistent and asymptotically normally distributed:

$$
\sqrt{N}\begin{bmatrix}\hat{\boldsymbol{\eta}} - \boldsymbol{\eta} \\ \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\end{bmatrix} \xrightarrow{D} N\left(\begin{bmatrix}\mathbf{0} \\ \mathbf{0}\end{bmatrix}, M^{-1}V(M^{-1})^T\right).
$$

In the covariance matrix, $M = \begin{bmatrix} c_1 & c_3 \\ d_1 & d_3 \end{bmatrix}$ and V is defined as $\begin{bmatrix} V_{11} & V_{12} \\ V_{12}^T & V_{22} \end{bmatrix}$, where

$$
\begin{aligned}
V_{11} &= \mathsf{V}(A_i) + p_0^{-1}\mathsf{V}_{y_0}(c_2 f_i) + \mathsf{Cov}_{y_0}(A_i, c_2 f_i) + \mathsf{Cov}_{y_0}(c_2 f_i, A_i), \\
V_{22} &= \mathsf{V}(B_i) + p_0^{-1}\mathsf{V}_{y_0}(d_2 f_i) + \mathsf{Cov}_{y_0}(B_i, d_2 f_i) + \mathsf{Cov}_{y_0}(d_2 f_i, B_i), \\
V_{12} &= \mathsf{Cov}(A_i, B_i) + \mathsf{Cov}_{y_0}(c_2 f_i, B_i) + \mathsf{Cov}_{y_0}(A_i, d_2 f_i) + p_0^{-1}\mathsf{Cov}_{y_0}(c_2 f_i, d_2 f_i),
\end{aligned}
$$

and V_{y_0} and Cov_{y_0} are variance and covariance taken with respect to $P(X, G|Y = 0)$, and V and Cov are taken with respect to $P(X, G)$.

4.3. Simulation Study

In this section, we conducted simulation studies to quantify bias in estimates of parameters in both the association and missingness model using the FS-WEL method and to evaluate the finite sample performance of FS-WEL under various scenarios.

4.3.1. Parameters

We considered a cohort of 10^6 individuals. First, we generated genotype G for each individual in the cohort based on the minor allele frequency $\theta = 0.2$ and $\theta = 0.5$ under the assumption of HWE, random mating and the Mendelian inheritance. That is, generate G taking on values 0, 1, and 2 from a multinomial distribution of probability $((1-\theta)^2, 2\theta(1-\theta), \theta^2)$. Then we generated genotype G^s for every individual's spouse in the same way as for generating G, and generated genotype G^c for their children from a multinomial distribution of probability presented in table 4.1 under the same three assumptions stated above. Second, we generated a binary variable X conditional on genotype G from a Bernoulli distribution of probability $P(X|G)$, i.e., $P(X = 1|G = 0) = 0.3$, $P(X = 1|G = 1) = 0.5$, and $P(X = 1|G = 2) = 0.35$. Then we generated a binary variable Y based on the logistic regression model

$$\text{logit}P(Y = 1) = \beta_0 + \beta_1 X + \beta_2 G, \tag{4.8}$$

where $\beta_1 = \log(1.2)$ and β_2 took on values $\log(1.2)$ and $\log(1.5)$ to simulate relatively weak and strong association between outcome Y and genotype G. The intercept β_0 was chosen to determine the outcome prevalence of 3% and 10%.

Then we generated R, the missingness status of genotype from model 4.3, where $\alpha_1 = \log(0.6)$, $\alpha_2 = \log(1.2)$, and α_3 takes on $\log(1.2)$ and $\log(1.5)$ to simulate relatively weak and strong association between missingness and genotype. We considered no interaction between Y and X, or Y and G ($\alpha_4 = \alpha_5 = 0$) and differential association of missingness and X, and missingness and G between cases and controls ($\alpha_4 = \alpha_5 = \log(1.5)$). The intercept α_0 was selected to determine 80% and 60% genotype availability. Last, we randomly selected 2000 cases and 2000 controls from the cohort to form a case-control sample. Naive estimates of β_1 and β_2 were obtained from fitting logistic regression model 4.8 restricted to complete observations under the assumption of missing

Table 4.1: Distribution of children's genotypes (G^c) conditional on parents' genotypes (G and G^s) under the assumption of Hardy-Weinberg equilibrium, random mating and the Mendelian inheritance.

G	G^s	G^c		
		0	1	2
0	0	1	0	0
0	1	0.5	0.5	0
0	2	0	1	0
1	0	0.5	0.5	0
1	1	0.25	0.5	0.25
1	2	0	0.5	0.5
2	0	0	1	0
2	1	0	0.5	0.5
2	2	0	0	1

completely at random. The naive estimate of θ was calculated as

$$\tilde{\theta} = \frac{2n_{02} + n_{01}}{2n_0}, \tag{4.9}$$

where n_{01} and n_{02} were numbers of individuals who have one and two copies of the minor allele among n_0 controls who have genotype available. We repeated the above steps 1000 times.

4.3.2. Results

Table 4.2 presents the point estimates, the mean asymptotic and empirical standard errors, and the coverage probability for the log odds ratio parameter $\boldsymbol{\beta}$ in the association model and MAF using the FS-WEL method. Point estimates are all close to true parameter values, two types of standard errors are close to each other and the coverage probability are around the nominal level of 0.95 across all simulated scenarios. Estimation results of the missingness model are presented in supplementary table 1.

Table 4.3 shows the mean bias in and the mean squared error (MSE) of log odds ratio estimates in the association model using FS-WEL and the naive method across 1000 iterations. Differences between the estimates and the true parameter values are within 2% of true parameter values in OR estimate $\hat{\beta}_1$, 4% in $\hat{\beta}_2$, and 1% in $\hat{\theta}$ in use of FS-WPL. In contrast, the naive method over-estimates β_1 by $47\% - 95\%$ and β_2 by $18\% - 82\%$ when the association of missingness and covariates is differential between cases and controls, and by $8\% - 15\%$ and $8\% - 24\%$, respectively, when the

association is non-differential in the missingness model. The naive method over-estimates $\hat{\theta}$ by $2\% - 12\%$. Moreover, MSE in the naive method is $1.3 - 8.1$ times greater for $\hat{\beta}_1$, $1.3 - 6.4$ times greater for $\hat{\beta}_2$, and $1.1 - 10.4$ times greater for $\hat{\theta}$ than MSE in FS-WEL. The FS-WEL method is able to correct for non-ignorable missing structure across all simulation settings, even when 40% of genotype data are missing, and MSE slightly increases when missingness increases from 20% to 40%. Naively ignoring the non-ignorable missing structure, however, leads to great bias in OR estimates, and the bias and MSE become even more prominent when the effect of covariates and genotype on missingness further differentiates between cases and control and when missing data proportion increases. We further investigated the performance of FS-WEL in scenarios with a higher MAF (0.5) and a larger prevalence (0.1). FS-WEL still presents a superior capacity of correcting for bias in OR estimates compared with the naive method and similar results are shown in supplementary table $2 - 7$. We also evaluated FS-WEL when the underlying missing mechanism is missing completely at random, and the proposed method produces consistent OR estimates and slightly improves efficiency for the OR estimators compared with the naive method (data now shown).

4.4. Real Data Example

In this section, we investigated a dataset derived from the ongoing Two Sister Study (https://sisterstudy.niehs.nih.gov/English/twosisterstudy.htm), a family-based study on young-onset breast cancer (under age 50). We drew part of the data from this transmission-based study to construct a case-control sample in order to investigate the association between young-onset breast cancer (defined by Y with $Y = 1$ denoting a case and $Y = 0$ a control) and genetic risk factors. The original analysis (O'Brien et al., 2016) estimated relative risks by applying a log-linear model on case-parent data that compared cases in a matched manner to the non-transmitted alleles carried by their parents. In this study, we aim to directly compare cases to unrelated controls and to estimate odds ratios of developing young onset breast cancer with respect to genetic and other risk factors.

The constructed dataset consists of 521 women who had been diagnosed with breast cancer before age 50 and had one or more first-degree family members previously diagnosed with breast cancer (often the mother), and 924 controls who had never been diagnosed with breast cancer, but had

an affected sister. Thus all cases and all controls had a first-degree family history of breast cancer. When multiple controls were available from the same family, we randomly selected one of them. Thus the study subjects used for this analysis are unrelated, but some with missing genotypes had first-degree relatives available to provide proxy genotype information.

Age at first full-term pregnancy (X_1) and age at menarche (X_2) are two standard risk predictors for breast cancer (Chen et al., 2006; Gail et al., 1989). In this case-control sample, 229 controls (25%) had their first full-term pregnancy at age 24 or younger, 294 (32%) had it between age 24 and 30, 160 (17%) had it after age 30, and 241 (26%) were nulliparous. Among cases, 110 (21%), 143 (28%), 110 (21%) and 158 (30%) were in the four categories of age at first full-term pregnancy, respectively. Age at menarche was 12 or before for 386 controls (42%), between 12 and 14 for 295 controls (32%), and 14 or later for 243 controls (26%). Among cases, there are $238(46\%)$, $171(33\%)$ and $112(21\%)$ in the three categories of age at menarche, respectively. In general, a greater proportion of controls than cases have a younger age at first full-term pregnancy and an older age at menarche. Genetic information on SNPs rs8050542, rs8046979, rs4784220, rs1420533 and rs43143 near the gene TOX3 on chromosome 16 were collected from all cases and 351 controls. Among 573 controls whose genetic information was missing, both parents' genotype data were collected for 328 controls and one parent's genotype is available for the remaining 245 controls.

The driving forces related to missingness of genotype data (G) is unknown and thus whether a missing at random assumption holds is uncertain; hence, a non-ignorable missing structure among controls (denoted by R with $R = 1$ indicating genotype available and $R = 0$ not available), i.e.,

$$\begin{aligned}\text{logit}P(R = 1|Y = 0, x_1, x_2, g) =&\alpha_0 + \alpha_{11}I(24 < x_1 \leq 30 \text{ or nulliparous}) + \alpha_{12}I(x_1 > 30)+\\ &\alpha_{21}I(12 < x_2 < 14) + \alpha_{22}I(x_2 \leq 12) + \alpha_3 g,\end{aligned}$$

is the most flexible and appropriate. For each of the five SNPs, we applied the FS-WPL method and investigated whether missingness depended on genotype at that locus and used parents' genetic information to infer an individual's missing genotype. We considered the association model

$$\begin{aligned}\text{logit}P(Y = 1|x_1, x_2, g) =&\beta_0 + \beta_{11}I(24 < x_1 \leq 30 \text{ or nulliparous}) + \beta_{12}I(x_1 > 30)+\\ &\beta_{21}I(12 < x_2 < 14) + \beta_{22}I(x_2 \leq 12) + \beta_3 g.\end{aligned}$$

In contrast, the naive estimates were obtained from fitting the association model above restricted to complete observations and MAF was estimated using complete controls based on equation 4.9. We bootstrapped this dataset 1000 times and repeated both methods. We noticed that the naive method tended to overestimate OR parameters for covariates in the association model; therefore, covariates that were not statistically significantly associated with the risk of breast cancer in a preliminary study using the naive model, i.e., number of pregnancies and the interactive effect of age at menarche and age at first full-term pregnancy, were not included in this analysis for the purpose of model parsimony.

Table 4.4 presents the estimated log OR parameters in the association model of the Two Sister Study using FS-WEL and the naive method. Difference in the estimated log ORs between FS-WEL and the naive method ranges from $1\% - 16\%$ for β_{11}, $2\% - 11\%$ for β_{12}, $1\% - 27\%$ for β_{13}, $28\% - 58\%$ for β_{21}, $16\% - 28\%$ for β_{22} and $4\% - 32\%$ for β_3. The estimates of θ are very close in use of the two methods ($1\% - 3\%$). Missingness of genotype data in controls does not depend on the SNPs under study (*P*-value is $0.667 - 0.976$), but only on covariate age at first full-term pregnancy i.e., *P*-value < 0.05 for most of the SNPs and is at the marginal level 0.051 for rs43143 (table 4.4). Therefore, this missing structure is not non-ignorable and the missing at random assumption holds in this study, that explains the consistency of MAF estimates in use of the two methods. Because the naive method assumes that genotype is missing completely at random while covariate age at first full-term pregnancy, in fact, is responsible for the missingness, a bias in estimated log OR for age at first full-term pregnancy in the association model is expected and biases in log OR estimates for other variables might be caused by the correlation between age at first full-term pregnancy and other variables. We further investigated the Two Sister Study under a missing at random assumption, and the results are consistent with those obtained in FS-WEL (data not shown).

Among women without young-onset breast cancer in the Two Sister Study, there is a greater probability of genotype data availability for those who gave birth to their first child after age 30, and one possible explanation is that those women might be of a higher education level and better social-economic status and more likely to consent to genetic information collection. The odds of developing young-onset breast cancer for healthy women with breast cancer family history who gave birth to their first child after age 30 are $80\% - 90\%$ greater than those whose first child was born before age 24, and the odds for those whose age at menarche is before 12 tend to be $40\% - 50\%$ greater

than those who had menarche after age 14. None of the five SNPs are found to be significant risk factors for young-onset breast cancer after adjusting for multiple comparisons, which is consistent with the results in the original Two Sister Study (O'Brien et al., 2016).

4.5. Discussion

Participation bias is a common issue in biomedical studies, which is difficult to explore or adjust for, especially when the responsible variables cannot be fully collected, that induces a non-ignorable missing structure. Imprudently assuming an MAR or MCAR mechanism would result in biased inference in association analysis. Diseased and healthy individuals may cause participation bias for different reasons, such as death, disease severity, and lack of motivation for participation. Our proposed family-supplemented weighted empirical likelihood method is able to correct for participation bias in case-control genetic association studies under a very general framework where both cases and controls can encounter genetic information missingness for reasons related to genotype, many other covariates, and their interactive effects. The FS-WEL method is especially effective at bias correction compared with the naive method that simply restricts analysis to observations with complete data when a relatively large amount of missingness of genotype data exists, association of missingness and covariates differentiates between cases and controls, and the disease under study is relatively strongly associated with genotype.

Of course, if special scenarios where missingness only exists among controls such as in the Two Sister Study, or among cases (unreported simulation studies), FS-WEL can be easily modified by restricting missingness model to controls or cases, and in the latter scenario, the estimating procedure can be further simplified because the nuisance parameter, the conditional probability of covariates on genotype, can be directly estimated among complete controls (equation 4.4) without the weighting step. More generally, the driving forces behind missingness are usually unknown in substantive studies based on electronic health record data or various secondary analyses derived from original clinical trials. FS-WEL provides a method to investigate whether an MCAR or MAR assumption holds, and is capable of adjusting for potential non-ignorable missingness. Furthermore, in unreported simulation studies, even when the underlying missing mechanism is MCAR, the FS-WEL method leads to consistent estimates of association parameters and improves the statistical efficiency compared with the naive method, probably due to a larger effective sample size

and parametric modeling of the genotype distribution in the estimating procedure of FS-WEL.

The FS-WEL method can effectively account for participation bias even when missingness is substantial. This great power results from the fact that genotype data from both parents or spouse and child provide the most informative proxy for an individual's missing genotype, which explains that the quantity and variability of bias in OR estimates under FS-WEL are consistently small but dramatically increase under the naive method as a greater amount of missingness happens. In reality, however, the most informative proxy might not be collected for every individual with missing genetic information, such as in the Two Sister Study, more than 40% of those without complete data were only supplemented with one parent's generic information. The FS-WEL method is able to flexibly handle different types of family's genetic supplements from one parent, children only, one parent and a sibling, etc. For example, integrating over the possible genotypes of the missing parent allows one to infer an individual's missing genotype based only on one parent's supplement. As family's supplements become less informative, consistent OR estimates with greater variability (MSE) might be expected as the proportion of missingness augments. A spouse's genetic information alone is non-informative and thus is not under consideration.

When applying FS-WEL in finite sample analyses, one may encounter the following numerical issue and we provide suggestions on how to overcome it. Theoretically, cases and controls share the same domain for covariates; however, in a real data analysis, distinction between cases and controls can be observed. One may face this issue when estimating the nuisance parameter, i.e., the conditional probability of covariates on genotype, because the nonparametric estimation is performed among controls. For distinctive covariate values that are only observed in cases, we suggest approximating its probability conditional genotype in use of the adjacent value, for example, equally splitting the probability of the adjacent value conditional on genotype into its own probability and probability for the unique value in cases. The issue of different empirical domains between cases and controls is less worrisome when one analyzes discrete covariates, such as the standard risk factors, age at first full-term pregnancy and age at menarche in the Two Sister Study, but more common with continuous covariates, especially in relatively small samples. Nevertheless, in unreported simulation studies, we considered continuous covariates where we treated each observed covariate value as a "discrete" variable level, and FS-WEL still successfully corrected for participation bias under nuisance parameter approximation.

The most effective strategy to deal with missing data is to further collect missing values themselves, which, however, is often not feasible, especially when missingness is caused by death and severe diseases. In genetic association studies, the missing genotype enjoys a valuable privilege of highly informative proxy from first-degree family members, and the proxy information is usually conveniently available in family-based studies. If not, researchers should take full advantage of this privilege and make extra effort to supplement the original dataset with family's genetic information to the possible extent. Beyond genetic association studies, multiphase designs with additional data collection of possible responsible variables for participation bias at the design level, and post hoc or parallel surveys with ongoing clinical trials on missing values themselves (at least a subset) and possible reasons for data incompleteness have been suggested (Haneuse and Chen, 2011; Haneuse et al., 2016). The main purpose is to exploit the distribution of incomplete variables, and to investigate the driving mechanisms for missingness, which provides the foundation of evaluation on study results and conclusions and application of novel approaches to correcting for potential biased inference.

Table 4.2: The estimated log OR of covariate X (β_1) and genotype G (β_2) in the association model and the estimated MAF (θ) using the family-supplemented weighted empirical likelihood method. The prevalence is 0.03, genotype availability is 0.8 and 0.6, the true value of β_1 is 0.182, of β_2 is 0.182 and 0.405, and MAF is 0.2. The true values of $(\alpha_3, \alpha_4, \alpha_5)$ in the three missingness models are: weak $= (0.182, 0.405, 0.405)$, strong $= (0.405, 0.405, 0.405)$, and No interaction (NI) $= (0.405, 0, 0)$. The mean asymptotic standard error ("asym"), empirical standard error ("emp"), and coverage probability ("coverage") of $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\theta}$ were calculated based on 1000 simulations.

$P(R=1)$	e^{α_2}	missing	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\theta}$	
			estimate (asy/emp)	coverage	estimate (asy/emp)	coverage	estimate (asy/emp)	coverage
0.8	1.2	weak	0.180 (0.071/0.069)	0.961	0.179 (0.059/0.062)	0.928	0.200 (0.007/0.007)	0.946
		strong	0.180 (0.071/0.068)	0.964	0.179 (0.058/0.057)	0.957	0.200 (0.007/0.007)	0.952
		NI	0.185 (0.071/0.068)	0.960	0.182 (0.060/0.061)	0.947	0.199 (0.007/0.007)	0.951
	1.5	weak	0.183 (0.073/0.070)	0.957	0.404 (0.057/0.058)	0.945	0.198 (0.007/0.007)	0.947
		strong	0.180 (0.073/0.069)	0.972	0.406 (0.056/0.060)	0.934	0.198 (0.007/0.007)	0.946
		NI	0.181 (0.073/0.067)	0.969	0.403 (0.058/0.062)	0.935	0.198 (0.007/0.007)	0.942
0.6	1.2	weak	0.179 (0.075/0.068)	0.967	0.179 (0.065/0.069)	0.936	0.200 (0.008/0.008)	0.950
		strong	0.180 (0.075/0.068)	0.969	0.179 (0.064/0.063)	0.961	0.200 (0.008/0.008)	0.953
		NI	0.185 (0.075/0.068)	0.963	0.175 (0.067/0.071)	0.937	0.200 (0.008/0.008)	0.952
	1.5	weak	0.183 (0.078/0.070)	0.967	0.406 (0.063/0.064)	0.946	0.198 (0.008/0.008)	0.937
		strong	0.179 (0.077/0.070)	0.975	0.406 (0.062/0.066)	0.927	0.198 (0.008/0.008)	0.949
		NI	0.182 (0.077/0.068)	0.976	0.403 (0.064/0.069)	0.934	0.198 (0.008/0.008)	0.942

Table 4.3: The mean bias in and the mean square error (MSE) of estimated log OR of covariate X (β_1) and genotype G (β_2) in the association model and the estimated MAF (θ) using the family-supplemented weighted empirical likelihood method (FS-WEL) and the naive method based on1000 simulations. The prevalence is 0.03, genotype availability is 0.8 and 0.6, the true value of β_1 is 0.182, of β_2 is 0.182 and 0.405, and MAF is 0.2. The true values of $(\alpha_3, \alpha_4, \alpha_5)$ in the three missingness models are: weak $= (0.182, 0.405, 0.405)$, strong $= (0.405, 0.405, 0.405)$, and No interaction (NI) $= (0.405, 0, 0)$. In each of the 12 settings, true values and estimates of coefficients β_1, β_2 and θ are presented in this order in the magnitude of 10^{-3}.

Model				Naive Method		FS-WPL Method	
$P(R=1)$	e^{β_2}	Missing	True	Bias[a]	MSE	Bias	MSE
0.8	1.2	weak	182	90.972	14.317	-2.659	4.775
			182	71.651	8.983	-3.543	3.864
			200	4.514	0.072	-0.471	0.050
		strong	182	86.399	13.224	-2.635	4.589
			182	69.068	8.442	-3.390	3.297
			200	11.074	0.177	-0.313	0.050
		NI	182	18.522	6.495	3.093	4.635
			182	36.302	4.920	0.003	3.775
			200	10.124	0.152	-0.870	0.049
	1.5	weak	182	91.532	14.456	0.374	4.855
			405	75.328	9.064	-0.969	3.333
			200	2.959	0.055	-1.883	0.051
		strong	182	88.835	14.071	-2.513	4.827
			405	73.935	9.050	0.813	3.584
			200	8.889	0.130	-2.051	0.052
		NI	182	14.350	5.949	-0.935	4.479
			405	34.134	4.948	-2.355	3.814
			200	9.087	0.135	-1.862	0.053
0.6	1.2	weak	182	169.575	37.065	-3.808	4.588
			182	149.636	27.067	-2.874	4.700
			200	9.246	0.156	-0.456	0.061
		strong	182	161.017	33.907	-2.497	4.621
			182	137.279	23.557	-3.223	3.974
			200	23.201	0.614	-0.309	0.059
		NI	182	27.175	9.309	3.040	4.691
			182	44.436	7.287	-7.215	5.104
			200	22.795	0.594	-0.300	0.062
	1.5	weak	182	172.247	37.875	0.532	4.938
			405	147.784	26.323	0.544	4.138
			200	8.001	0.127	-2.036	0.063
		strong	182	160.941	33.795	-3.009	4.906
			405	139.789	24.250	0.697	4.364
			200	21.001	0.512	-2.006	0.064
		NI	182	19.376	8.870	-0.635	4.562
			405	46.625	7.218	-2.549	4.723
			200	21.003	0.513	-1.772	0.063

[a] Bias is defined as the mean estimate minus the true value

Table 4.4: Estimated log OR parameters in the association model of the Two Sister Study using the family-supplemented weighted empirical likelihood method (FS-WEL) and the naive method. Mean asymptotic ("asy") and bootstrap ("bts") standard errors are calculated on 1000 bootstrap iterations. *p*-value is resulted from a Wald test in use of the bootstrap standard error in both association and missingness models.

	Association Model				Missingness Model
Model	FS-WPL (asy/bts)	*P*-value	Naive (bts)	*P*-value	*P*-value
rs8051542	0.261 (0.096/0.112)	0.020	0.251 (0.105)	0.016	0.667
X_1 (24, 30]	0.167 (0.204/0.203)	0.411	0.167 (0.188)	0.374	0.116
> 30	0.605 (0.230/0.233)	0.009	0.666 (0.222)	0.003	0.040
NL[a]	0.236 (0.306/0.204)	0.247	0.237 (0.190)	0.211	0.380
X_2 (12, 14)	0.264 (0.201/0.203)	0.194	0.396 (0.185)	0.032	0.226
≤ 12	0.360 (0.227/0.193)	0.063	0.445 (0.177)	0.012	0.275
MAF	0.401 (0.011/0.012)		0.390 (0.020)		
rs8046979	-0.196 (0.114/0.117)	0.092	-0.134 (0.097)	0.165	0.777
X_1 (24, 30]	0.175 (0.202/0.210)	0.404	0.190 (0.195)	0.330	0.126
> 30	0.612 (0.228/0.235)	0.009	0.672 (0.228)	0.003	0.047
NL	0.183 (0.293/0.211)	0.385	0.215 (0.188)	0.251	0.405
X_2 (12, 14)	0.303 (0.195/0.197)	0.125	0.390 (0.185)	0.035	0.205
≤ 12	0.357 (0.215/0.185)	0.054	0.451 (0.173)	0.009	0.245
MAF	0.498 (0.011/0.012)		0.486 (0.019)		
rs4784220	0.223 (0.093/0.101)	0.027	0.252 (0.110)	0.023	0.894
X_1 (24, 30]	0.207 (0.209/0.203)	0.308	0.175 (0.185)	0.345	0.111
> 30	0.667 (0.233/0.226)	0.003	0.677 (0.223)	0.002	0.041
NL	0.257 (0.314/0.206)	0.212	0.219 (0.191)	0.253	0.351
X_2 (12, 14)	0.234 (0.204/0.206)	0.256	0.370 (0.194)	0.057	0.207
≤ 12	0.368 (0.228/0.191)	0.054	0.452 (0.184)	0.014	0.238
MAF	0.439 (0.008/0.009)		0.439 (0.016)		
rs1420533	-0.209 (0.114/0.114)	0.066	-0.148 (0.095)	0.119	0.976
X_1 (24, 30]	0.167 (0.202/0.210)	0.427	0.189 (0.197)	0.336	0.122
> 30	0.605 (0.229/0.237)	0.011	0.673 (0.230)	0.003	0.047
NL	0.167 (0.293/0.206)	0.416	0.213 (0.186)	0.254	0.388
X_2 (12, 14)	0.302 (0.195/0.193)	0.118	0.387 (0.183)	0.034	0.185
≤ 12	0.352 (0.215/0.186)	0.058	0.451 (0.174)	0.009	0.242
MAF	0.496 (0.011/0.011)		0.489 (0.019)		
rs43143	-0.091 (0.099/0.106)	0.392	-0.082 (0.093)	0.380	0.925
X_1 (24, 30]	0.218 (0.202/0.208)	0.295	0.194 (0.194)	0.317	0.120
> 30	0.633 (0.229/0.242)	0.009	0.675 (0.231)	0.004	0.051
NL	0.251 (0.297/0.205)	0.221	0.240 (0.192)	0.211	0.375
X_2 (12, 14)	0.289 (0.198/0.205)	0.159	0.392 (0.186)	0.035	0.216
≤ 12	0.384 (0.224/0.193)	0.047	0.445 (0.177)	0.012	0.254
MAF	0.440 (0.011/0.012)		0.446 (0.020)		

[a] NL: nulliparous

CHAPTER 5

CONCLUSION

In this dissertation, we focus on the two main issues under the outcome-dependent sampling framework, efficiency and bias. We have developed two novel outcome-dependent sampling designs that take advantage of the biased sampling scheme to improve statistical efficiency for estimating association parameters, especially with respect to the incomplete phase II covariates. We also have developed a new estimating equation approach to correcting for participation bias, a type of outcome-dependent selection bias, in genetic association studies.

Our work provides a new perspective to define informative subjects in efficient sampling designs. Existing efficient designs usually consider subjects in rare groups of an outcome or covariates at phase I informative. We found that subjects who have worse goodness-of-fit based on an external model relating the outcome only to phase I covariates are more informative with respect to phase II covariates, and our proposed sampling designs greatly improve efficiency for estimating phase II covariates by oversampling these informative subjects. Instead of relying on an existing preliminary model that relates the outcome variable with phase I covariates, using internal phase I data to derive such a model is considered for future work, which needs to deal with the nontrivial complication that sample selection and subsequent statistical inference are based on data of the same individuals. The proposed designs assume that phase I data are a cross-sectional sample, and we are interested in extending them to Two-phase case-control sampling, where phase I consists of a case-control sample as well. Last, extending the estimating equation approach to adjusting for participation bias beyond genetic association studies, where the informative proxy from family members for missing data are not necessarily available, is of interest in future work.

APPENDIX A

THEORETICAL DERIVATION FOR CHAPTER 3

The proof of consistency and asymptotic normality of the pseudo-likelihood estimator $\hat{\boldsymbol{\beta}}$ can follow similar steps for Propositions 1 and 2 in Breslow and Cain (1988). The proof requires accommodation of the BGOF sampling, which is different from the prospective (Scott and Wild, 1997) and retrospective (Breslow and Cain, 1988) two-phase designs. The $M \equiv \sum_{i=1}^{N} R_i$ "phase I" subjects in the GOF subsample is a biased random sample selected using pre-specified outcome-dependent sampling probabilities. A key to the proof, which accommodates this sampling feature, is that the pseudo-model for the GOF sub-sample, $P(Y = 1|\boldsymbol{w}; R = 1)$ is a genuine probabilitiy function so that E$(Y|\boldsymbol{w}, R = 1) = P(Y = 1|\boldsymbol{w}; R = 1)$ and var$(Y|\boldsymbol{w}; R = 1) = P(Y = 1|\boldsymbol{w}; R = 1)P(Y = 0|\boldsymbol{w}; R = 1)$. Let $\hat{\psi}_{1l}$ denote M_{1l}/N, $\hat{\psi}_{0l}$ denote M_{0l}/N, $\hat{\psi}_l$ denote $(\hat{\psi}_{0l}, \hat{\psi}_{1l})$, and $\hat{\boldsymbol{\psi}}$ denote the vector of all these $2L$ sampling probabilities, $\hat{\boldsymbol{\psi}} = \{\hat{\psi}^l : \ l = 1, \dots L\}$. For subject i in the BGOF sample who belongs to sampling stratum l, $i = 1, \dots, m_{1l} + m_{0l}$, we re-write $\mu(\boldsymbol{w}_i; \boldsymbol{\beta}, \pi_{0l}, \pi_{1l})$ as $\mu(\boldsymbol{w}_i; \boldsymbol{\beta}, \hat{\psi}_l)$:

$$\mu(\boldsymbol{w}_i; \boldsymbol{\beta}, \hat{\boldsymbol{\psi}}_l) \quad = \quad \frac{\exp\left\{\boldsymbol{w}_i\boldsymbol{\beta} + o(\boldsymbol{x}_i) + \log m_{1l}/m_{0l} - \log \hat{\psi}_{1l}/\hat{\psi}_{0l}\right\}}{1 + \exp\left\{\boldsymbol{w}_i\boldsymbol{\beta} + o(\boldsymbol{x}_i) + \log m_{1l}/m_{0l} - \log \hat{\psi}_{1l}/\hat{\psi}_{0l}\right\}}.$$

Let $p^*(Y|\boldsymbol{w}; \boldsymbol{\beta}, \hat{\boldsymbol{\psi}}_l)$ denote the distribution of Y conditional on $\boldsymbol{w}$ in the BGOF subsample. The pseudo-likeihood estimator is obtained by maximizing the following pseudo log-likelihood function over $\boldsymbol{\beta}$:

$$\begin{aligned} l(\boldsymbol{\beta}; \hat{\boldsymbol{\psi}}) \quad &\equiv \quad m^{-1} \sum_{y=1}^{1} \sum_{l=1}^{L} \sum_{i=1}^{m_{yl}} \log p^*(y|\boldsymbol{w}_i; \boldsymbol{\beta}, \hat{\boldsymbol{\psi}}_l) \\ &= \quad m^{-1} \sum_{y=0}^{1} \sum_{l=1}^{L} \sum_{i=1}^{m_{yl}} \left[y \log \mu(\boldsymbol{w}_i; \boldsymbol{\beta}, \hat{\boldsymbol{\psi}}_l) + (1 - y) \log\{1 - \mu(\boldsymbol{w}_i; \boldsymbol{\beta}, \hat{\boldsymbol{\psi}}_l)\} \right]. \end{aligned}$$

Let $p^*_{yl}(\boldsymbol{w})$ denote the distribution of $\boldsymbol{W}$ in the (y, l) stratum, and E^*_{yl} denote the expectation taken with respect to $p^*_{yl}(\boldsymbol{w})$. Further let E^*_l denote the expectation taken with respect to the distribution of $\boldsymbol{W}$ within stratum l. Assume that m_{yl}/m converges to γ_{yl} as $m \to \infty$, and m/N converges to a

constant $\theta > 0$ as $N \to \infty$. To prove consistency of $\hat{\boldsymbol{\beta}}$, clearly $l(\boldsymbol{\beta}; \boldsymbol{\psi})$ converges to

$$l^*(\boldsymbol{\beta}; \boldsymbol{\psi}) = \sum_{y=0}^{1} \sum_{l=1}^{L} \gamma_{yl} \mathrm{E}^*_{yl}\{\log p^*(y|\boldsymbol{w})\}$$

almost surely in a neighborhoodd of $\boldsymbol{\beta}$. Since components of $\boldsymbol{w}$ are bounded, the convergence is uniform by Theorem 2 in Jennrich (1969). Then $l(\boldsymbol{\beta}; \hat{\boldsymbol{\psi}})$ converges to $l^*(\boldsymbol{\beta}; \boldsymbol{\psi})$ uniformly since $\hat{\psi}_{yl} \stackrel{a.s}{\to} \psi_{yl}$ as $N \to \infty$, $y = 0, 1$, $l = 1, \ldots, L$. Further, $l^*(\boldsymbol{\beta}; \boldsymbol{\psi})$ is uniquely maximized at the true value of $\boldsymbol{\beta}$ assuming identifiability of $\boldsymbol{\beta}$ since it is an expected likelihood function. Consistency of $\hat{\boldsymbol{\beta}}$ follows from Lemma (1) in Manski and McFadden (1981).

To derive the large sample distribution of $\hat{\boldsymbol{\beta}}$, we perform Taylor's series expansion at the true value $\boldsymbol{\beta}$ on the pseudo-likelihood score function as follows. Below we use μ_i^l as the shorthand for $\mu(\boldsymbol{w}_i; \boldsymbol{\beta}, \boldsymbol{\psi}_l)$.

$$\begin{aligned}
0 &= U(\hat{\boldsymbol{\beta}}; \hat{\boldsymbol{\psi}}) \equiv \frac{\partial l(\boldsymbol{\beta}; \hat{\boldsymbol{\psi}})}{\partial \boldsymbol{\beta}} = m^{-1} \sum_{y=0}^{1} \sum_{l=1}^{L} \sum_{i=1}^{m_{yl}} \boldsymbol{w}_i^T \left\{ y - \mu(\boldsymbol{w}_i; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}_l) \right\} \\
&\approx m^{-1} \sum_{y=0}^{1} \sum_{l=1}^{L} \sum_{i=1}^{m_{yl}} \boldsymbol{w}_i^T \left\{ y - \mu(\boldsymbol{w}_i; \boldsymbol{\beta}, \boldsymbol{\psi}_l) \right\} \\
&+ m^{-1} \sum_{y=0}^{1} \sum_{l=1}^{L} \sum_{i=1}^{m_{yl}} \boldsymbol{w}_i^T \boldsymbol{w}_i \mu_i^l \left(1 - \mu_i^l\right) (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&- m^{-1} \sum_{l=1}^{L} \sum_{y=0}^{1} \sum_{i=1}^{m_{yl}} \boldsymbol{w}_i^T \mu_i^l \left(1 - \mu_i^l\right) \left(\log \hat{\psi}_{1l} / \hat{\psi}_{0l} - \log \psi_{1l} / \psi_{0l} \right).
\end{aligned}$$

It is straightforward to show that

$$\begin{aligned}
&m^{-1} \sum_{y=0}^{1} \sum_{l=1}^{L} \sum_{i=1}^{m_{yl}} \boldsymbol{w}_i^T \boldsymbol{w}_i \mu_i^l (1 - \mu_i^l) \stackrel{p}{\to} H \equiv \sum_{y=0}^{1} \sum_{l=1}^{L} \mathrm{E}^*_{yl}\{\boldsymbol{w}_i^T \boldsymbol{w}_i \mu_i^l (1 - \mu_i^l)\}, \\
&m^{-1} \sum_{y=0}^{1} \sum_{i=1}^{m_{yl}} \boldsymbol{w}_i^T \mu_i^l (1 - \mu_i^l)\} \stackrel{p}{\to} \mathrm{E}^*_l\{\boldsymbol{w}_i^T \mu_i^l \left(1 - \mu_i^l\right)\}.
\end{aligned}$$

Re-arranging the terms above leads to

$$\begin{aligned}
\sqrt{m}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= -H^{-1} \left\{ m^{-1/2} \sum_{y=0}^{1} \sum_{l=1}^{L} \sum_{i=1}^{m_{yl}} \boldsymbol{w}_i^T \left(y - \mu_i^l\right) - \right. \\
&\left. \sqrt{m} \sum_{l=1}^{L} \mathrm{E}^*_l\{\boldsymbol{w}_i^T \mu_i^l \left(1 - \mu_i^l\right)\} \left(\log \hat{\psi}_{1l} / \hat{\psi}_{0l} - \log \psi_{1l} / \psi_{0l} \right) \right\}. \qquad \text{(A.1)}
\end{aligned}$$

Note that the first and second terms in the curly braces are independent, and that $\hat{\psi}_l$'s are independent. For the first term, because its expectation equals zero, it can be written as

$$m^{-1/2} \sum_{y=0}^{1} \sum_{l=1}^{L} \sum_{i=1}^{m_{yl}} \left[\boldsymbol{w}_i^T \left(y - \mu_i^l \right) - \mathrm{E}_{yl}^* \left\{ \boldsymbol{w}_i^T \left(y - \mu_i^l \right) \right\} \right].$$

An application of Lindeberg-Feller Central limit theorem leads to a result that was proved in Breslow and Cain (1988):

$$\frac{1}{\sqrt{m}} \sum_{y=0}^{1} \sum_{l=1}^{L} \sum_{i=1}^{m_{yl}} \boldsymbol{w}_i^T \left(y - \mu_i^l \right) \xrightarrow{D} N(0, V).$$

where

$$V = H - \sum_{y=0}^{1} \sum_{l=1}^{L} \gamma_{yl}^{-1} \mathrm{E}_l^* \left\{ \boldsymbol{w}^T \mu^l (1 - \mu^l) \right\} \mathrm{E}_l^* \left\{ \boldsymbol{w} \mu^l (1 - \mu^l) \right\}.$$

Using standard theory for the multinomial distribution, it can be shown that

$$\sqrt{m} \left(\log \hat{\psi}_{1l} / \hat{\psi}_{0l} - \log \psi_{1l} / \psi_{0l} \right) \xrightarrow{D} N(0, \psi_{0l}^{-1} + \psi_{1l}^{-1}).$$

Therefore, the second term in the curly brace in equation (A.1) converges in distribution to

$$N \left[0, \ \sum_{y=0}^{1} \sum_{l=1}^{L} \theta \psi_{yl}^{-1} E_l^* \left\{ \boldsymbol{w}^T \mu^l \left(1 - \mu^l \right) \right\} E_l^* \left\{ \boldsymbol{w} \mu^l \left(1 - \mu^l \right) \right\} \right].$$

Putting the above results together, we have the following results as stated in the Corollary:

$$\sqrt{N} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N \left\{ 0, H^{-1} (H - G) H^{-1} \right\},$$

where G is a matrix defined as

$$G \equiv \sum_{y=0}^{1} \sum_{l=1}^{L} \left(\gamma_{yl}^{-1} - \theta \psi_{yl}^{-1} \right) E_l^* \left\{ \boldsymbol{w}^T \mu^l \left(1 - \mu^l \right) \right\} E_l^* \left\{ \boldsymbol{w} \mu^l \left(1 - \mu^l \right) \right\}.$$

The asymptotic variance-covariance matrix $H^{-1}(H - G)H^{-1}$ can be consistently estimated empirically, with

$$\hat{H} = \frac{1}{m} \sum_{y=0}^{1} \sum_{l=1}^{L} \sum_{i=1}^{m_{yl}} \boldsymbol{w}_i^T \boldsymbol{w}_i \mu_i^l \left(1 - \mu_i^l \right),$$

and

$$\hat{G} = \frac{1}{m} \sum_{y=0}^{1} \sum_{l=1}^{L} \left[\left(m_{yl}^{-1} - M_{yl}^{-1} \right) \sum_{y=0}^{1} \sum_{i=0}^{m_{yl}} \left\{ \boldsymbol{w}_i^T \mu_i^l \left(1 - \mu_i^l \right) \right\} \sum_{y=0}^{1} \sum_{i=0}^{m_{yl}} \left\{ \boldsymbol{w}_i \mu_i^l \left(1 - \mu_i^l \right) \right\} \right].$$

APPENDIX B

THEORETICAL DERIVATION FOR CHAPTER 4

B.1. Derivation

In this section, we show the detailed derivation of the conditional expectation of the score function of the missingness model on observed data for a deceased individual, i.e., $R = 0$.

$$
\begin{aligned}
E\big(S_i(\boldsymbol{\alpha})|\boldsymbol{ob}_i\big) =& - E(\boldsymbol{d}_i P(R=1|y_i,x_i,g_i)|\boldsymbol{ob}_i) \\
=& -\sum_g \boldsymbol{d}_i P(R=1|y_i,x_i,g) P(G=g|\boldsymbol{ob}_i) \\
=& -\sum_g \frac{\boldsymbol{d}_i P(R=1|y_i,x_i,g) P(r_i|y_i,x_i,g,g_i^f) P(g|y_i,x_i,g_i^f)}{P(r_i|y_i,x_i,g_i^f)} \\
=& -\sum_g \frac{\boldsymbol{d}_i P(R=1|y_i,x_i,g) P(r_i|y_i,x_i,g,g_i^f) P(g,x_i,g_i^f|y_i)}{P(r_i|y_i,x_i,g_i^f)\sum_g P(g,x_i,g_i^f|y_i)} \\
=& -\sum_g \frac{\boldsymbol{d}_i P(R=1|y_i,x_i,g) P(r_i|y_i,x_i,g,g_i^f) P(g,x_i,g_i^f|y_i)}{\sum_g P(r_i,g|y_i,x_i,g_i^f)\sum_g P(g,x_i,g_i^f|y_i)} \\
=& -\sum_g \frac{\boldsymbol{d}_i P(R=1|y_i,x_i,g) P(r_i|y_i,x_i,g,g_i^f) P(g,x_i,g_i^f|y_i)}{\sum_g \{P(r_i|g,y_i,x_i,g_i^f) P(g|y_i,x_i,g_i^f)\}\sum_g P(g,x_i,g_i^f|y_i)} \\
=& -\sum_g \frac{\boldsymbol{d}_i P(R=1|y_i,x_i,g) P(r_i|y_i,x_i,g,g_i^f) P(g,x_i,g_i^f|y_i)}{\sum_g \{P(r_i|y_i,x_i,g,g_i^f)\frac{P(x_i,g,g_i^f|y_i)}{\sum_g P(x_i,g,g_i^f|y_i)}\}\sum_g P(x_i,g,g_i^f|y_i)} \\
=& -\sum_g \frac{\boldsymbol{d}_i P(R=1|y_i,x_i,g) P(r_i|y_i,x_i,g,g_i^f) P(g,x_i,g_i^f|y_i)}{\sum_g \{P(r_i|y_i,x_i,g,g_i^f) P(x_i,g,g_i^f|y_i)}
\end{aligned}
$$

Following the result of Satten and Kupper (1993) for a rare outcome to relate $P(X,G,G^f|Y=1)$ and $P(X,G,G^f|Y=0)$, i.e.,

$$
P(x,g,g^f|Y=1) = \frac{e^{f_\beta(x,g)} P(x,g,g^f|Y=0)}{\sum_x \sum_g e^{f_\beta(x,g)} P(x,g,g^f|Y=0)},
$$

the conditional expectation score function can be further derived as

$$
\begin{aligned}
&-E(\boldsymbol{d}_i P(R=1|y_i,x_i,g_i)|\boldsymbol{ob}_i) \\
\approx &-\sum_g \frac{\boldsymbol{d}_i P(R=1|y_i,x_i,g)P(r_i|y_i,x_i,g,g_i^f)e^{y_i f_\beta(x_i,g)}P(x_i,g,g_i^f|Y=0)}{\sum\limits_g P(r_i|y_i,x_i,g,g_i^f)e^{y_i f_\beta(x_i,g)}P(x_i,g,g_i^f|Y=0)} \\
= &-\sum_g \frac{\boldsymbol{d}_i P(R=1|y_i,x_i,g)P(R=0|y_i,x_i,g)e^{y_i f_\beta(x_i,g)}\delta_{x_i g}P(g,g_i^f|Y=0)}{\sum\limits_g P(R=0|y_i,x_i,g)e^{y_i f_\beta(x_i,g)}\delta_{x_i g}P(g,g_i^f|Y=0)} \\
\approx &-\sum_g \frac{\boldsymbol{d}_i P(R=1|y_i,x_i,g)P(R=0|y_i,x_i,g)e^{y_i f_\beta(x_i,g)}\delta_{x_i g}P_\theta(g,g_i^f)}{\sum\limits_g P(R=0|y_i,x_i,g)e^{y_i f_\beta(x_i,g)}\delta_{x_i g}P_\theta(g,g_i^f)}
\end{aligned}
$$

Note that under the assumption of a rare disease, $P(g,g_i^f|Y=0)=P_\theta(g,g_i^f)$. Conditional on the disease status Y, covariate X, and genotype information G, one's missingness does not depend on one's family member's genotype G^f, i.e. $P(r_i|y_i,x_i,g,g_i^f)=P(r_i|y_i,x_i,g)$. Conditional on genotype G, distribution of covariate X among controls does not depend on G^f, i.e., $P(x|g,g^f,Y=0)=P(x|g,Y=0)=\delta_{xg}$.

B.2. Asymptotic Properties

In this section, we derive the asymptotic property of the family-supplemented weighted empirical likelihood method. Let p_1 and p_0 denote the proportion of cases and controls in the sample. First we applied Taylor expansion to the score function of the association model around the true parameter $(\boldsymbol{\eta}^T,\boldsymbol{\delta}_{xg}^T,\boldsymbol{\alpha}^T)^T$, i.e.,

$$
\begin{aligned}
\mathbf{0}=&\sum_{i=1}^N \frac{R_i}{\pi_i(\hat{\boldsymbol{\alpha}})}U_i(\hat{\boldsymbol{\eta}},\hat{\boldsymbol{\delta}}_{xg}(\hat{\boldsymbol{\alpha}})) \\
\mathbf{0}\approx&\frac{1}{N}\sum_{i=1}^N \frac{R_i}{\pi_i(\boldsymbol{\alpha})}U_i(\boldsymbol{\eta},\boldsymbol{\delta}_{xg})+\frac{1}{N}\sum_{i=1}^N \frac{R_i}{\pi_i(\boldsymbol{\alpha})}\frac{\partial U_i(\boldsymbol{\eta},\boldsymbol{\delta}_{xg})}{\partial\boldsymbol{\eta}}(\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}) \\
&+\frac{1}{N}\sum_{i=1}^N \frac{R_i}{\pi_i(\boldsymbol{\alpha})}\frac{\partial U_i(\boldsymbol{\eta},\boldsymbol{\delta}_{xg})}{\partial\boldsymbol{\delta}_{xg}}(\hat{\boldsymbol{\delta}}_{xg}(\boldsymbol{\alpha})-\boldsymbol{\delta}_{xg}) \\
&+\frac{1}{N}\sum_{i=1}^N \left\{\frac{R_i}{\pi_i(\boldsymbol{\alpha})}\frac{\partial U_i(\boldsymbol{\eta},\boldsymbol{\delta}_{xg})}{\partial\boldsymbol{\delta}_{xg}}\frac{\partial\hat{\boldsymbol{\delta}}_{xg}(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}-U_i(\boldsymbol{\eta},\boldsymbol{\delta}_{xg})\frac{R_i}{\pi_i^2(\boldsymbol{\alpha})}\frac{\partial\pi_i(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}\right\}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})
\end{aligned}
\tag{B.1}
$$

In equation (B.1),

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\frac{R_i}{\pi_i(\boldsymbol{\alpha})}\frac{\partial U_i(\boldsymbol{\eta},\boldsymbol{\delta}_{xg})}{\partial\boldsymbol{\eta}} &= \frac{1}{N}\sum_{y}\sum_{i=1}^{N_y}\frac{R_i}{\pi_i(\boldsymbol{\alpha})}\left\{\frac{\partial U_i(\boldsymbol{\eta},\boldsymbol{\delta}_{xg})}{\partial\boldsymbol{\eta}}\right\}\\ &= \sum_{y}\left\{\frac{N_y}{N}\frac{1}{N_y}\sum_{i=1}^{N_y}\frac{R_i}{\pi_i(\boldsymbol{\alpha})}\left\{\frac{\partial U_i(\boldsymbol{\eta},\boldsymbol{\delta}_{xg})}{\partial\boldsymbol{\eta}}\right\}\right\}\\ &\xrightarrow{P}\sum_{y}p_yE_y\left\{\frac{\partial U_i(\boldsymbol{\eta},\boldsymbol{\delta}_{xg})}{\partial\boldsymbol{\eta}}\right\}\\ &= E_{\{y,x,g\}}\left\{\frac{\partial U_i(\boldsymbol{\eta},\boldsymbol{\delta}_{xg})}{\partial\boldsymbol{\eta}}\right\} =: c_1,\end{aligned}$$

where E_y is the expectation taken with respect to $P(R,X,G|Y)$ and $E_{\{y,x,g\}}$ is taken with respect to $P(Y,X,G)$ in the sample.

Similarly,

$$\frac{1}{N}\sum_{i=1}^{N}\frac{R_i}{\pi_i(\boldsymbol{\alpha})}\frac{\partial U_i(\boldsymbol{\eta},\boldsymbol{\delta}_{xg})}{\partial\boldsymbol{\delta}_{xg}}\xrightarrow{P}E_{\{y,x,g\}}\left\{\frac{\partial U_i(\boldsymbol{\eta},\boldsymbol{\delta}_{xg})}{\partial\boldsymbol{\delta}_{xg}}\right\} =: c_2$$

$$\frac{1}{N}\sum_{i=1}^{N}\frac{R_i}{\pi_i(\boldsymbol{\alpha})}\frac{\partial U_i(\boldsymbol{\eta},\boldsymbol{\delta}_{xg})}{\partial\boldsymbol{\delta}_{xg}}\frac{\partial\hat{\boldsymbol{\delta}}_{xg}(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}\xrightarrow{P}E_{\{y,x,g\}}\left\{\frac{\partial U_i(\boldsymbol{\eta},\boldsymbol{\delta}_{xg})}{\partial\boldsymbol{\delta}_{xg}}\frac{\partial\hat{\boldsymbol{\delta}}_{xg}(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}\right\} =: c_{31}$$

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}U(\boldsymbol{\eta},\boldsymbol{\delta}_{xg})\frac{R_i}{\pi_i^2(\boldsymbol{\alpha})}\frac{\partial\pi_i(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}} &\xrightarrow{P}E_{\{y,x,g\}}\left\{U(\boldsymbol{\eta},\boldsymbol{\delta}_{xg})\frac{1}{\pi_i(\boldsymbol{\alpha})}\frac{\partial\pi_i(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}\right\}\\ &= E_{\{y,x,g\}}\left\{U(\boldsymbol{\eta},\boldsymbol{\delta}_{xg})\frac{\partial\log\pi_i(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}\right\} =: c_{32}\end{aligned}$$

We denote $c_3 = c_{31} - c_{32}$ and therefore,

$$\mathbf{0} = \frac{1}{N}\sum_{i=1}^{N}\frac{R_i}{\pi_i(\boldsymbol{\alpha})}U_i(\boldsymbol{\eta},\boldsymbol{\delta}_{xg}) + c_1(\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}) + c_2(\hat{\boldsymbol{\delta}}_{xg}(\boldsymbol{\alpha})-\boldsymbol{\delta}_{xg}) + c_3(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}) \tag{B.2}$$

Then we performed Taylor expansion on the score function of the missingness model around the

true parameter $(\boldsymbol{\eta}^T, \boldsymbol{\delta}_{xg}^T, \boldsymbol{\alpha}^T)^T$, i.e.,

$$
\begin{aligned}
\mathbf{0} &= \sum_{i=1}^{N} U_i^s(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\delta}}_{xg}(\hat{\boldsymbol{\alpha}}), \hat{\boldsymbol{\eta}}) \\
\mathbf{0} &\approxeq \frac{1}{N}\sum_{i=1}^{N} U_i^s(\boldsymbol{\alpha}, \boldsymbol{\delta}_{xg}, \boldsymbol{\eta}) + \frac{1}{N}\sum_{i=1}^{N} \frac{\partial U_i^s(\boldsymbol{\alpha}, \boldsymbol{\delta}_{xg}, \boldsymbol{\eta})}{\partial \boldsymbol{\eta}}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \\
&+ \frac{1}{N}\sum_{i=1}^{N} \frac{\partial U_i^s(\boldsymbol{\alpha}, \boldsymbol{\delta}_{xg}, \boldsymbol{\eta})}{\partial \boldsymbol{\delta}_{xg}}(\hat{\boldsymbol{\delta}}_{xg}(\boldsymbol{\alpha}) - \boldsymbol{\delta}_{xg}) \\
&+ \frac{1}{N}\sum_{i=1}^{N} \left(\frac{\partial U_i^s(\boldsymbol{\alpha}, \boldsymbol{\delta}_{xg}, \boldsymbol{\eta})}{\partial \boldsymbol{\alpha}} + \frac{\partial U_i^s(\boldsymbol{\alpha}, \boldsymbol{\delta}_{xg}, \boldsymbol{\eta})}{\partial \boldsymbol{\delta}_{xg}} \frac{\partial \hat{\boldsymbol{\delta}}_{xg}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \right) (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})
\end{aligned} \tag{B.3}
$$

In equation (B.3),

$$
\frac{1}{N}\sum_{i=1}^{N} \frac{\partial U_i^s(\boldsymbol{\alpha}, \boldsymbol{\delta}_{xg}, \boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \xrightarrow{P} E\left\{ \frac{\partial U_i^s(\boldsymbol{\alpha}, \boldsymbol{\delta}_{xg}, \boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \right\} =: d_1,
$$

$$
\frac{1}{N}\sum_{i=1}^{N} \frac{\partial U_i^s(\boldsymbol{\alpha}, \boldsymbol{\delta}_{xg}, \boldsymbol{\eta})}{\partial \boldsymbol{\delta}_{xg}} \xrightarrow{P} E\left\{ \frac{\partial U_i^s(\boldsymbol{\alpha}, \boldsymbol{\delta}_{xg}, \boldsymbol{\eta})}{\partial \boldsymbol{\delta}_{xg}} \right\} =: d_2,
$$

$$
\frac{1}{N}\sum_{i=1}^{N} \frac{\partial U_i^s(\boldsymbol{\alpha}, \boldsymbol{\delta}_{xg}, \boldsymbol{\eta})}{\partial \boldsymbol{\alpha}} \xrightarrow{P} E\left\{ \frac{\partial U_i^s(\boldsymbol{\alpha}, \boldsymbol{\delta}_{xg}, \boldsymbol{\eta})}{\partial \boldsymbol{\alpha}} \right\} =: d_{31},
$$

$$
\frac{1}{N}\sum_{i=1}^{N} \frac{\partial U_i^s(\boldsymbol{\alpha}, \boldsymbol{\delta}_{xg}, \boldsymbol{\eta})}{\partial \boldsymbol{\delta}_{xg}} \frac{\partial \hat{\boldsymbol{\delta}}_{xg}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \xrightarrow{P} E\left\{ \frac{\partial U_i^s(\boldsymbol{\alpha}, \boldsymbol{\delta}_{xg}, \boldsymbol{\eta})}{\partial \boldsymbol{\delta}_{xg}} \frac{\partial \hat{\boldsymbol{\delta}}_{xg}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \right\} =: d_{32},
$$

where these expectations are taken with respect to the joint distribution $P(R, Y, X, G, G^f)$ in the sample. Let $d_3 = d_{31} + d_{32}$ and equation (B.3) can be written as below

$$
\mathbf{0} = \frac{1}{N}\sum_{i=1}^{N} U_i^s(\boldsymbol{\alpha}, \boldsymbol{\delta}_{xg}, \boldsymbol{\eta}) + d_1(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) + d_2(\hat{\boldsymbol{\delta}}_{xg}(\boldsymbol{\alpha}) - \boldsymbol{\delta}_{xg}) + d_3(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \tag{B.4}
$$

The nonparametric estimator of each element δ_{xg} in the vector $\boldsymbol{\delta}_{xg}$ is

$$\hat{\delta}_{xg}(\boldsymbol{\alpha}) = \frac{1}{N_0}\sum_{i=1}^{N_0} \frac{I(R_i=1, X_i=x, G_i=g)}{\pi_\alpha(1,0,x_i,g_i)\frac{1}{N_0}\sum_{i=1}^{N_0}\frac{I(R_i=1,G_i=g)}{\pi_\alpha(1,0,x_i,g_i)}},$$

where

$$\begin{aligned}\pi_\alpha(1,0,x_i,g_i)\frac{1}{N_0}\sum_{i=1}^{N_0}\frac{I(R_i=1,G_i=g)}{\pi_\alpha(1,0,x_i,g_i)} &\xrightarrow{P} \pi_\alpha(1,0,x_i,g_i)E\left\{I(G_i=g)|Y=0\right\} \\ &= \pi_\alpha(1,0,x_i,g_i)P(G=g|Y=0) =: C_i.\end{aligned}$$

Therefore,

$$\hat{\delta}_{xg}(\boldsymbol{\alpha}) - \delta_{xg}(\boldsymbol{\alpha}) = \frac{1}{N_0}\sum_{i=1}^{N_0}\left\{\frac{I(R_i=1,X_i=x,G_i=g)}{C_i} - \delta_{xg}\right\} =: \frac{1}{N_0}\sum_{i=1}^{N_0} f_i^{xg}$$

Let f_i denote the vector of $\{f_i^{xg}, x=1,2,...,J-1$ and $g=0,1,2\}$ and thus

$$\hat{\boldsymbol{\delta}}_{xg}(\boldsymbol{\alpha}) - \boldsymbol{\delta}_{xg} = \frac{1}{N_0}\sum_{i=1}^{N_0} f_i \tag{B.5}$$

Then we substitute the expression of $\hat{\boldsymbol{\delta}}_{xg}(\boldsymbol{\alpha}) - \boldsymbol{\delta}_{xg}$ in equation (B.5) into equation (B.2) and equation (B.4), i.e.,

$$M\begin{bmatrix}\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}\\ \hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}\end{bmatrix} + \begin{bmatrix}\frac{1}{N}\sum_{i=1}^{N}A_i + c_2\cdot\frac{1}{N_0}\sum_{i=1}^{N_0}f_i\\ \frac{1}{N}\sum_{i=1}^{N}B_i + d_2\cdot\frac{1}{N_0}\sum_{i=1}^{N_0}f_i\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix},$$

where

$$\begin{aligned}M &= \begin{bmatrix}c_1 & c_3\\ d_1 & d_3\end{bmatrix},\\ A_i &= \frac{R_i}{\pi_i(\boldsymbol{\alpha})}U_i(\boldsymbol{\eta},\boldsymbol{\delta}_{xg}),\\ B_i &= U_i^s(\boldsymbol{\alpha},\boldsymbol{\delta}_{xg},\boldsymbol{\eta}).\end{aligned}$$

The influence functions of $(\boldsymbol{\eta}^T, \boldsymbol{\alpha}^T)^T$ can be written out as

$$\sqrt{N}\begin{bmatrix}\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}\\ \hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}\end{bmatrix} = -M^{-1}\begin{bmatrix}\sum_y \frac{1}{\sqrt{N_y}}\sum_{i=1}^{N_y} p_y^{1/2}A_i + \frac{1}{\sqrt{N_0}}\sum_{i=1}^{N_0} p_0^{-1/2}c_2 f_i\\ \frac{1}{\sqrt{N}}\sum_{i=1}^{N} B_i + \frac{1}{\sqrt{N_0}}\sum_{i=1}^{N_0} p_0^{-1/2}d_2 f_i\end{bmatrix}.$$

Finally under regularity conditions, $(\hat{\boldsymbol{\eta}}^T, \hat{\boldsymbol{\alpha}}^T)^T$ is consistently and asymptotically normally distributed as

$$\sqrt{N}\begin{bmatrix}\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}\\ \hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}\end{bmatrix} \xrightarrow{D} N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, M^{-1}V(M^{-1})^T\right).$$

The covariance matrix V is defined

$$V = \begin{bmatrix}V_{11} & V_{12}\\ V_{12}^T & V_{22}\end{bmatrix},$$

where

$$\begin{aligned}
V_{11} &= \mathsf{V}(A_i) + p_0^{-1}\mathsf{V}_{y_0}(c_2 f_i) + \mathsf{Cov}_{y_0}(A_i, c_2 f_i) + \mathsf{Cov}_{y_0}(c_2 f_i, A_i),\\
V_{22} &= \mathsf{V}(B_i) + p_0^{-1}\mathsf{V}_{y_0}(d_2 f_i) + \mathsf{Cov}_{y_0}(B_i, d_2 f_i) + \mathsf{Cov}_{y_0}(d_2 f_i, B_i),\\
V_{12} &= \mathsf{Cov}(A_i, B_i) + \mathsf{Cov}_{y_0}(c_2 f_i, B_i) + \mathsf{Cov}_{y_0}(A_i, d_2 f_i) + p_0^{-1}\mathsf{Cov}_{y_0}(c_2 f_i, d_2 f_i),
\end{aligned}$$

and V_{y_0} and Cov_{y_0} are variance and covariance taken conditional on $Y = 0$, and V_{y_1} and Cov_{y_1} are taken conditional on $Y = 1$.

APPENDIX C

ADDITIONAL SIMULATION RESULTS FOR CHAPTER 2 AND 3

In this APPENDIX, we present additional simulation studies in Chapter 2 and 3. Results under a small cohort scenario of size 3000, are presented in table C.1 - C.6, and results under a big cohort scenario of size 2×10^4 are shown in table C.7 - C.14.

Table C.1: The estimated log OR of phase I covariate ($\hat{\beta}_1$) under balanced goodness-of-fit based design (BGOF), the goodness-of-fit based design (GOF), the case-control design (CC), and the balanced design (BD). The phase I cohort size was 3000, the prevalence was 0.05 and 0.10, the correlation parameter ρ for phase I variables was 0 and 0.3, and the true value of β_4 was 0.5, 0.7, and 0.9. The correlation between phase II variable X_4 and phase I variable X_1, X_2, and X_3 are 0.6, 0.5, and 0.3, respectively. X_1 was the stratifying variable in BGOF. The true value of β_1 is 0.5. The mean asymptotic standard error ("asym"), empirical standard error ("emp"), and coverage probability ("coverage") of $\hat{\beta}_4$ were calculated based on 1000 simulations.

$P(Y=1)$	ρ	β_4	BGOF (asym/emp)	coverage	GOF (asym/emp)	coverage	CC (asym)	BD (asym)
0.05	0	0.5	0.44 (0.56/0.55)	0.939	0.45 (0.60/0.61)	0.950	0.48 (0.65)	0.46 (0.60)
		0.7	0.48 (0.58/0.59)	0.948	0.45 (0.61/0.62)	0.950	0.52 (0.67)	0.48 (0.63)
		0.9	0.47 (0.60/0.60)	0.952	0.46 (0.64/0.67)	0.948	0.51 (0.69)	0.47 (0.66)
	0.3	0.5	0.47 (0.50/0.52)	0.955	0.48 (0.54/0.55)	0.962	0.49 (0.57)	0.47 (0.52)
		0.7	0.46 (0.52/0.50)	0.950	0.46 (0.56/0.56)	0.946	0.52 (0.59)	0.49 (0.55)
		0.9	0.46 (0.54/0.56)	0.942	0.45 (0.59/0.60)	0.949	0.49 (0.61)	0.46 (0.58)
0.10	0	0.5	0.47 (0.39/0.39)	0.942	0.47 (0.42/0.42)	0.954	0.49 (0.45)	0.48 (0.41)
		0.7	0.50 (0.40/0.40)	0.953	0.51 (0.43/0.46)	0.955	0.52 (0.46)	0.51 (0.43)
		0.9	0.47 (0.41/0.41)	0.947	0.48 (0.44/0.43)	0.952	0.49 (0.48)	0.50 (0.45)
	0.3	0.5	0.50 (0.35/0.35)	0.951	0.50 (0.37/0.39)	0.947	0.52 (0.39)	0.50 (0.36)
		0.7	0.50 (0.36/0.36)	0.943	0.50 (0.38/0.39)	0.948	0.51 (0.40)	0.51 (0.38)
		0.9	0.48 (0.37/0.38)	0.947	0.47 (0.40/0.41)	0.955	0.50 (0.42)	0.48 (0.40)

Table C.2: The estimated log OR of phase I covariate ($\hat{\beta}_2$) under balanced goodness-of-fit based design (BGOF), the goodness-of-fit based design (GOF), the case-control design (CC), and the balanced design (BD). The phase I cohort size was 3000, the prevalence was 0.05 and 0.10, the correlation parameter ρ for phase I variables was 0 and 0.3, and the true value of β_4 was 0.5, 0.7, and 0.9. The correlation between phase II variable X_4 and phase I variable X_1, X_2, and X_3 are 0.6, 0.5, and 0.3, respectively. X_1 was the stratifying variable in BGOF. The true value of β_2 is 0.6. The mean asymptotic standard error ("asym"), empirical standard error ("emp"), and coverage probability ("coverage") of $\hat{\beta}_4$ were calculated based on 1000 simulations.

$P(Y=1)$	ρ	β_4	BGOF (asym/emp)	coverage	GOF (asym/emp)	coverage	CC (asym)	BD (asym)
0.05	0.0	0.5	0.59 (0.16/0.17)	0.938	0.59 (0.16/0.16)	0.950	0.61 (0.18)	0.60 (0.18)
		0.7	0.60 (0.16/0.17)	0.946	0.59 (0.16/0.17)	0.949	0.62 (0.18)	0.62 (0.18)
		0.9	0.59 (0.17/0.17)	0.952	0.59 (0.17/0.17)	0.947	0.61 (0.19)	0.61 (0.19)
	0.3	0.5	0.60 (0.14/0.14)	0.955	0.60 (0.14/0.14)	0.962	0.61 (0.16)	0.61 (0.16)
		0.7	0.60 (0.15/0.15)	0.950	0.60 (0.15/0.15)	0.944	0.62 (0.16)	0.62 (0.16)
		0.9	0.60 (0.15/0.15)	0.940	0.60 (0.15/0.15)	0.949	0.61 (0.17)	0.62 (0.17)
0.10	0.0	0.5	0.60 (0.11/0.11)	0.942	0.60 (0.11/0.11)	0.954	0.61 (0.12)	0.61 (0.12)
		0.7	0.60 (0.11/0.11)	0.952	0.60 (0.12/0.11)	0.951	0.61 (0.13)	0.61 (0.13)
		0.9	0.59 (0.12/0.12)	0.947	0.60 (0.12/0.12)	0.951	0.61 (0.13)	0.61 (0.13)
	0.3	0.5	0.61 (0.10/0.10)	0.949	0.60 (0.10/0.11)	0.946	0.61 (0.11)	0.61 (0.11)
		0.7	0.60 (0.10/0.11)	0.942	0.60 (0.10/0.11)	0.948	0.61 (0.11)	0.61 (0.11)
		0.9	0.59 (0.11/0.11)	0.946	0.60 (0.11/0.11)	0.954	0.61 (0.12)	0.60 (0.11)

Table C.3: The estimated log OR of phase I covariate ($\hat{\beta}_3$) under balanced goodness-of-fit based design (BGOF), the goodness-of-fit based design (GOF), the case-control design (CC), and the balanced design (BD). The phase I cohort size was 3000, the prevalence was 0.05 and 0.10, the correlation parameter ρ for phase I variables was 0 and 0.3, and the true value of β_4 was 0.5, 0.7, and 0.9. The correlation between phase II variable X_4 and phase I variable X_1, X_2, and X_3 are 0.6, 0.5, and 0.3, respectively. X_1 was the stratifying variable in BGOF. The true value of β_3 is -0.7. The mean asymptotic standard error ("asym"), empirical standard error ("emp"), and coverage probability ("coverage") of $\hat{\beta}_4$ were calculated based on 1000 simulations.

$P(Y=1)$	ρ	β_4	BGOF (asym/emp)	coverage	GOF (asym/emp)	coverage	CC (asym)	BD (asym)
0.05	0.0	0.5	-0.71 (0.31/0.31)	0.955	-0.71 (0.31/0.31)	0.945	-0.71 (0.33)	-0.73 (0.32)
		0.7	-0.71 (0.30/0.30)	0.941	-0.71 (0.30/0.31)	0.942	-0.72 (0.33)	-0.71 (0.33)
		0.9	-0.71 (0.30/0.30)	0.955	-0.71 (0.30/0.30)	0.935	-0.71 (0.34)	-0.71 (0.33)
	0.3	0.5	-0.70 (0.28/0.28)	0.943	-0.71 (0.28/0.28)	0.952	-0.72 (0.31)	-0.71 (0.29)
		0.7	-0.70 (0.28/0.27)	0.951	-0.70 (0.28/0.27)	0.956	-0.72 (0.31)	-0.73 (0.30)
		0.9	-0.70 (0.28/0.29)	0.959	-0.70 (0.28/0.29)	0.954	-0.72 (0.32)	-0.72 (0.30)
0.10	0.0	0.5	-0.71 (0.21/0.22)	0.940	-0.71 (0.21/0.21)	0.941	-0.71 (0.22)	-0.70 (0.22)
		0.7	-0.69 (0.21/0.21)	0.957	-0.69 (0.21/0.21)	0.953	-0.69 (0.23)	-0.69 (0.22)
		0.9	-0.72 (0.21/0.21)	0.944	-0.72 (0.21/0.21)	0.950	-0.71 (0.24)	-0.71 (0.23)
	0.3	0.5	-0.71 (0.19/0.20)	0.959	-0.71 (0.20/0.20)	0.952	-0.72 (0.21)	-0.71 (0.21)
		0.7	-0.69 (0.19/0.20)	0.951	-0.70 (0.20/0.20)	0.947	-0.70 (0.22)	-0.70 (0.21)
		0.9	-0.70 (0.20/0.19)	0.954	-0.70 (0.20/0.20)	0.946	-0.70 (0.22)	-0.71 (0.21)

Table C.4: Asymptotic variance of $\hat{\boldsymbol{\beta}}$ under the balanced goodness-of-fit design (BGOF) and its efficiency relative to the goodness-of-fit based design (GOF), the balanced design (BD) and the case-control design (CC). The phase I cohort size was 3000. The prevalence was 0.05 and 0.10, the correlation parameter ρ for phase I variables was 0 and 0.3, and the true value of β_4 was 0.5, 0.7, and 0.9. The correlation between phase II variable X_4 and phase I variable X_1, X_2, and X_3 are 0.6, 0.5, and 0.3, respectively. X_3 was the stratifying variable in BGOF.

$P(Y=1)$	ρ	β_4	Var($\hat{\beta}_4$)	RE[a] of BGOF vs.			Var($\hat{\beta}_1$)	RE of BGOF vs.		
				GOF	BD	CC		GOF	BD	CC
0.05	0	0.5	0.042	1.00	1.22	1.23	0.36	1.00	1.15	1.16
		0.7	0.043	1.01	1.28	1.29	0.37	1.01	1.18	1.18
		0.9	0.046	1.01	1.34	1.35	0.40	1.01	1.19	1.20
	0.3	0.5	0.029	1.00	1.21	1.21	0.30	1.00	1.10	1.11
		0.7	0.030	1.01	1.24	1.25	0.32	1.00	1.10	1.11
		0.9	0.034	1.01	1.28	1.29	0.35	1.01	1.10	1.10
0.10	0	0.5	0.020	1.01	1.20	1.22	0.17	1.01	1.15	1.16
		0.7	0.021	1.01	1.24	1.25	0.18	1.01	1.17	1.17
		0.9	0.023	1.01	1.29	1.29	0.19	1.01	1.18	1.19
	0.3	0.5	0.014	1.00	1.18	1.18	0.14	1.00	1.10	1.11
		0.7	0.015	1.00	1.21	1.21	0.15	1.00	1.11	1.11
		0.9	0.017	1.01	1.24	1.24	0.16	1.00	1.11	1.11
			Var($\hat{\beta}_2$)	RE of BGOF vs.			Var($\hat{\beta}_3$)	RE of BGOF vs.		
				GOF	BD	CC		GOF	BD	CC
0.05	0	0.5	0.026	1.01	1.24	1.26	0.070	1.35	1.10	1.52
		0.7	0.026	1.01	1.30	1.31	0.068	1.33	1.19	1.62
		0.9	0.027	1.02	1.35	1.36	0.069	1.32	1.28	1.72
	0.3	0.5	0.021	1.00	1.24	1.24	0.058	1.37	1.10	1.65
		0.7	0.022	1.01	1.28	1.29	0.058	1.35	1.17	1.71
		0.9	0.023	1.01	1.31	1.33	0.060	1.33	1.25	1.78
0.10	0	0.5	0.013	1.01	1.21	1.21	0.033	1.33	1.10	1.51
		0.7	0.013	1.02	1.24	1.24	0.033	1.32	1.16	1.57
		0.9	0.014	1.03	1.27	1.27	0.034	1.30	1.23	1.64
	0.3	0.5	0.010	1.01	1.19	1.20	0.029	1.36	1.09	1.60
		0.7	0.011	1.02	1.21	1.22	0.029	1.34	1.15	1.65
		0.9	0.012	1.02	1.24	1.24	0.030	1.33	1.20	1.69

[a] Relative Efficiency (RE): calculated as the asymptotic variance under each design over that of BGOF

Table C.5: Asymptotic variance of $\hat{\boldsymbol{\beta}}$ in the balanced goodness-of-fit design (BGOF) and its efficiency relative to the goodness-of-fit based design (GOF), the balanced design (BD) and the case-control design (CC). The phase I cohort size was 3000. The prevalence was 0.05 and 0.10, the correlation parameter ρ for phase I variables was 0 and 0.3, and the true value of β_4 was 0.5, 0.7, and 0.9. The correlation between phase II and phase I variables was 0. X_1 was the stratifying variable in BGOF.

$P(Y=1)$	ρ	β_4	Var($\hat{\beta}_4$)	RE[a] of BGOF vs.			Var($\hat{\beta}_1$)	RE of BGOF vs.		
				GOF	BD	CC		GOF	BD	CC
0.05	0	0.5	0.016	1.00	1.10	1.11	0.143	1.27	0.99	1.40
		0.7	0.018	1.00	1.09	1.10	0.153	1.25	0.98	1.37
		0.9	0.021	1.00	1.08	1.08	0.164	1.22	0.99	1.33
	0.3	0.5	0.016	1.00	1.08	1.09	0.165	1.24	0.99	1.33
		0.7	0.018	1.00	1.07	1.08	0.176	1.22	0.99	1.31
		0.9	0.021	1.01	1.07	1.08	0.192	1.20	0.99	1.28
0.10	0	0.5	0.008	1.00	1.09	1.10	0.070	1.27	1.00	1.39
		0.7	0.009	1.00	1.08	1.09	0.074	1.25	1.00	1.37
		0.9	0.010	1.00	1.08	1.08	0.080	1.23	1.00	1.34
	0.3	0.5	0.008	1.00	1.08	1.09	0.080	1.23	0.99	1.33
		0.7	0.009	1.00	1.07	1.08	0.086	1.22	1.00	1.31
		0.9	0.010	1.00	1.06	1.07	0.092	1.20	0.99	1.29
			Var($\hat{\beta}_2$)	RE of BGOF vs.			Var($\hat{\beta}_3$)	RE of BGOF vs.		
				GOF	BD	CC		GOF	BD	CC
0.05	0	0.5	0.016	1.00	1.18	1.18	0.100	1.00	0.93	0.93
		0.7	0.016	1.01	1.17	1.18	0.104	1.00	0.93	0.94
		0.9	0.017	1.00	1.16	1.16	0.109	1.00	0.93	0.93
	0.3	0.5	0.017	1.00	1.15	1.16	0.086	1.01	1.02	1.04
		0.7	0.018	1.00	1.15	1.16	0.092	1.00	1.01	1.04
		0.9	0.019	1.01	1.14	1.15	0.098	1.00	1.01	1.03
0.10	0	0.5	0.008	1.01	1.16	1.16	0.047	1.00	0.94	0.95
		0.7	0.008	1.00	1.15	1.15	0.049	1.00	0.94	0.95
		0.9	0.009	1.00	1.15	1.14	0.051	1.00	0.94	0.95
	0.3	0.5	0.009	1.00	1.14	1.14	0.042	1.01	1.02	1.05
		0.7	0.009	1.00	1.13	1.14	0.044	1.00	1.02	1.05
		0.9	0.010	1.00	1.12	1.13	0.047	1.00	1.01	1.04

[a] Relative Efficiency (RE): calculated as the asymptotic variance under each design over that of BGOF

Table C.6: Asymptotic variance of $\hat{\boldsymbol{\beta}}$ in balanced goodness-of-fit design (BGOF) and its efficiency relative to the goodness-of-fit based design (GOF), the balanced design (BD) and the case-control design (CC). The phase I cohort size was 3000. The prevalence was 0.05 and 0.10, the correlation parameter ρ for phase I variables was 0 and 0.3, and the true value of β_4 was 0.5, 0.7, and 0.9. The correlation between phase II and phase I variables was 0. X_3 was the stratifying variable in BGOF.

$P(Y=1)$	ρ	β_4	Var($\hat{\beta}_4$)	RE[a] of BGOF vs.			Var($\hat{\beta}_1$)	RE of BGOF vs.		
				GOF	BD	CC		GOF	BD	CC
0.05	0.0	0.5	0.016	1.01	1.09	1.11	0.180	1.00	1.08	1.11
		0.7	0.018	1.00	1.07	1.10	0.190	1.00	1.07	1.10
		0.9	0.021	1.00	1.07	1.07	0.201	1.00	1.07	1.09
	0.3	0.5	0.016	1.01	1.09	1.10	0.202	1.01	1.07	1.08
		0.7	0.018	1.00	1.07	1.08	0.215	1.00	1.06	1.08
		0.9	0.021	1.01	1.07	1.08	0.230	1.00	1.06	1.07
0.10	0.0	0.5	0.008	1.00	1.08	1.10	0.088	1.00	1.08	1.10
		0.7	0.009	1.00	1.07	1.09	0.092	1.00	1.08	1.10
		0.9	0.010	1.00	1.06	1.08	0.098	1.00	1.07	1.09
	0.3	0.5	0.008	1.00	1.08	1.09	0.099	1.00	1.07	1.08
		0.7	0.009	1.00	1.07	1.08	0.104	1.00	1.07	1.08
		0.9	0.010	1.00	1.06	1.07	0.110	1.00	1.06	1.07
			Var($\hat{\beta}_2$)	RE of BGOF vs.			Var($\hat{\beta}_3$)	RE of BGOF vs.		
				GOF	BD	CC		GOF	BD	CC
0.05	0.0	0.5	0.016	1.00	1.16	1.19	0.072	1.39	0.91	1.29
		0.7	0.016	1.00	1.15	1.17	0.078	1.35	0.91	1.26
		0.9	0.017	1.00	1.14	1.16	0.083	1.32	0.92	1.22
	0.3	0.5	0.017	1.01	1.16	1.16	0.064	1.35	0.93	1.40
		0.7	0.018	1.00	1.15	1.16	0.070	1.32	0.94	1.37
		0.9	0.020	1.00	1.15	1.15	0.076	1.30	0.94	1.34
0.10	0.0	0.5	0.008	1.00	1.14	1.16	0.034	1.37	0.92	1.30
		0.7	0.008	1.00	1.13	1.16	0.036	1.35	0.93	1.28
		0.9	0.009	1.00	1.12	1.14	0.039	1.31	0.93	1.24
	0.3	0.5	0.009	1.01	1.14	1.15	0.031	1.35	0.96	1.40
		0.7	0.009	1.00	1.13	1.14	0.034	1.32	0.96	1.37
		0.9	0.010	1.00	1.12	1.13	0.036	1.29	0.96	1.34

[a] Relative Efficiency (RE): calculated as the asymptotic variance under each design over that of BGOF

Table C.7: The estimated log OR of phase I covariate ($\hat{\beta}_1$) under balanced goodness-of-fit based design (BGOF), the goodness-of-fit based design (GOF), the case-control design (CC), and the balanced design (BD). The phase I cohort size was 2×10^4, the prevalence was 0.05 and 0.10, the correlation parameter ρ for phase I variables was 0 and 0.3, and the true value of β_4 was 0.5, 0.7, and 0.9. The correlation between phase II variable X_4 and phase I variable X_1, X_2, and X_3 are 0.6, 0.5, and 0.3, respectively. X_1 was the stratifying variable in BGOF. The true value of β_1 is 0.5. The mean asymptotic standard error ("asym"), empirical standard error ("emp"), and coverage probability ("coverage") of $\hat{\beta}_4$ were calculated based on 1000 simulations.

$P(Y=1)$	ρ	β_4	BGOF (asym/emp)	coverage	GOF (asym/emp)	coverage	CC (asym)	BD (asym)
0.05	0.0	0.5	0.49 (0.35/0.34)	0.960	0.49 (0.42/0.41)	0.958	0.51 (0.45)	0.48 (0.38)
		0.7	0.48 (0.36/0.36)	0.950	0.48 (0.43/0.44)	0.942	0.48 (0.46)	0.48 (0.39)
		0.9	0.48 (0.37/0.37)	0.949	0.48 (0.44/0.46)	0.939	0.52 (0.48)	0.49 (0.41)
	0.3	0.5	0.50 (0.30/0.30)	0.948	0.50 (0.38/0.37)	0.952	0.53 (0.39)	0.49 (0.31)
		0.7	0.49 (0.31/0.31)	0.950	0.47 (0.39/0.39)	0.939	0.52 (0.41)	0.51 (0.33)
		0.9	0.47 (0.32/0.32)	0.936	0.48 (0.41/0.41)	0.939	0.50 (0.42)	0.50 (0.34)
0.10	0.0	0.5	0.50 (0.33/0.32)	0.944	0.47 (0.41/0.41)	0.952	0.52 (0.45)	0.49 (0.37)
		0.7	0.46 (0.34/0.35)	0.941	0.46 (0.42/0.41)	0.950	0.50 (0.46)	0.47 (0.38)
		0.9	0.48 (0.35/0.35)	0.943	0.47 (0.43/0.43)	0.965	0.53 (0.47)	0.47 (0.40)
	0.3	0.5	0.49 (0.28/0.28)	0.950	0.48 (0.37/0.36)	0.955	0.52 (0.39)	0.48 (0.30)
		0.7	0.49 (0.29/0.28)	0.967	0.48 (0.38/0.40)	0.942	0.51 (0.40)	0.48 (0.31)
		0.9	0.50 (0.30/0.29)	0.953	0.49 (0.40/0.40)	0.956	0.51 (0.42)	0.48 (0.33)

Table C.8: The estimated log OR of phase I covariate ($\hat{\beta}_2$) under balanced goodness-of-fit based design (BGOF), the goodness-of-fit based design (GOF), the case-control design (CC), and the balanced design (BD). The phase I cohort size was 2×10^4, the prevalence was 0.05 and 0.10, the correlation parameter ρ for phase I variables was 0 and 0.3, and the true value of β_4 was 0.5, 0.7, and 0.9. The correlation between phase II variable X_4 and phase I variable X_1, X_2, and X_3 are 0.6, 0.5, and 0.3, respectively. X_1 was the stratifying variable in BGOF. The true value of β_2 is 0.6. The mean asymptotic standard error ("asym"), empirical standard error ("emp"), and coverage probability ("coverage") of $\hat{\beta}_4$ were calculated based on 1000 simulations.

$P(Y=1)$	ρ	β_4	BGOF (asym/emp)	coverage	GOF (asym/emp)	coverage	CC (asym)	BD (asym)
0.05	0.0	0.5	0.59 (0.11/0.11)	0.956	0.59 (0.11/0.11)	0.960	0.61 (0.12)	0.60 (0.12)
		0.7	0.59 (0.11/0.12)	0.950	0.59 (0.11/0.11)	0.958	0.60 (0.13)	0.61 (0.13)
		0.9	0.60 (0.12/0.12)	0.961	0.60 (0.12/0.12)	0.938	0.62 (0.13)	0.61 (0.13)
	0.3	0.5	0.59 (0.10/0.10)	0.944	0.60 (0.10/0.10)	0.946	0.60 (0.11)	0.60 (0.11)
		0.7	0.60 (0.10/0.10)	0.947	0.59 (0.10/0.10)	0.953	0.61 (0.11)	0.60 (0.11)
		0.9	0.59 (0.10/0.10)	0.963	0.60 (0.11/0.11)	0.952	0.61 (0.12)	0.61 (0.11)
0.10	0.0	0.5	0.60 (0.11/0.11)	0.945	0.60 (0.11/0.11)	0.947	0.61 (0.12)	0.60 (0.12)
		0.7	0.59 (0.12/0.11)	0.949	0.59 (0.12/0.12)	0.949	0.61 (0.13)	0.60 (0.13)
		0.9	0.60 (0.12/0.12)	0.940	0.59 (0.12/0.12)	0.955	0.61 (0.13)	0.60 (0.13)
	0.3	0.5	0.60 (0.10/0.10)	0.943	0.60 (0.10/0.10)	0.959	0.61 (0.11)	0.60 (0.11)
		0.7	0.60 (0.10/0.11)	0.947	0.60 (0.10/0.10)	0.939	0.60 (0.11)	0.60 (0.11)
		0.9	0.59 (0.11/0.11)	0.942	0.59 (0.11/0.11)	0.954	0.61 (0.12)	0.61 (0.11)

Table C.9: The estimated log OR of phase I covariate ($\hat{\beta}_3$) under balanced goodness-of-fit based design (BGOF), the goodness-of-fit based design (GOF), the case-control design (CC), and the balanced design (BD). The phase I cohort size was 2×10^4, the prevalence was 0.05 and 0.10, the correlation parameter ρ for phase I variables was 0 and 0.3, and the true value of β_4 was 0.5, 0.7, and 0.9. The correlation between phase II variable X_4 and phase I variable X_1, X_2, and X_3 are 0.6, 0.5, and 0.3, respectively. X_1 was the stratifying variable in BGOF. The true value of β_3 is -0.7. The mean asymptotic standard error ("asym"), empirical standard error ("emp"), and coverage probability ("coverage") of $\hat{\beta}_4$ were calculated based on 1000 simulations.

$P(Y=1)$	ρ	β_4	BGOF (asym/emp)	coverage	GOF (asym/emp)	coverage	CC (asym)	BD (asym)
0.05	0.0	0.5	-0.71 (0.21/0.21)	0.944	-0.71 (0.21/0.21)	0.946	-0.70 (0.23)	-0.71 (0.22)
		0.7	-0.70 (0.21/0.22)	0.947	-0.71 (0.21/0.21)	0.953	-0.71 (0.23)	-0.71 (0.23)
		0.9	-0.71 (0.21/0.22)	0.963	-0.71 (0.21/0.21)	0.952	-0.71 (0.24)	-0.71 (0.23)
	0.3	0.5	-0.70 (0.20/0.20)	0.957	-0.70 (0.19/0.20)	0.948	-0.70 (0.21)	-0.71 (0.21)
		0.7	-0.70 (0.20/0.21)	0.960	-0.70 (0.19/0.20)	0.935	-0.73 (0.22)	-0.70 (0.22)
		0.9	-0.70 (0.20/0.20)	0.951	-0.70 (0.19/0.19)	0.956	-0.69 (0.22)	-0.70 (0.22)
0.10	0.0	0.5	-0.70 (0.21/0.20)	0.943	-0.71 (0.21/0.21)	0.959	-0.70 (0.22)	-0.71 (0.22)
		0.7	-0.70 (0.21/0.21)	0.947	-0.71 (0.21/0.21)	0.939	-0.70 (0.23)	-0.71 (0.23)
		0.9	-0.71 (0.21/0.20)	0.942	-0.70 (0.21/0.21)	0.954	-0.72 (0.23)	-0.71 (0.23)
	0.3	0.5	-0.71 (0.20/0.19)	0.948	-0.70 (0.19/0.19)	0.946	-0.70 (0.21)	-0.70 (0.21)
		0.7	-0.70 (0.20/0.21)	0.946	-0.69 (0.19/0.19)	0.955	-0.71 (0.22)	-0.70 (0.21)
		0.9	-0.70 (0.20/0.21)	0.956	-0.71 (0.20/0.20)	0.954	-0.71 (0.22)	-0.72 (0.22)

Table C.10: The estimated log OR of phase I covariate ($\hat{\beta}_4$) under balanced goodness-of-fit based design (BGOF), the goodness-of-fit based design (GOF), the case-control design (CC), and the balanced design (BD). The phase I cohort size was 2×10^4, the prevalence was 0.05 and 0.10, the correlation parameter ρ for phase I variables was 0 and 0.3, and the true value of β_4 was 0.5, 0.7, and 0.9. The correlation between phase II variable X_4 and phase I variable X_1, X_2, and X_3 are 0.6, 0.5, and 0.3, respectively. The mean asymptotic standard error ("asym"), empirical standard error ("emp"), and coverage probability ("coverage") of $\hat{\beta}_4$ were calculated based on 1000 simulations. X_1 was the stratifying variable in BGOF.

$P(Y=1)$	ρ	β_4	BGOF (asym/emp)	coverage	GOF (asym/emp)	coverage	CC (asym)	BD (asym)
0.05	0	0.5	0.51 (0.14/0.14)	0.947	0.50 (0.14/0.14)	0.953	0.51 (0.16)	0.51 (0.15)
		0.7	0.71 (0.15/0.15)	0.949	0.71 (0.15/0.15)	0.952	0.71 (0.16)	0.71 (0.16)
		0.9	0.91 (0.15/0.15)	0.946	0.92 (0.15/0.15)	0.949	0.91 (0.17)	0.91 (0.17)
	0.3	0.5	0.50 (0.12/0.12)	0.959	0.50 (0.12/0.12)	0.960	0.50 (0.13)	0.50 (0.13)
		0.7	0.71 (0.12/0.12)	0.953	0.71 (0.12/0.12)	0.950	0.70 (0.13)	0.70 (0.13)
		0.9	0.91 (0.13/0.12)	0.959	0.91 (0.13/0.13)	0.939	0.91 (0.14)	0.90 (0.14)
0.10	0	0.5	0.50 (0.14/0.14)	0.958	0.51 (0.14/0.14)	0.957	0.50 (0.16)	0.51 (0.16)
		0.7	0.72 (0.15/0.15)	0.941	0.71 (0.15/0.14)	0.950	0.70 (0.16)	0.72 (0.16)
		0.9	0.91 (0.15/0.15)	0.954	0.91 (0.15/0.15)	0.986	0.91 (0.17)	0.92 (0.17)
	0.3	0.5	0.51 (0.12/0.12)	0.938	0.50 (0.12/0.11)	0.957	0.50 (0.13)	0.51 (0.13)
		0.7	0.71 (0.12/0.12)	0.955	0.71 (0.12/0.12)	0.948	0.71 (0.13)	0.72 (0.13)
		0.9	0.91 (0.13/0.12)	0.958	0.91 (0.13/0.13)	0.957	0.91 (0.14)	0.91 (0.14)

Table C.11: Asymptotic variance of $\hat{\boldsymbol{\beta}}$ under the balanced goodness-of-fit design (BGOF) and its efficiency relative to the goodness-of-fit based design (GOF), the balanced design (BD) and the case-control design (CC). The phase I cohort size was 2×10^4. The prevalence was 0.05 and 0.10, the correlation parameter ρ for phase I variables was 0 and 0.3, and the true value of β_4 was 0.5, 0.7, and 0.9. The correlation between phase II variable X_4 and phase I variable X_1, X_2, and X_3 are 0.6, 0.5, and 0.3, respectively. X_1 was the stratifying variable in BGOF.

$P(Y=1)$	ρ	β_4	Var($\hat{\beta}_4$)	RE of BGOF vs.			Var($\hat{\beta}_1$)	RE of BGOF vs.		
				GOF	BD	CC		GOF	BD	CC
0.05	0	0.5	0.020	1.00	1.20	1.23	0.12	1.41	1.15	1.65
		0.7	0.021	1.00	1.24	1.28	0.13	1.42	1.19	1.67
		0.9	0.022	1.00	1.28	1.34	0.14	1.42	1.22	1.70
	0.3	0.5	0.014	1.00	1.15	1.20	0.09	1.57	1.10	1.74
		0.7	0.015	1.00	1.17	1.23	0.10	1.58	1.12	1.74
		0.9	0.016	1.01	1.19	1.28	0.10	1.58	1.14	1.73
0.10	0	0.5	0.020	1.01	1.19	1.22	0.11	1.55	1.21	1.81
		0.7	0.021	1.00	1.22	1.26	0.12	1.54	1.24	1.82
		0.9	0.023	1.00	1.26	1.31	0.12	1.54	1.28	1.84
	0.3	0.5	0.014	1.00	1.15	1.19	0.08	1.79	1.18	2.00
		0.7	0.015	1.01	1.17	1.22	0.08	1.80	1.22	2.00
		0.9	0.016	1.01	1.19	1.26	0.09	1.79	1.24	1.99
			Var($\hat{\beta}_2$)	RE of BGOF vs.			Var($\hat{\beta}_3$)	RE of BGOF vs.		
				GOF	BD	CC		GOF	BD	CC
0.05	0	0.5	0.045	1.00	1.11	1.14	0.045	1.00	1.11	1.14
		0.7	0.044	1.00	1.18	1.22	0.044	1.00	1.18	1.22
		0.9	0.043	1.00	1.25	1.31	0.043	1.00	1.25	1.31
	0.3	0.5	0.040	0.94	1.11	1.12	0.040	0.94	1.11	1.12
		0.7	0.041	0.93	1.15	1.17	0.041	0.93	1.15	1.17
		0.9	0.041	0.92	1.19	1.22	0.041	0.92	1.19	1.22
0.10	0	0.5	0.043	1.00	1.13	1.15	0.043	1.00	1.13	1.15
		0.7	0.043	1.00	1.18	1.22	0.043	1.00	1.18	1.22
		0.9	0.043	1.00	1.23	1.29	0.043	1.00	1.23	1.29
	0.3	0.5	0.040	0.95	1.12	1.13	0.040	0.95	1.12	1.13
		0.7	0.040	0.94	1.15	1.18	0.040	0.94	1.15	1.18
		0.9	0.041	0.94	1.18	1.21	0.041	0.94	1.18	1.21

Table C.12: Asymptotic variance of $\hat{\boldsymbol{\beta}}$ under the balanced goodness-of-fit design (BGOF) and its efficiency relative to the goodness-of-fit based design (GOF), the balanced design (BD) and the case-control design (CC). The phase I cohort size was 2×10^4. The prevalence was 0.05 and 0.10, the correlation parameter ρ for phase I variables was 0 and 0.3, and the true value of β_4 was 0.5, 0.7, and 0.9. The correlation between phase II variable X_4 and phase I variable X_1, X_2, and X_3 are 0.6, 0.5, and 0.3, respectively. X_3 was the stratifying variable in BGOF.

$P(Y=1)$	ρ	β_4	Var($\hat{\beta}_4$)	RE of BGOF vs.			Var($\hat{\beta}_1$)	RE of BGOF vs.		
				GOF	BD	CC		GOF	BD	CC
0.05	0	0.5	0.020	0.99	1.22	1.22	0.170	0.99	1.15	1.16
		0.7	0.021	1.00	1.28	1.27	0.180	1.00	1.18	1.18
		0.9	0.022	1.00	1.33	1.33	0.200	1.00	1.19	1.19
	0.3	0.5	0.014	1.01	1.20	1.20	0.150	0.96	1.09	1.06
		0.7	0.015	1.01	1.23	1.24	0.160	0.96	1.08	1.05
		0.9	0.016	1.01	1.27	1.28	0.170	0.96	1.08	1.05
0.10	0	0.5	0.020	0.99	1.21	1.21	0.170	0.99	1.16	1.16
		0.7	0.021	0.99	1.27	1.26	0.180	0.99	1.18	1.18
		0.9	0.023	1.00	1.31	1.30	0.190	1.00	1.20	1.19
	0.3	0.5	0.014	1.01	1.19	1.19	0.140	0.97	1.10	1.08
		0.7	0.015	1.00	1.22	1.22	0.150	0.96	1.10	1.08
		0.9	0.016	1.01	1.25	1.25	0.160	0.96	1.11	1.08
			Var($\hat{\beta}_2$)	RE of BGOF vs.			Var($\hat{\beta}_3$)	RE of BGOF vs.		
				GOF	BD	CC		GOF	BD	CC
0.05	0	0.5	0.013	1.00	1.22	1.23	0.019	2.40	1.02	2.73
		0.7	0.013	1.00	1.27	1.26	0.019	2.29	1.15	2.80
		0.9	0.014	1.00	1.31	1.31	0.020	2.17	1.26	2.86
	0.3	0.5	0.010	1.00	1.21	1.22	0.015	2.46	1.07	2.95
		0.7	0.011	1.00	1.24	1.25	0.016	2.34	1.18	2.95
		0.9	0.011	1.01	1.28	1.28	0.017	2.23	1.29	2.96
0.10	0	0.5	0.013	1.00	1.20	1.20	0.013	3.40	1.25	3.93
		0.7	0.013	1.00	1.23	1.23	0.013	3.20	1.40	3.93
		0.9	0.014	1.00	1.26	1.25	0.014	2.98	1.52	3.85
	0.3	0.5	0.010	1.01	1.19	1.20	0.011	3.58	1.30	4.24
		0.7	0.011	1.00	1.21	1.22	0.011	3.30	1.42	4.11
		0.9	0.011	1.01	1.23	1.23	0.013	3.04	1.52	3.94

Table C.13: Asymptotic variance of $\hat{\boldsymbol{\beta}}$ in balanced goodness-of-fit design (BGOF) and its efficiency relative to case-control designs (CC), balanced design (BD), and goodness-of-fit design (GOF). The phase I cohort size is 2×10^4. The prevalence is 0.05 and 0.10, the correlation ρ among phase I variables is 0 and 0.3, and the true value of β_4 is 0.5, 0.7, and 0.9. The correlation between phase II variable X_4 and phase I variable X_1, X_2, and X_3 is 0. X_1 is the stratifying variable in BGOF.

$P(Y=1)$	ρ	β_4	Var($\hat{\beta}_4$)	RE of BGOF vs.			Var($\hat{\beta}_1$)	RE of BGOF vs.		
				GOF	BD	CC		GOF	BD	CC
0.05	0	0.5	0.008	1.01	1.10	1.11	0.046	1.91	0.97	2.10
		0.7	0.009	1.01	1.09	1.09	0.051	1.82	0.98	2.00
		0.9	0.010	1.01	1.09	1.09	0.057	1.75	0.98	1.90
	0.3	0.5	0.008	1.00	1.08	1.09	0.056	1.78	0.97	1.92
		0.7	0.009	1.00	1.07	1.08	0.061	1.72	0.98	1.84
		0.9	0.010	1.00	1.07	1.08	0.068	1.64	0.98	1.76
0.10	0	0.5	0.008	1.00	1.10	1.10	0.037	2.38	1.09	2.61
		0.7	0.009	1.00	1.09	1.09	0.041	2.23	1.07	2.43
		0.9	0.010	1.00	1.08	1.08	0.047	2.09	1.06	2.27
	0.3	0.5	0.008	1.00	1.07	1.09	0.046	2.15	1.07	2.33
		0.7	0.009	1.01	1.07	1.08	0.051	2.03	1.06	2.19
		0.9	0.010	1.01	1.05	1.06	0.058	1.91	1.04	2.03
			Var($\hat{\beta}_2$)	RE of BGOF vs.			Var($\hat{\beta}_3$)	RE of BGOF vs.		
				GOF	BD	CC		GOF	BD	CC
0.05	0	0.5	0.008	1.00	1.16	1.17	0.047	1.00	0.94	0.94
		0.7	0.008	1.00	1.16	1.16	0.049	1.01	0.94	0.95
		0.9	0.008	1.01	1.15	1.16	0.053	1.00	0.94	0.94
	0.3	0.5	0.008	1.00	1.14	1.15	0.044	0.97	1.01	1.01
		0.7	0.009	1.00	1.14	1.14	0.046	0.96	1.01	1.00
		0.9	0.010	1.00	1.12	1.14	0.049	0.97	1.01	1.00
0.10	0	0.5	0.008	1.00	1.14	1.15	0.045	1.00	0.96	0.96
		0.7	0.008	1.00	1.14	1.15	0.047	1.00	0.96	0.97
		0.9	0.009	1.01	1.13	1.14	0.050	1.01	0.97	0.97
	0.3	0.5	0.009	1.01	1.13	1.14	0.043	0.97	1.03	1.02
		0.7	0.009	1.01	1.12	1.13	0.045	0.97	1.02	1.02
		0.9	0.010	1.00	1.11	1.11	0.048	0.98	1.02	1.01

Table C.14: Asymptotic variance of $\hat{\boldsymbol{\beta}}$ in balanced goodness-of-fit design (BGOF) and its efficiency relative to case-control designs (CC), balanced design (BD), and goodness-of-fit design (GOF). The phase I cohort size is 2×10^4. The prevalence is 0.05 and 0.10, the correlation ρ among phase I variables is 0 and 0.3, and the true value of β_4 is 0.5, 0.7, and 0.9. The correlation between phase II variable X_4 and phase I variable X_1, X_2, and X_3 is 0. X_3 is the stratifying variable in BGOF.

$P(Y=1)$	ρ	β_4	Var($\hat{\beta}_4$)	RE of BGOF vs.			Var($\hat{\beta}_1$)	RE of BGOF vs.		
				GOF	BD	CC		GOF	BD	CC
0.05	0.0	0.5	0.008	1.04	1.08	1.11	0.09	1.03	1.08	1.10
		0.7	0.009	1.03	1.08	1.10	0.09	1.03	1.08	1.10
		0.9	0.010	1.03	1.06	1.09	0.10	1.03	1.08	1.09
	0.3	0.5	0.008	1.01	1.09	1.10	0.10	0.97	1.06	1.05
		0.7	0.009	1.01	1.08	1.09	0.11	0.96	1.05	1.03
		0.9	0.010	1.00	1.07	1.08	0.12	0.97	1.05	1.03
0.10	0.0	0.5	0.008	1.02	1.08	1.10	0.09	1.02	1.08	1.10
		0.7	0.009	1.03	1.07	1.09	0.09	1.02	1.08	1.09
		0.9	0.010	1.02	1.06	1.09	0.10	1.01	1.07	1.09
	0.3	0.5	0.008	1.00	1.08	1.09	0.10	0.97	1.06	1.05
		0.7	0.009	1.00	1.07	1.08	0.11	0.97	1.06	1.04
		0.9	0.010	1.01	1.06	1.06	0.11	0.97	1.05	1.03
			Var($\hat{\beta}_2$)	RE of BGOF vs.			Var($\hat{\beta}_3$)	RE of BGOF vs.		
				GOF	BD	CC		GOF	BD	CC
0.05	0.0	0.5	0.007	1.04	1.15	1.18	0.014	3.66	0.85	3.31
		0.7	0.008	1.03	1.15	1.18	0.015	3.44	0.87	3.12
		0.9	0.008	1.03	1.15	1.18	0.017	3.25	0.89	2.93
	0.3	0.5	0.008	1.01	1.15	1.15	0.017	2.52	0.84	2.63
		0.7	0.009	1.00	1.15	1.14	0.019	2.39	0.84	2.49
		0.9	0.010	1.00	1.13	1.14	0.021	2.28	0.86	2.35
0.10	0.0	0.5	0.008	1.03	1.15	1.17	0.008	6.07	1.03	5.66
		0.7	0.008	1.03	1.14	1.16	0.009	5.38	1.03	4.96
		0.9	0.009	1.02	1.13	1.15	0.011	4.68	1.02	4.36
	0.3	0.5	0.009	1.01	1.13	1.14	0.011	3.65	0.97	3.83
		0.7	0.009	1.00	1.12	1.13	0.013	3.32	0.97	3.47
		0.9	0.010	1.00	1.11	1.11	0.015	3.06	0.98	3.17

APPENDIX D

ADDITIONAL SIMULATION RESULTS FOR CHAPTER 4

Table D.1: The estimated log OR ($\boldsymbol{\alpha}$) in the missingness model and the estimated MAF (θ) using the family-supplemented weighted empirical likelihood method. The prevalence is 0.03, genotype availability is 0.8 and 0.6, the true value of β_2 is 0.182 and 0.405, and MAF is 0.2. The true values of $(\alpha_3, \alpha_4, \alpha_5)$ in the three missingness models are: weak $= (0.182, 0.405, 0.405)$, strong $= (0.405, 0.405, 0.405)$, and No interaction (NI) $= (0.405, 0, 0)$. The mean asymptotic standard error ("asym") and empirical standard error ("emp") of $\hat{\boldsymbol{\alpha}}$ were calculated based on 1000 simulations.

$P(R=1)$	e^{β_2}	missing	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_2$
0.8	1.2	weak	1.213 (0.115/0.114)	-0.513 (0.162/0.165)	0.180 (0.119/0.116)
		strong	1.123 (0.113/0.113)	-0.506 (0.160/0.156)	0.192 (0.119/0.116)
		NI	1.139 (0.113/0.113)	-0.512 (0.156/0.162)	0.190 (0.120/0.117)
	1.5	weak	1.211 (0.114/0.115)	-0.509 (0.165/0.162)	0.180 (0.119/0.118)
		strong	1.138 (0.113/0.117)	-0.519 (0.163/0.165)	0.175 (0.119/0.119)
		NI	1.140 (0.113/0.116)	-0.508 (0.158/0.156)	0.186 (0.120/0.120)
0.6	1.2	weak	0.224 (0.092/0.094)	-0.517 (0.136/0.136)	0.182 (0.097/0.097)
		strong	0.135 (0.092/0.094)	-0.509 (0.137/0.140)	0.188 (0.098/0.097)
		NI	0.160 (0.092/0.095)	-0.514 (0.133/0.138)	0.180 (0.098/0.097)
	1.5	weak	0.223 (0.092/0.093)	-0.511 (0.138 /0.141)	0.179 (0.097/0.098)
		strong	0.144 (0.092/0.092)	-0.513 (0.139/0.133)	0.180 (0.098/0.097)
		NI	0.159 (0.092/0.094)	-0.506 (0.135/0.137)	0.187 (0.098/0.099)

$P(R=1)$	e^{β_2}	missing	$\hat{\alpha_3}$	$\hat{\alpha_4}$	$\hat{\alpha_5}$
0.8	1.2	weak	0.194 (0.152/0.158)	0.410 (0.167/0.169)	0.400 (0.220/0.220)
		strong	0.423 (0.161/0.161)	0.401 (0.169/0.162)	0.408 (0.234/0.222)
		NI	0.414 (0.162/0.166)	-0.007 (0.161/0.162)	0.0001 (0.208/0.208)
	1.5	weak	0.189 (0.152/0.150)	0.407 (0.168/0.164)	0.406 (0.215/0.213)
		strong	0.408 (0.161/0.164)	0.418 (0.170/0.163)	0.407 (0.228/0.230)
		NI	0.413 (0.162/0.164)	-0.008 (0.161/0.161)	-0.003 (0.204/0.198)
0.6	1.2	weak	0.181 (0.113/ 0.112)	0.406 (0.140/0.138)	0.414 (0.164/0.156)
		strong	0.411 (0.117/0.120)	0.404 (0.142/0.141)	0.412 (0.171/0.168)
		NI	0.409 (0.118/0.119)	0.004 (0.139/0.139)	0.003 (0.157/0.156)
	1.5	weak	0.188 (0.113/0.115)	0.408 (0.140/0.145)	0.400 (0.160/0.158)
		strong	0.406 (0.118/0.118)	0.408 (0.142/0.137)	0.404 (0.167/0.166)
		NI	0.408 (0.118/0.120)	-0.008 (0.138/0.144)	-0.005 (0.154/0.149)

Table D.2: The estimated log OR of covariate X (β_1) and genotype G (β_2) in the association model and the estimated MAF (θ) using the family-supplemented weighted empirical likelihood method. The prevalence is 0.03, genotype availability is 0.8 and 0.6, the true value of β_2 is 0.182 and 0.405, and MAF is 0.5. The true values of $(\alpha_3, \alpha_4, \alpha_5)$ in the three missingness models are: weak $= (0.182, 0.405, 0.405)$, strong $= (0.405, 0.405, 0.405)$, and No interaction (NI) $= (0.405, 0, 0)$. The mean asymptotic standard error ("asym"), empirical standard error ("emp"), and coverage probability ("coverage") of $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\theta}$ were calculated based on 1000 simulations.

$P(R=1)$	e^{α_2}	missing	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\theta}$	
			estimate (asy/emp)	coverage	estimate (asy/emp)	coverage	estimate (asy/emp)	coverage
0.8	1.2	weak	0.182 (0.068/0.061)	0.978	0.183 (0.050/0.051)	0.942	0.498 (0.009/0.009)	0.960
		strong	0.183 (0.068/0.064)	0.964	0.182 (0.049/0.050)	0.954	0.499 (0.009/0.009)	0.949
		NI	0.184 (0.068/0.063)	0.962	0.185 (0.051/0.052)	0.938	0.498 (0.009/0.009)	0.955
	1.5	weak	0.186 (0.070/0.067)	0.956	0.405 (0.050/0.051)	0.951	0.497 (0.009/0.009)	0.943
		strong	0.180 (0.069/0.065)	0.965	0.403 (0.050/0.047)	0.961	0.497 (0.009/0.009)	0.939
		NI	0.182 (0.069/0.066)	0.963	0.404 (0.052/0.054)	0.938	0.497 (0.009/0.009)	0.953
0.6	1.2	weak	0.182 (0.073/0.061)	0.983	0.183 (0.055/0.056)	0.953	0.498 (0.010/0.010)	0.965
		strong	0.183 (0.073/0.064)	0.975	0.181 (0.056/0.057)	0.951	0.499 (0.010/0.010)	0.947
		NI	0.185 (0.073/0.063)	0.971	0.184 (0.058/0.057)	0.948	0.498 (0.010/0.010)	0.954
	1.5	weak	0.186 (0.074/0.067)	0.967	0.404 (0.056/0.056)	0.954	0.497 (0.010/0.010)	0.953
		strong	0.180 (0.074/0.066)	0.975	0.402 (0.056/0.055)	0.952	0.497 (0.010/0.010)	0.949
		NI	0.182 (0.074/0.066)	0.974	0.405 (0.059/0.059)	0.949	0.497 (0.010/0.010)	0.943

Table D.3: The estimated log OR ($\boldsymbol{\alpha}$) in the missingness model and the estimated MAF (θ) using the family-supplemented weighted empirical likelihood method. The prevalence is 0.03, genotype availability is 0.8 and 0.6, the true value of β_2 is 0.182 and 0.405, and MAF is 0.5. The true values of $(\alpha_3, \alpha_4, \alpha_5)$ in the three missingness models are: weak $= (0.182, 0.405, 0.405)$, strong $= (0.405, 0.405, 0.405)$, and No interaction (NI) $= (0.405, 0, 0)$. The mean asymptotic standard error ("asym") and empirical standard error ("emp") of $\hat{\boldsymbol{\alpha}}$ were calculated based on 1000 simulations.

$P(R=1)$	e^{β_2}	missing	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_2$
0.8	1.2	weak	1.097 (0.144/0.148)	-0.515 (0.211/0.212)	0.180 (0.113/0.111)
		strong	0.892 (0.140/0.143)	-0.512 (0.207/0.201)	0.182 (0.114/0.113)
		NI	0.926 (0.142/0.148)	-0.504 (0.195/0.200)	0.187 (0.116/0.113)
	1.5	weak	1.090 (0.144/0.149)	-0.514 (0.220/0.223)	0.184 (0.113/0.108)
		strong	0.891 (0.140/0.146)	-0.508 (0.216/0.221)	0.181 (0.114/0.116)
		NI	0.931 (0.142/0.149)	-0.506 (0.201/0.210)	0.184 (0.116/0.119)
0.6	1.2	weak	0.120 (0.114/0.120)	-0.511 (0.171/0.172)	0.183 (0.093/ 0.093)
		strong	-0.098 (0.113/0.114)	-0.510 (0.172/0.171)	0.182 (0.094/0.093)
		NI	-0.095 (0.114/0.115)	-0.502 (0.164/0.164)	0.188 (0.094/0.093)
	1.5	weak	0.119 (0.114/ 0.111)	-0.510 (0.177/0.173)	0.185 (0.093/0.091)
		strong	-0.096 (0.113/0.115)	-0.517 (0.178/0.178)	0.181 (0.094/0.093)
		NI	-0.090 (0.113/0.113)	-0.504 (0.169/0.166)	0.185 (0.094/0.093)

$P(R=1)$	e^{β_2}	missing	$\hat{\alpha_3}$	$\hat{\alpha_4}$	$\hat{\alpha_5}$
0.8	1.2	weak	0.179 (0.120/0.122)	0.415 (0.165/0.167)	0.403 (0.179/0.175)
		strong	0.404 (0.121/0.123)	0.407 (0.168/0.165)	0.407 (0.183/0.175)
		NI	0.410 (0.124/0.131)	0.001 (0.155/0.154)	-0.007 (0.163/0.168)
	1.5	weak	0.184 (0.120/0.123)	0.409 (0.167/0.164)	0.407 (0.181/0.176)
		strong	0.408 (0.122/0.127)	0.405 (0.169/0.172)	0.405 (0.185/0.187)
		NI	0.409 (0.124/0.126)	0.001 (0.156/0.159)	-0.005 (0.164/0.167)
0.6	1.2	weak	0.182 (0.090/0.094)	0.406 (0.136/0.135)	0.405 (0.135/0.133)
		strong	0.404 (0.091/0.094)	0.408 (0.139/0.135)	0.407 (0.139/0.139)
		NI	0.408 (0.092/0.094)	-0.006 (0.133/0.133)	-0.004 (0.127/0.126)
	1.5	weak	0.182 (0.090/0.090)	0.400 (0.136/0.138)	0.409 (0.136/0.131)
		strong	0.402 (0.092/0.093)	0.407 (0.139/0.141)	0.413 (0.140/0.138)
		NI	0.405 (0.092/0.092)	-0.007 (0.133/0.128)	-0.004 (0.127/0.125)

Table D.4: The mean bias in and the mean square error (MSE) of estimated log OR of covariate X (β_1) and genotype G (β_2) in the association model and the estimated MAF (θ) using the family-supplemented weighted empirical likelihood method (FS-WEL) and the naive method based on 1000 simulations. The prevalence is 0.03, genotype availability is 0.8 and 0.6, the true value of β_2 is 0.182 and 0.405, and MAF is 0.5. The true values of $(\alpha_3, \alpha_4, \alpha_5)$ in the three missingness models are: weak $= (0.182, 0.405, 0.405)$, strong $= (0.405, 0.405, 0.405)$, and No interaction (NI) $= (0.405, 0, 0)$. In each of the 12 settings, true values and estimates of coefficients β_1, β_2 and θ are presented in this order in the magnitude of 10^{-3}.

Model				Naive Method		FS-WPL Method	
$P(R=1)$	e^{β_2}	Missing	True	Bias[a]	MSE	Bias	MSE
0.8	1.2	weak	182	72.324	10.124	0.033	3.739
			182	67.926	7.017	1.042	2.566
			500	6.948	0.128	-1.725	0.080
		strong	182	67.882	9.920	0.288	4.083
			182	62.074	6.475	-0.631	2.506
			500	18.470	0.421	-1.190	0.083
		NI	182	18.951	5.832	2.062	3.984
			182	37.357	4.048	2.326	2.737
			500	17.588	0.386	-1.636	0.083
	1.5	weak	182	72.487	10.654	3.643	4.445
			405	66.778	7.044	-0.168	2.595
			500	6.040	0.111	-2.865	0.090
		strong	182	64.075	9.092	-2.146	4.196
			405	59.557	6.001	-2.480	2.258
			500	16.818	0.360	-2.908	0.088
		NI	182	15.933	5.963	-0.193	4.409
			405	34.628	4.129	-1.024	2.946
			500	16.351	0.343	-2.841	0.086
0.6	1.2	weak	182	138.307	25.742	0.090	3.746
			182	133.852	21.147	0.527	3.087
			500	15.832	0.360	-1.699	0.098
		strong	182	131.333	23.961	0.624	4.102
			182	120.563	18.015	-1.693	3.245
			500	38.113	1.558	-1.086	0.105
		NI	182	22.227	8.218	2.412	3.999
			182	51.096	6.351	2.069	3.282
			500	37.579	1.514	-1.724	0.101
	1.5	weak	182	136.611	25.988	3.697	4.506
			405	131.872	20.655	-1.853	3.127
			500	14.835	0.322	-2.625	0.110
		strong	182	126.493	22.998	-2.285	4.301
			405	120.704	17.906	-3.164	2.981
			500	36.126	1.408	-2.717	0.107
		NI	182	19.535	8.242	0.092	4.395
			405	48.261	6.026	-0.392	3.442
			500	36.311	1.412	-2.771	0.109

[a] Bias is defined as the mean estimate minus the true value

Table D.5: The estimated log OR of covariate X (β_1) and genotype G (β_2) in the association model and the estimated MAF (θ) using the family-supplemented weighted empirical likelihood method. The prevalence is 0.10, genotype availability is 0.8 and 0.6, the true value of β_2 is 0.182 and 0.405, and MAF is 0.5. The true values of $(\alpha_3, \alpha_4, \alpha_5)$ in the three missingness models are: weak $= (0.182, 0.405, 0.405)$, strong $= (0.405, 0.405, 0.405)$, and No interaction (NI) $= (0.405, 0, 0)$. The mean asymptotic standard error ("asym"), empirical standard error ("emp"), and coverage probability ("coverage") of $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\theta}$ were calculated based on 1000 simulations.

$P(R=1)$	e^{α_2}	missing	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\theta}$	
			estimate (asy/emp)	coverage	estimate (asy/emp)	coverage	estimate (asy/emp)	coverage
0.8	1.2	weak	0.181 (0.068/0.066)	0.958	0.181 (0.049/0.052)	0.939	0.495 (0.0093/0.0090)	0.915
		strong	0.181 (0.068/0.066)	0.948	0.182 (0.049/0.052)	0.933	0.495 (0.0093/0.0092)	0.927
		NI	0.183 (0.068/0.063)	0.966	0.185 (0.051/0.053)	0.933	0.495 (0.0092/0.0091)	0.931
	1.5	weak	0.182 (0.070/0.065)	0.963	0.402 (0.050/0.050)	0.954	0.490 (0.0093/0.0090)	0.824
		strong	0.181 (0.070/0.063)	0.968	0.406 (0.050/0.051)	0.943	0.490 (0.0093/0.0091)	0.802
		NI	0.186 (0.069/0.066)	0.959	0.402 (0.051/0.051)	0.954	0.490 (0.0092/0.0087)	0.818
0.6	1.2	weak	0.181 (0.073/0.066)	0.960	0.181 (0.056/0.057)	0.939	0.495 (0.0105/0.0101)	0.925
		strong	0.180 (0.073/0.066)	0.966	0.181 (0.056/0.058)	0.931	0.495 (0.0105/0.0105)	0.930
		NI	0.183 (0.072/0.064)	0.973	0.184 (0.058/0.059)	0.941	0.495 (0.0104/0.0101)	0.930
	1.5	weak	0.182 (0.074/0.066)	0.975	0.401 (0.056/0.056)	0.954	0.490 (0.0105/0.0099)	0.864
		strong	0.181 (0.074/0.063)	0.977	0.406 (0.056/0.056)	0.949	0.490 (0.0105/0.0101)	0.853
		NI	0.186 (0.074/0.066)	0.968	0.402 (0.058/0.057)	0.962	0.490 (0.0104/0.0097)	0.851

Table D.6: The estimated log OR ($\boldsymbol{\alpha}$) in the missingness model and the estimated MAF (θ) using the family-supplemented weighted empirical likelihood method. The prevalence is 0.10, genotype availability is 0.8 and 0.6, the true value of β_2 is 0.182 and 0.405, and MAF is 0.5. The true values of $(\alpha_3, \alpha_4, \alpha_5)$ in the three missingness models are: weak $= (0.182, 0.405, 0.405)$, strong $= (0.405, 0.405, 0.405)$, and No interaction (NI) $= (0.405, 0, 0)$. The mean asymptotic standard error ("asym") and empirical standard error ("emp") of $\hat{\boldsymbol{\alpha}}$ were calculated based on 1000 simulations.

$P(R=1)$	e^{β_2}	missing	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_2$
0.8	1.2	weak	1.089 (0.143/0.142)	-0.512 (0.210/0.206)	0.178 (0.113/0.115)
		strong	0.883 (0.140/0.137)	-0.508 (0.206/0.197)	0.190 (0.114/0.113)
		NI	0.950 (0.142/0.142)	-0.511 (0.195/0.196)	0.180 (0.116/0.115)
	1.5	weak	1.090 (0.143/0.147)	-0.515 (0.219/0.223)	0.189 (0.113/0.115)
		strong	0.891 (0.140/0.137)	-0.520 (0.215/0.218)	0.185 (0.114/0.114)
		NI	0.946 (0.142/0.149)	-0.506 (0.201/0.203)	0.185 (0.116/0.120)
0.6	1.2	weak	0.100 (0.113/0.115)	-0.519 (0.170/0.177)	0.184 (0.093/0.091)
		strong	-0.108 (0.113/0.113)	-0.520 (0.171/0.170)	0.180 (0.094/0.093)
		NI	-0.045 (0.114/0.116)	-0.511 (0.164/0.160)	0.180 (0.095/0.091)
	1.5	weak	0.102 (0.113/0.122)	-0.519 (0.175/0.184)	0.181 (0.093/0.097)
		strong	-0.111 (0.113/0.115)	-0.507 (0.177/0.177)	0.185 (0.094/0.095)
		NI	-0.047 (0.113/0.116)	-0.511 (0.168 /0.169)	0.181 (0.095/0.097)
$P(R=1)$	e^{β_2}	missing	$\hat{\alpha_3}$	$\hat{\alpha_4}$	$\hat{\alpha_5}$
0.8	1.2	weak	0.187 (0.120/0.120)	0.412 (0.165/0.164)	0.407 (0.179/0.179)
		strong	0.417 (0.122/0.123)	0.404 (0.168/0.167)	0.397 (0.184/0.177)
		NI	0.410 (0.125/0.127)	0.002 (0.156/0.154)	-0.002 (0.164/0.166)
	1.5	weak	0.184 (0.121/0.124)	0.403 (0.166/0.168)	0.411 (0.182/0.186)
		strong	0.408 (0.123/0.124)	0.407 (0.169/0.176)	0.415 (0.186/0.185)
		NI	0.411 (0.125/0.129)	0.003 (0.156/0.161)	-0.005 (0.166/0.168)
0.6	1.2	weak	0.180 (0.089/0.093)	0.411 (0.136/0.133)	0.414 (0.135/0.143)
		strong	0.409 (0.091/0.096)	0.414 (0.138/0.136)	0.407 (0.139/0.139)
		NI	0.407 (0.093/0.095)	0.004 (0.133/0.129)	-0.001 (0.128/0.129)
	1.5	weak	0.184 (0.090/0.096)	0.412 (0.136/0.137)	0.408 (0.136/0.144)
		strong	0.406 (0.092/0.095)	0.404 (0.138/0.137)	0.406 (0.140/0.139)
		NI	0.408 (0.093/0.095)	-0.001 (0.133/0.139)	0.002 (0.128/0.126)

Table D.7: The mean bias in and the mean square error (MSE) of estimated log OR of covariate X (β_1) and genotype G (β_2) in the association model and the estimated MAF (θ) using the family-supplemented weighted empirical likelihood method (FS-WEL) and the naive method based on 1000 simulations. The prevalence is 0.03, genotype availability is 0.8 and 0.6, the true value of β_2 is 0.182 and 0.405, and MAF is 0.2. The true values of $(\alpha_3, \alpha_4, \alpha_5)$ in the three missingness models are: weak $= (0.182, 0.405, 0.405)$, strong $= (0.405, 0.405, 0.405)$, and No interaction (NI) $= (0.405, 0, 0)$. In each of the 12 settings, true values and estimates of coefficients β_1, β_2 and θ are presented in this order in the magnitude of 10^{-3}.

Model				Naive Method		FS-WPL Method	
$P(R=1)$	e^{β_2}	Missing	True	Bias[a]	MSE	Bias	MSE
0.8	1.2	weak	182	70.375	10.634	-1.484	4.325
			182	66.544	6.952	-0.919	2.669
			500	4.121	0.091	-4.955	0.105
		strong	182	67.080	9.962	-1.729	4.412
			182	61.865	6.395	-0.154	2.697
			500	15.142	0.307	-4.934	0.109
		NI	182	16.610	5.638	0.217	4.015
			182	38.474	4.203	2.395	2.856
			500	13.799	0.274	-5.146	0.109
	1.5	weak	182	69.026	10.241	-0.304	4.260
			405	65.704	6.753	-3.477	2.486
			500	-0.985	0.077	-9.807	0.176
		strong	182	66.699	9.686	-1.199	3.995
			405	66.490	7.217	0.892	2.587
			500	9.217	0.165	-10.431	0.192
		NI	182	21.966	6.250	3.630	4.406
			405	33.733	3.781	-3.053	2.614
			500	9.023	0.161	-9.923	0.175
0.6	1.2	weak	182	140.213	26.990	-1.713	4.398
			182	136.738	22.004	-1.427	3.288
			500	12.732	0.257	-4.788	0.124
		strong	182	134.456	25.394	-2.097	4.407
			182	122.709	18.591	-1.605	3.365
			500	35.114	1.338	-4.597	0.131
		NI	182	24.989	7.967	0.230	4.071
			182	51.520	6.623	1.195	3.494
			500	33.260	1.211	-4.896	0.127
	1.5	weak	182	140.514	27.382	-0.416	4.310
			405	132.061	20.802	-4.365	3.206
			500	8.218	0.169	-9.628	0.191
		strong	182	128.610	23.314	-1.694	3.977
			405	124.718	18.915	0.635	3.159
			500	29.057	0.946	-10.227	0.207
		NI	182	26.604	8.726	3.285	4.425
			405	49.575	6.249	-3.828	3.214
			500	28.185	0.898	-9.964	0.193

[a] Bias is defined as the mean estimate minus the true value

BIBLIOGRAPHY

Agresti, A and Kateri, M (2011). *Categorical data analysis*. Springer.

Allison, PD (2000). Multiple imputation for missing data: A cautionary tale. *Sociological methods & research* 28.3, 301–309.

Anderson, CD, Nalls, MA, Biffi, A, Rost, NS, Greenberg, SM, Singleton, AB, Meschia, JF, and Rosand, J (2011). The Effect of Survival Bias on Case-Control Genetic Association Studies of Highly Lethal DiseasesClinical Perspective. *Circulation: Cardiovascular Genetics* 4.2, 188–196.

Aschengrau, A and Seage, GR (2013). *Essentials of epidemiology in public health*. Jones & Bartlett Publishers.

Breslow, NE and Cain, KC (1988). Logistic regression for two-stage case-control data. *Biometrika* 75, 11–20.

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

Breslow, N and Day, N (1980). *Statistical methods in cancer research: the analysis of case-control studies*. Lyon, France: International Agency for Research on Cancer.

Chard, T (1991). Frequency of implantation and early pregnancy loss in natural cycles. *Bailliere's clinical obstetrics and gynaecology* 5.1, 179–189.

Chen, J, Pee, D, Ayyagari, R, Graubard, B, Schairer, C, Byrne, C, Benichou, J, and Gail, MH (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.

Chen, L, Weinberg, CR, and Chen, J (2016). Using family members to augment genetic case–control studies of a life-threatening disease. *Statistics in medicine* 30, 2815–2830.

Chiu, C-J, Conley, YP, Gorin, MB, Gensler, G, Lai, C-Q, Shang, F, and Taylor, A (2011). Associations between genetic polymorphisms of insulin-like growth factor axis genes and risk for age-related macular degeneration. *Investigative ophthalmology & visual science* 52.12, 9099–9107.

Falcone, GJ, Biffi, A, Devan, WJ, Brouwers, HB, Anderson, CD, Valant, V, Ayres, AM, Schwab, K, Rost, NS, Goldstein, JN, et al. (2013). Burden of Blood Pressure-Related Alleles Is Associated With Larger Hematoma Volume and Worse Outcome in Intracerebral Hemorrhage. *Stroke* 44.2, 321–326.

Fithian, W and Hastie, T (2014). Local case-control sampling: efficient sampling in unbalanced datasets. *Annals of Statistics* 42, 1693–1724.

Gail, MH, Brinton, LA, Byar, DP, Corle, DK, Green, SB, Schairer, C, and Mulvihill, JJ (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *JNCI: Journal of the National Cancer Institute* 81.24, 1879–1886.

Haneuse, S, Saegusa, T, and Lumley, T (2011). osDesign: an R package for the analysis, evaluation, and design of two-phase and case-control studies. *Journal of Statistical Software* 43.

Haneuse, S and Chen, J (2011). A multiphase design strategy for dealing with participation bias. *Biometrics* 67.1, 309–318.

Haneuse, S, Bogart, A, Jazic, I, Westbrook, EO, Boudreau, D, Theis, MK, Simon, GE, and Arterburn, D (2016). Learning about missing data mechanisms in electronic health records-based research: a survey-based approach. *Epidemiology (Cambridge, Mass.)* 27.1, 82.

Holcroft, C and Spiegelman, D (1999). Design of validation studies for estimating the odds ratio of exposure-disease relationships when exposure is misclassified. *Biometrics* 55, 1193–1201.

Horsfall, LJ, Nazareth, I, and Petersen, I (2012). Cardiovascular events as a function of serum bilirubin levels in a large statin-treated cohort. *Circulation* 126, 2556–2564,

Jennrich, RI (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics* 40.2, 633–643.

Lacour, RA, Westin, SN, Meyer, LA, Wingo, SN, Schorge, JO, Brooks, R, Mutch, D, Molina, A, Sutphen, R, Barnes, M, et al. (2011). Improved survival in non-Ashkenazi Jewish ovarian cancer patients with BRCA1 and BRCA2 gene mutations. *Gynecologic oncology* 121.2, 358–363.

Lawless, JF, Kalbfleisch, JD, and Wild, CJ (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society, Series B* 61, 413–438.

Liew, Z, Olsen, J, Cui, X, Ritz, B, and Arah, OA (2015). Bias from conditioning on live birth in pregnancy cohorts: an illustration based on neurodevelopment in children after prenatal exposure to organic pollutants. *International journal of epidemiology* 44.1, 345–354.

Manski, CF and McFadden, D (1981). Alternative estimators and sample designs for discrete choice analysis. *Structural Analysis of Discrete Data with Econometric Applications*, 2–50.

McIsaac, MA and Cook, R (2014). Response-dependent two-phase sampling designs for biomarker studies. *Canadian Journal of Statistics* 42, 268–284.

McNamee, R (2005). Optimal design and efficiency of two-phase case-control studies with error-prone and error-free exposure measures. *Biostatistics* 6, 590–603.

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

O'Brien, KM, Shi, M, Sandler, DP, Taylor, JA, Zaykin, DV, Keller, J, Wise, AS, and Weinberg, CR (2016). A family-based, genome-wide association study of young-onset breast cancer: inherited variants and maternally mediated effects. *European Journal of Human Genetics* 24.9, 1316.

Qin, J and Lawless, J (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* 22, 300–325.

Reilly, M (1996). Optimal sampling strategies for two-stage studies. *American Journal of Epidemiology* 143, 92–100.

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

Rubin, DB (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association* 91.434, 473–489.

Satten, GA and Kupper, LL (1993). Inferences about exposure-disease associations using probability-of-exposure information. *Journal of the American Statistical Association* 88.421, 200–208.

Scott, AJ and Wild, CJ (1991). Fitting regression models in stratified case-control studies. *Biometrics* 47, 497–510.

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

White, JE (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.

Wild, P, Andrieu, N, Goldstein, AM, and Schill, W (2008). Flexible two-phase studies for rare exposures: feasibility, planning and efficiency issues of a new variant. *Epidemiologic Perspectives & Innovations* 5.1, 4.

Williams, P, Pendyala, L, and Superko, R (2011). Survival bias and drug interaction can attenuate cross-sectional case-control comparisons of genes with health outcomes. An example of the kinesin-like protein 6 (KIF6) Trp719Arg polymorphism and coronary heart disease. *BMC medical genetics* 12.1, 42.

Zhu, Y, Mendola, P, Albert, PS, Bao, W, Hinkle, SN, Tsai, M, et al. (2016). Longitudinal study of insulin-like growth factor 1 and binding proteins 2 and 3 and subsequent risk of gestational diabetes among women in a multiracial cohort. *Diabetes* 65, 3495–3504.