Published in final edited form as:

Biometrics. 2020 March ; 76(1): 210–223. doi:10.1111/biom.13140.

Novel Two-Phase Sampling Designs for Studying Binary Outcomes

Le Wang^{1,2,*}, Matthew L Williams³, Yong Chen¹, Jinbo Chen^{1,**}

¹Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA

²Department of Mathematics and Statistics, Villanova University, Villanova, PA 19085, USA

³Division of Cardiovascular Surgery, Department of Surgery, University of Pennsylvania, Philadelphia, PA 19104, USA

Summary:

In a biomedical cohort study for assessing the association between an 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. When the outcome is binary, one may adopt a case-control sampling design or a balanced case-control design where cases and controls are further matched on a small number of complete discrete covariates. While the latter achieves success in estimating odds ratio (OR) parameters for the matching covariates, similar two-phase design options have not been explored for the remaining covariates, especially the incompletely collected ones. This is of great importance in studies where the covariates of interest cannot be completely collected. To this end, assuming that an external model is available relating the outcome and complete covariates, we propose a novel sampling scheme that oversamples cases and controls with worse goodness-of-fit based on the external model and further matches them on complete covariates similarly to the balanced design. We develop a pseudolikelihood method for estimating OR parameters. Through simulation studies and explorations in a real cohort study, we find that our design 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.

Keywords

Goodness-of-fit; Odds ratio estimation; Pseudolikelihood; Relative efficiency; Two-phase design

* l.wang@villanova.edu. ** jinboche@mail.med.upenn.edu.

Supporting Information

Web Appendix and Supplementary Figure and Tables referenced in Sections 4, 5 and 7, and the sample code are available with this article at the Biometrics website on Wiley Online Library. The R package “TwoPhaseReg” referenced in section 5 is available on github - <https://github.com/dragontaoran/TwoPhaseReg>.

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 of 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 are 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 the case where sampling is 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 the sampling proportion within each stratum have to be determined a priori. Besides “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.

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 computational 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 a 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 can be nearly as efficient as if the full data were used even under model mis-specification.

Our current work started with an investigation of 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. We found that GOF has a unique advantage of increasing efficiency for estimating OR parameters for the incomplete phase II covariates. Comparing the efficiency of GOF and balanced sampling motivated us to propose a new two-phase sampling scheme, balanced goodness-of-fit based sampling (BGOF), which performs GOF sampling first and then balanced sampling from 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 the paper is organized as follows. In Section 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 3. In Section 4, we describe the sampling and estimation procedures of BGOF and compare its efficiency with the other three designs. In Section 5, we assess the efficiency of GOF and BGOF when an alternative method is adopted for analysis (Tao et al., 2017; Zeng and Lin, 2014). In Section 6, we demonstrate implementation and efficiency gain of GOF and BGOF using data from a study of 30-day all-cause hospital readmission for patients who underwent coronary bypass graft surgery. We make final remarks and discuss future extensions in Section 7.

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 X denote phase I covariates that are available for all subjects, and 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 X and Z ,

$$\text{logit}P(Y = 1 | x, z) = x\beta_1 + z\beta_2, \quad (1)$$

where β_1 and β_2 are the OR parameters of interest. Note here that to simplify notation, a variable with value equal to one is implicitly included in X , with the corresponding regression coefficient in β_1 being the intercept parameter. Let β denote the vector of all parameters $(\beta_1^T, \beta_2^T)^T$. The outcome and phase I variables (Y, 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 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 the completely collected data (Y, X) . In a balanced design, discrete sampling strata are defined based on Y and X .

2.1 A Brief Review of the LCC Sampling Design

In the LCC design (Fithian and Hastie, 2014), the regression variables Y , X , and Z are fully observed on all N subjects. A preliminary model that is in the same form as equation (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 $P^{le}(Y=1|x, z)$, where superscripts “l” and “e” represent “local” and “external” to the study dataset, respectively. This design selects subjects with probability

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

That is, a Bernoulli experiment is conducted for each of the N subjects, with the success probability for subject i equal to $|y_i - P^{le}(Y=1|x_i, z_i)|$, $i = 1, 2, \dots, N$. 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 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 Z on a subset of subjects selected based only on (Y, X) . In our extension, we assume that an external model exists for relating Y to phase I covariates X , i.e.,

$$\text{logit} P^e(Y=1|x) = x\eta_1,$$

where the parameters η_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 X can be involved depending on the value of η_1 . We similarly define a quantity $S(y, x)$ that measures the difference between Y and $P^e(Y=1|x)$, i.e.,

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

Let R denote whether a phase I subject is selected into phase II, with $R=1$ indicating selection and $R=0$ non-selection, and $V = \{i: R_i=1, i=1, \dots, N\}$ 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 $S(y, x)$ 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 $P(R=1/y, x)$ to be $S(y, x)$ multiplied by a constant $c_1(c_1 > 0)$ for cases and $c_0(c_0 > 0)$ for controls, i.e.,

$$P(R = 1 | Y = 1, \mathbf{x}) = \min\{1, c_1 S(1, \mathbf{x}_1)\},$$

$$P(R = 1 | Y = 0, \mathbf{x}) = \min\{1, c_0 S(0, \mathbf{x}_1)\}.$$

Because function $S(y, \mathbf{x})$ informs the goodness-of-fit of the external model $P^e(Y = 1 | \mathbf{x})$, we term our extension to LCC as the goodness-of-fit based sampling, with the notion of “goodness-of-fit” further explored in Section 3.

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 $P(R = 1 | y, \mathbf{x})$ 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,

$$P(R = 1 | y, \mathbf{x}) = \min\{1, c S(y, \mathbf{x})\}.$$

This is equivalent to sampling in two steps. First a subset V is selected using sampling probabilities $S(y, \mathbf{x})$. 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 of 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:

$$P(R = 1 | Y = 1, \mathbf{x}) = 1,$$

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

2.3 Statistical Inference for the GOF Sampling Design

Because GOF selects phase II subjects based on both Y and $\mathbf{X}$, fitting model (1) to phase II data naively ignoring the sampling scheme can lead to biased estimates of parameters β . Note that the probability of observing a case given covariates $(\mathbf{X}, \mathbf{Z})$ in phase II is

$$P(Y = 1 | \mathbf{x}, \mathbf{z}; R = 1) = \frac{P(Y = 1 | \mathbf{x}, \mathbf{z}) P(R = 1 | Y = 1, \mathbf{x})}{P(R = 1 | \mathbf{x}, \mathbf{z})}. \quad (2)$$

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

$$\text{logit} P(Y = 1 | \mathbf{x}, \mathbf{z}; R = 1) = \mathbf{x}\beta_1 + \mathbf{z}\beta_2 + \alpha(\mathbf{x}). \quad (3)$$

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

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

where $\mathbf{w}_i = (\mathbf{x}_i, \mathbf{z}_i)$. We note that this is a pseudolikelihood 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 o set term $\alpha(\mathbf{x})$ is a specified function. Parameters $\boldsymbol{\beta}$ can be solved from the pseudolikelihood score equation

$$U(\boldsymbol{\beta}) \equiv \partial \ell(\boldsymbol{\beta}) / \partial \boldsymbol{\beta} = \sum_{i=1}^m (y_i - \mu_i^g) \mathbf{w}_i^T = 0, \quad (4)$$

with $\mu_i^g = \text{expit}\{\mathbf{w}_i \boldsymbol{\beta} + \alpha(\mathbf{x}_i)\}$. Note that $E\{U(\boldsymbol{\beta})\} = 0$ because $E(Y|\mathbf{w}_i, R_i = 1) = \mu_i^g$.

Therefore, by standard Z-estimation theory, the estimated coefficients $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normally distributed, i.e.,

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

with $V = \text{diag}\{\mu_i^g(1 - \mu_i^g)\}$ and $\mathbf{W}$ 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 $P(Y = 1|\mathbf{x}, \mathbf{z}; R = 1)$. This suggests that $\hat{\boldsymbol{\beta}}$ and its asymptotic variance estimates can be directly obtained from standard statistical software by fitting logistic regression model (1) to phase II data with the o set term $\alpha(\mathbf{x})$. This is the same case for LCC.

2.4 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, $\mathbf{X} = (X_1, X_2, X_3)$. 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 of 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

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

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 of size 10^6 , which served as the “external data” to fit a logistic regression model for Y given only X , i.e.,

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

The fitted model $P^e(Y = 1 | \mathbf{x})$ was then treated as the known external model and was used in all simulation iterations. $S(y, \mathbf{x})$ was calculated for each of the 3000 subjects. We included all cases in phase II and sampled controls with probability $\min\{1, cS(0, \mathbf{x})\}$. The constant c was chosen such that 2–2.5 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 mean estimates of the log OR parameters, and the empirical variances and mean asymptotic variances of $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$.

Figure 1 presents the mean asymptotic variance of $\hat{\beta}_4$ (Panels 1a and 1b), the estimated log OR of phase II covariate Z , and $\hat{\beta}_1$ (Panels 1c and 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 1c).

However, BD achieved smaller variance when the phase I covariates were moderately correlated (Panel 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. We defer detailed numerical results to Table 2.

3. Insight into the Efficiency of GOF

Figure 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 2.4 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 $\mathcal{S}(y, \mathbf{x})$ 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 $\mathcal{S}(y, \mathbf{x})$ 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 .

Because $\mathcal{S}(y, \mathbf{x})$ indicated the goodness-of-fit of the external model $P^e(Y = 1|\mathbf{x})$, we used the standardized Pearson residuals to investigate the nature of 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 $P^e(Y = 1|\mathbf{x}_i)$. If the external model $P^e(Y = 1|\mathbf{x})$ were the true model, the standardized Pearson residual for the i th 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 i th subject's estimated leverage. Larger absolute values of d indicates worse goodness-of-fit of $P^e(Y = 1|\mathbf{x})$.

For illustration purposes, we present in Figure 2 results from 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 2a and 2b demonstrate the relationship between $|d|$ and the sampling probability $P(R = 1|y, \mathbf{x})$ in GOF. As $P(R = 1|y, \mathbf{x})$ 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 2c and 2d display the distribution of the empirical mean $|d|$ in 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 (1) was the *true* model, lack of fit based on P^e would be suggestive of the necessity to include 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 .

4. Balanced Goodness-of-Fit Based Design

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 remaining covariates, 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 $V = \{i: R_i = 1, i = 1, \dots, N\}$ 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 . To do this, discrete strata L are formed based on phase I variables 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 subjects from cell $Y = y$ and $L = l$ through a Bernoulli experiment with probability m_{yl}/M_{yl} ($m_{yl} < M_{yl}$). 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.

4.1 Statistical Inference for the BGOF Design

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, heuristically, BGOF can be seen as a balanced design in cohort V which is randomly generated from the distribution $P(Y = 1 | x, z, R = 1)$ as defined in equation (3). Therefore, motivated by statistical inference methods for standard two-phase designs (Breslow and Cain, 1988; Scott and Wild, 1991), we propose to estimate OR parameters β by solving an estimating equation that takes the same form as (4) but with μ_i^g replaced by μ_i^l :

$$\mu_i^l = \expit\left\{w_i\beta + o(x_i) + \log \frac{\pi_{1l}}{\pi_{0l}}\right\},$$

where the i th subject is in stratum l . Similar to the estimation under GOF, the corresponding estimator $\hat{\beta}$ can be obtained using standard software for fitting logistic regression models with offset term $\alpha(x_i) + \log(\pi_{1l}/\pi_{0l})$ for the i th observation in stratum l . We present in the

corollary below the large sample properties of $\hat{\beta}$ and present the proof in APPENDIX. Let E_{yl}^* and E_l^* denote expectations with respect to the distribution of $\mathbf{W}$ in cell (y, l) and stratum l , respectively. Let ψ_{yl} denote the large sample limit of M_{yl}/N as $N \rightarrow \infty$, γ_{yl} that of m_{yl}/m as $m \rightarrow \infty$, and θ that of m/N as $N \rightarrow \infty$.

Corollary: Suppose that the true value β lies inside a compact space, and that component of $(\mathbf{X}, \rightarrow \infty)$ are bounded. Define matrices H and G as

$$H \equiv \sum_{y=0}^1 \sum_{l=1}^L E_{yl}^* \{ \mathbf{w}_i^T \mathbf{w}_i \mu_i^l (1 - \mu_i^l) \}$$

$$G \equiv \sum_{y=0}^1 \sum_{l=1}^L (\gamma_{yl}^{-1} - \theta \psi_{yl}^{-1}) E_l^* \{ \mathbf{w}_i^T \mu_i^l (1 - \mu_i^l) \} E_l^* \{ \mathbf{w}_i \mu_i^l (1 - \mu_i^l) \}.$$

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

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

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

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

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

The asymptotic variance of $\hat{\beta}$ differs 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 between 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 et al., 2011) for the standard two-phase prospective design (option “cohort=TRUE”) can be used directly for obtaining variance estimates for $\hat{\beta}$, with the GOF subsample treated as phase I sample. In other words, although GOF generates an outcome-dependent subsample because the sampling probability depends also on Y , in terms of variance estimation, it can be safely treated as a prospective sample. As explained in APPENDIX, the key to this interesting result is that the pseudo-model for the GOF subsample, $P(Y=1|\mathbf{x}, z; R=1)$, is a genuine probability function.

4.2 Simulation Studies

We conducted simulation studies to compare the efficiency of BGOF, GOF, BD, and CC using the same setup as above for GOF. The results on estimation of β_4 are presented in Table 1. Under all four 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.

Table 2 presents the mean asymptotic variance of $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$ under BGOF and its efficiency relative to the other three 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. Compared with the full cohort, BGOF lost 40–50% efficiency for estimating the OR for phase II covariate due to incomplete measurement (Table 1), and the extent of efficiency loss was similar for phase I covariates (data not shown). For comparison purposes, we used the binary variable X_3 as the stratifying variable in BGOF and BD, and the results were largely similar and are presented in Supplementary Table 1.

Table 2 showed that the correlation among phase I variables might affect the efficiency of BGOF relative to BD, that is, the relative efficiency decreased with the correlation. Then a question arose as to 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 X with the remaining setup identical to that in Section 2.4. 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., $\rho=0.3$ (Supplementary Table 2). 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 up efficiency of BGOF for estimating its OR relative to BD, but it did not help (Supplementary Table 3).

Lastly, because GOF and BGOF requires 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 external model reflects the true relationship between Y and X in the data. To this end, we deliberately changed η_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 (supplementary figure 1). It turned

out that although the predicted sampling probabilities calculated based on the “inaccurate” external model deviated from the true probabilities, the relative magnitude of the sampling probabilities still maintained. That is, individuals who had greater sampling probabilities calculated based on the true external model were still more likely to be selected based on the “inaccurate” external model. The efficiency advantage of GOF and BGOF may decrease as the external model becomes far less accurate.

5. Alternative Analytical Approaches

We considered alternative approaches to analyzing data arising from GOF and BGOF designs and evaluated their relative efficiency to CC and BD. Weighting individuals selected into phase II by the inverse of their sampling probabilities is an obvious option, but the statistical efficiency of this approach is generally lower for analyzing two-phase data (Breslow and Chatterjee, 1999). Recently, Tao and colleagues (2017) developed an efficient semiparametric maximum likelihood approach under general two-phase designs. They modeled the conditional probability of phase II covariates given phase I covariates in the likelihood function using B-spline sieve approximation and therefore named their approach the sieve maximum likelihood approach (“SMLE”). Compared with other maximum likelihood methods for analyzing two-phase data (Breslow and Holubkov, 1997; Scott and Wild, 1997) which can only handle discrete phase I covariates, SMLE has an important advantage of accommodating continuous phase I covariates, although the number of phase I covariates has to be necessarily small (Tao et al., 2017). When feasible, SMLE is expected to have higher efficiency than our pseudolikelihood method since it is fully efficient. In this section, we evaluated the relative efficiency of our proposed sampling designs compared with CC and BD when this maximum-likelihood approach is used for analysis.

We followed the same simulation setup as in Tao et al. (2017) and implemented SMLE using R package “TwoPhaseReg”. First, we generated three independent standard normal variables, U_1 , U_2 , and Z . Then we generated phase I variable X_1 as $X_1 = U_1 + 0.3Z$ and X_2 as $X_2 = U_2$, and Z was treated as the phase II variable. Under this setup, the correlation between X_1 and Z was 0.3, and X_2 was independent of Z . The binary outcome variable Y was generated from the logistic regression model

$$\text{logit}P(Y = 1 | X_1, X_2, Z) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 Z.$$

We set $\beta_1 = 0.6$, $\beta_2 = 0.5$, and β_3 equal to 0.5 or 0.9 corresponding to weak or strong effects of Z . The intercept β_0 was chosen to achieve outcome prevalence $P(Y = 1)$ of 5% or 10%. The phase II sampling procedures under the CC, BD, GOF, and BGOF schemes were the same as in Sections 2.4 and 4.2. We set the phase I cohort size to be 3000, and repeated the simulation 1000 times.

The results on estimation of β_3 using SMLE are presented in Supplementary Table 4. When analyzed by SMLE, BGOF and GOF still achieved the highest efficiency for estimating the OR for the phase II variable Z . BGOF was 21–26% more efficient than CC and 10–11% more efficient than BD. The relative efficiency of BGOF ranged from 0.94–1.44 compared with CC and 0.99–1.21 compared with BD for estimating β_1 . BGOF was 1–32% more

efficient than CC and 1–18% more efficient than BD for estimating β_2 (data not shown). The relative efficiency of GOF ranged from 0.96–1.28 compared with CC and 0.96–1.10 compared with BD for estimating β_1 , and ranged from 1.00–1.18 compared with CC and 1.00–1.17 compared with BD for β_2 . We also compared SMLE and the pseudolikelihood approach with respect to efficiency for estimating the OR parameters. SMLE yielded similar variances for the phase II variable, but reduced variances for phase I variables substantially by 32–91%.

6. Illustration of GOF and BGOF Using a Study of 30-day Hospital Readmission

We use data from a study of 30-day hospital readmission to illustrate the implementation of GOF and BGOF and to evaluate their performance in a real study setting. Data were acquired from the Pennsylvania Health Care Cost Containment Council (PHC4), an independent state agency that collects data on inpatient hospital discharge and ambulatory/outpatient procedure records from hospitals and freestanding ambulatory surgery centers in Pennsylvania. The study consisted of 4161 patients who underwent coronary bypass graft procedure in years 2014 and 2015, and 535 (13%) were readmitted within 30 days after hospital discharge. Short-term hospital readmission after major surgeries incurs substantial cost, and its rate is an important measure of the quality of care (Liao et al., 2008; Jencks et al., 2009). The ultimate goal is to develop a model for predicting individual level risk of readmission. In the study cohort, the median age was 67 years with an interquartile range of 59–74 years. Age was used as number of years greater than 65 at unit of 5 years in the model and age less than 65 was defined as 0. One thousand and nineteen patients (25%) had heart failure, 701 (17%) had peripheral vascular disease, 331 (8%) had cerebrovascular disease, and 256 (6%) had cardiomyopathy. Chronic kidney disease was treated as a three level ordinal variable with 148 (4%) patients having stage 5 or end-stage renal disease, 604 (14%) having stage 1–4 disease, and 3409 (82%) disease free. One hundred and twenty-six (3%) patients experienced chronic liver disease, 714 (17%) experienced mental disorder, and 514 (12%) experienced sleep apnea. Six hundred and five (15%) patients had high white blood cell count (WBC) ($\geq 11.0 \times 10^9/L$), 370 (9%) had high creatinine (≥ 1.5 mg/dL), and 378 (9%) had low hemoglobin (< 11.1 g/dL). We developed a preliminary prediction model using these variables as predictors, where $I()$ denotes the indicator function:

$$\begin{aligned} \text{logit}P(Y = 1) = & -2.77 + 0.036 \times \text{age} + 0.403 \times I(\text{chronic kidney disease}) + \\ & 0.499 \times I(\text{heart failure}) + 0.391 \times I(\text{mental disorder}) + \\ & 0.303 \times I(\text{peripheral vascular disease}) + 0.353 \times I(\text{high WBC}) + \\ & 0.313 \times I(\text{cerebrovascular disease}) - 0.447 \times I(\text{high creatinine}) + \\ & 0.356 \times I(\text{low hemoglobin}) + 0.541 \times I(\text{sleep apnea}) - \\ & 0.514 \times I(\text{chronic liver disease}) + 0.549 \times I(\text{cardiomyopathy}). \end{aligned}$$

We performed GOF and BGOF sampling in the study cohort, and used age dichotomized at year 65 as the stratifying variable in BD and BGOF. We considered each predictor except age in the above model in turn as the phase II variable, which allowed us to evaluate the efficiency of GOF and BGOF in eleven settings with different phase II effect sizes and different correlation structure between phase I and phase II covariates. The external logistic regression models for sampling, including all phase I variables as predictors, were obtained

from an earlier PHC4 dataset that included 6953 patients who underwent coronary bypass graft procedure in years 2011–2013. The study investigator was interested in using more recent data in years 2014 and 2015 to develop the prediction model. We selected all 535 cases and the same number of controls into phase II under all four designs. For BGOF, in order to ensure the targeted number of controls, we selected roughly twice as many controls as the cases into subset V . Therefore, we estimated the constant c_0 roughly as the ratio of the targeted number of candidate controls 1070 over the empirical mean number of controls, obtained as the sum of the sampling probability P^* for all the controls in the full cohort. Because results from one-time subsampling cannot accurately reflect the performance of each study design, we bootstrapped phase I cohort 1,000 times, and repeated the above steps in each iteration.

Table 3 presents the estimated log OR parameters for phase II variables averaged over 1,000 bootstrapped samples in the eleven settings. The mean estimates were all close to the estimates obtained from the full cohort analysis as in the model above, and the coverages were all around 0.95. The mean relative efficiency for both GOF and BGOF was 0.99–1.22 compared with CC, and 1.01–1.19 compared with BD. We did not observe an increasing trend in the relative efficiency with respect to increasing effect sizes of phase II covariates, probably because the difference in log OR parameters was small in this study such that the correlation between phase I and phase II covariates played a more important role in relative efficiency. Phase II covariates that were more strongly correlated with phase I variables achieved greater efficiency gain under GOF and BGOF. For example, when high creatinine was treated as the phase II covariate, the highest four correlation coefficient with phase I covariates ranged from 0.2 to 0.55. The largest correlation coefficient between chronic liver disease and corresponding phase I covariates was 0.09 with all other pairwise correlation coefficients less than 0.03. These results were consistent with our simulation studies, where the efficiency gain by GOF and BGOF became much smaller when phase I and phase II variables were independent (Supplementary Table 2).

7. 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 and BGOF, we offer a new perspective to define “informative” subjects for efficient sampling, especially with respect to phase II covariates. In contrast, 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. Our proposed GOF and BGOF designs rely on an existing external model for sampling. It does not require that the external model conform to a similar model for the outcome and phase I variables in the study cohort, although the efficiency is expected to be higher if this happens to be the case.

The advantage of BGOF lies in the fact that it makes full use of all phase I covariates to determine the sampling probability without coarsening. Of course, instead of four sampling

strata considered in the simulation studies, sampling strata for BD can also be based on composite categories created based on multiple phase I variables. As suggested by one reviewer, we compared the designs where the strata for BD was defined based on all three phase I covariates. Specifically, using the same simulation setup as in Section 2.4, we dichotomized X_1 and X_2 at their respective medians, thereby creating 16 phase I sampling strata based on all three phase I variables and the outcome variable. We needed to increase phase I cohort size substantially to ensure a sufficient number of subjects in each of the 16 phase I strata (Supplementary table 5). BGOF remained the most efficient for estimating ORs of phase II covariates (Supplementary table 6), but the efficiency advantage relative to BD decreased as more phase I covariates were used to form the sampling strata in BD. Indeed, if phase I sample size is sufficiently large to allow full stratification on all phase I covariates, BD and BGOF will have similar statistical efficiency. However, how finely phase I data can be stratified is often constrained by the number of cases, because the number of cases within each sampling stratum should not be smaller than 5 (rule of thumb) in BD. In practice it is often impractical to use all phase I variables to form the sampling strata due to difficulty in finding a good stratification strategy or insufficient number of individuals in certain sampling strata. Therefore, we presented results when phase I strata were stratified only on X_1 , with X_2 and X_3 representing non-stratifying phase I variables. This allowed us to compare efficiency of different study designs with respect to both stratifying and non-stratifying phase I variables. The efficiency advantage and feasibility render BGOF an attractive sampling method for reducing study cost while maintaining statistical efficiency. BGOF was designed for sampling in typical epidemiological study cohorts, but is also expected to be useful for studies of small sample sizes.

We used the pseudolikelihood method for analysis of data collected under BGOF and BD. The SMLE method (Tao et al., 2017) can be a more efficient alternative when the number of phase I covariates is small. A key reason for this efficiency difference lies in differential usage of phase I data. When the pseudolikelihood method is applied, phase I data only for those who are selected into phase II are used in analysis. The higher efficiency of BGOF remains under SMLE in simulation studies. Intuitively, GOF and BGOF are designed to utilize phase I information to select more informative phase II individuals, such that the design efficiency does not rely on the adopted analytical approaches. Balanced design also enjoys this same property. In contrast, a design that is derived assuming a given analytical method, e.g., the “optimal” design (McNamee, R, 2005) mentioned earlier, may not maintain high efficiency should alternative analytical methods be adopted. We recognize that the pseudolikelihood approach has its own advantages over SMLE. GOF and BGOF are the most useful when there are many phase I covariates, when SMLE may not be applicable (Tao et al., 2017). SMLE can potentially accommodate more phase I variables if some variables are assumed to be independent. The method would enjoy full efficiency and be more efficient than the pseudolikelihood method if such independence holds, but the estimates would be biased otherwise. Readers can choose the most appropriate method for analysis given specific studies. We plan to evaluate the relative performance of MLE and the pseudolikelihood method in our future work particularly in the setting of GOF/BGOF.

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 appealing to use internal phase I

data to derive such a model. In contrast to our current sampling schemes where the sampling probability $\mathcal{S}(y_i, x_i)$ only depends on each subject's own data, by using the internal preliminary model, sampling across different subjects becomes dependent. That is, the probability that a phase I subject gets selected into phase II depends on other subjects' data, which can potentially be accommodated by modification on the large sample theory. For a preliminary investigation of this extension of BGOF, we performed additional simulation studies using the same setup as in Section 2.4, except that the full phase I cohort was used to derive the model for sampling. The pseudolikelihood estimates under both GOF and BGOF sampling designs seemed unbiased (Supplementary table 7), and both designs appeared to have similar or slightly higher efficiency than when an external preliminary model was used (Supplementary table 8). Note that SMLE is also applicable when such internal sampling model is adopted. We would like to investigate this potential extension in our future work. In the current 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.

Lastly, in this work, the external information regarding phase I covariates was utilized merely for sampling, and was involved in estimation only through the offset term in the pseudolikelihood function. A reviewer suggested that the external model be potentially used as constraints to improve estimation efficiency along the research line of Chatterjee et al. (2016). These methods assume that the external information regarding the relationship between the outcome and phase I covariates is consistent with that in the study cohort. If violated, the estimation will be biased. When such consistency is plausible, it is attractive to make further use of the external information to improve estimation efficiency for analyzing GOF and BGOF data. Nevertheless, these methods are likelihood-based and developed for analyzing cohort or case-control data. They cannot yet accommodate data collected through two-phase sampling. It will be interesting to extend our work along this line.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by NIH grants R01-ES016626. The authors would like to thank Dr. Xinglei Chai for her help with the computational implementation of the balanced design in R and Ms. Jill Schnall for her help with processing the dataset for the 30-day all-cause readmission study. The authors also would like to thank the associate editor and two reviewers of Biometrics for their insightful comments and suggestions on this work. *Conflict of Interest:* None declared.

References

- Agresti A and Kateri M (2011). Categorical data analysis. Springer.
- 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.
- 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.
- Chatterjee N, Chen Y-H, Maas P, and Carroll RJ (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association* 111, 107–117. [PubMed: 27570323]
- Fithian W and Hastie T (2014). Local case-control sampling: efficient sampling in unbalanced datasets. *Annals of Statistics* 42, 1693–1724. [PubMed: 25492979]
- 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,.
- 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. [PubMed: 11315067]
- Jencks SF, Williams MV, and Coleman EA (2009). Rehospitalizations among patients in the medicare fee-for-service program. *New England Journal of Medicine* 360, 1418–1428. [PubMed: 19339721]
- Liao L, Allen LA, and Whellan DJ (2008). Economic burden of heart failure in the elderly. *Pharmacoeconomics* 26, 447–462. [PubMed: 18489197]
- 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. [PubMed: 15860543]
- Neyman J (1938). Contribution to the theory of sampling from human populations. *Journal of the American Statistical Association* 33, 101–116.
- Reilly M (1996). Optimal sampling strategies for two-stage studies. *American Journal of Epidemiology* 143, 92–100. [PubMed: 8533752]
- Scott AJ and Wild CJ (1991). Fitting regression models in stratified case-control studies. *Biometrics* 47, 497–510.
- Tao R, Zeng D, and Lin D-Y (2017). Efficient semiparametric inference under two-phase sampling, with applications to genetic association studies. *Journal of the American Statistical Association* 112, 1468–1476. [PubMed: 29479125]
- 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. [PubMed: 7055123]
- 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, 4. [PubMed: 18828892]
- Zeng D and Lin D (2014). Efficient estimation of semiparametric transformation models for two-phase cohort studies. *Journal of the American Statistical Association* 109, 371–383. [PubMed: 24659837]

Figure 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 is 3000 with $P(Y=1) = 0.05$, and the true value of β_4 is between 0.5–0.9. Phase I variables are uncorrelated in panels a and c and modestly correlated in panels b and d.

Figure 2:

Insights into the relative efficiency of GOF. Panels a and b show the relationship between $|d|$ and $P(R = 1/y, x)$ 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 is 2×10^4 with $P(Y = 1) = 0.05$, and the log OR of the phase II variable Z is 0.9.

Table 1:

The estimated log OR of phase II covariate ($\hat{\beta}_4$) under the 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 is 3000, the prevalence is 0.05 and 0.10, the correlation parameter ρ for phase I variables is 0 and 0.3, and the true value of β_4 is 0.5, 0.7, and 0.9. The mean asymptotic standard error (“asym”), empirical standard error (“emp”), coverage probability (“coverage”), and the mean asymptotic standard error based on the full cohort (“full”) of $\hat{\beta}_4$ are calculated based on 1000 simulations.

$P(Y = 1)$	ρ	β_4 (full)	BGOF (asym/emp)	coverage	GOF (asym/emp)	coverage	CC (asym)	BD (asym)
0.05	0	0.5 (0.15)	0.52 (0.20/0.21)	0.949	0.52 (0.21/0.21)	0.951	0.52 (0.23)	0.53 (0.22)
		0.7 (0.15)	0.70 (0.21/0.21)	0.949	0.72 (0.21/0.22)	0.948	0.71 (0.24)	0.72 (0.23)
		0.9 (0.16)	0.92 (0.22/0.21)	0.946	0.93 (0.22/0.22)	0.947	0.92 (0.25)	0.93 (0.24)
	0.3	0.5 (0.12)	0.52 (0.17/0.17)	0.955	0.52 (0.17/0.17)	0.948	0.51 (0.18)	0.52 (0.18)
		0.7 (0.12)	0.72 (0.17/0.18)	0.947	0.72 (0.17/0.18)	0.941	0.71 (0.19)	0.72 (0.19)
		0.9 (0.13)	0.92 (0.18/0.19)	0.945	0.93 (0.18/0.19)	0.935	0.93 (0.21)	0.93 (0.20)
0.10	0	0.5 (0.11)	0.51 (0.14/0.14)	0.950	0.51 (0.14/0.14)	0.951	0.51 (0.17)	0.51 (0.16)
		0.7 (0.11)	0.71 (0.15/0.14)	0.943	0.71 (0.15/0.15)	0.939	0.70 (0.16)	0.71 (0.16)
		0.9 (0.12)	0.91 (0.15/0.15)	0.954	0.91 (0.15/0.15)	0.953	0.91 (0.17)	0.91 (0.17)
	0.3	0.5 (0.09)	0.50 (0.12/0.12)	0.948	0.50 (0.12/0.12)	0.943	0.50 (0.13)	0.50 (0.13)
		0.7 (0.09)	0.70 (0.12/0.13)	0.938	0.71 (0.12/0.12)	0.951	0.71 (0.13)	0.71 (0.13)
		0.9 (0.10)	0.91 (0.13/0.13)	0.944	0.91 (0.13/0.13)	0.944	0.92 (0.14)	0.91 (0.14)

Table 2:

Asymptotic variance of $\hat{\beta}$ 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 is 3000. The prevalence is 0.05 and 0.10, the correlation parameter ρ for phase I variables is 0 and 0.3, and the true value of β_4 is 0.5, 0.7, and 0.9. Relative Efficiency (RE) is calculated as the mean asymptotic variance under each design over that of BGOF.

$P(Y = 1)$	ρ	β_4	$\text{Var}(\hat{\beta}_4)$	RE of BGOF vs.			$\text{Var}(\hat{\beta}_1)$	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	0.5	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
			$\text{Var}(\hat{\beta}_2)$	RE of BGOF vs.			$\text{Var}(\hat{\beta}_3)$	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	0.5	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

Table 3:

The estimates of the log OR parameters and asymptotic standard errors (SE) for phase II covariates averaged over 1000 bootstrapped samples in the 30-day all-cause readmission study under the balanced goodness-of-fit based design (BGOF) and the goodness-of-fit based design (GOF). Relative efficiency (RE) is calculated as the mean of asymptotic variance under CC or BD over that of BGOF or GOF. “Cohort” log OR parameters are obtained from fitting the logistic regression model on the full 2014–2015 dataset.

Name	Cohort	BGOF			GOF		
		Estimate (SE)	Coverage	RE vs. CC	Estimate (SE)	Coverage	RE vs. CC
				vs. BD			vs. BD
Cardiomyopathy	0.549	0.579 (0.244)	0.945	1.09	0.576 (0.243)	0.955	1.09
Sleep apnea	0.541	0.588 (0.189)	0.953	0.99	0.583 (0.189)	0.947	0.99
Chronic liver disease	0.514	0.556 (0.371)	0.954	1.00	0.555 (0.370)	0.944	1.00
Heart failure	0.499	0.529 (0.146)	0.938	1.07	0.532 (0.146)	0.938	1.07
High creatinine	-0.447	-0.467 (0.246)	0.953	1.22	-0.468 (0.247)	0.954	1.22
Chronic kidney disease	0.403	0.410 (0.142)	0.944	1.11	0.411 (0.142)	0.943	1.11
Mental disorder	0.391	0.401 (0.165)	0.950	1.02	0.394 (0.165)	0.948	1.02
Low hemoglobin	0.356	0.365 (0.212)	0.949	1.18	0.366 (0.212)	0.940	1.18
High WBC	0.353	0.338 (0.175)	0.948	1.09	0.348 (0.175)	0.955	1.08
Cerebrovascular disease	0.313	0.316 (0.217)	0.949	1.12	0.315 (0.218)	0.941	1.11
Peripheral vascular disease	0.303	0.310 (0.160)	0.943	1.10	0.309 (0.160)	0.940	1.10