Optimal sampling for design-based estimators of regression models

Tong Chen | Thomas Lumley

Department of Statistics, University of Auckland, Auckland, New Zealand

Correspondence
Tong Chen, Department of Statistics, University of Auckland, Auckland 1010, New Zealand.
Email: tche929@aucklanduni.ac.nz

Funding information
National Institute of Allergy and Infectious Diseases, Grant/Award Number: R01-AI131771; Patient-Centered Outcomes Research Institute, Grant/Award Number: R-1609-36207

Two-phase designs measure variables of interest on a subcohort where the outcome and covariates are readily available or cheap to collect on all individuals in the cohort. Given limited resource availability, it is of interest to find an optimal design that includes more informative individuals in the final sample. We explore the optimal designs and efficiencies for analyses by design-based estimators. Generalized raking is an efficient class of design-based estimators, and they improve on the inverse-probability weighted (IPW) estimator by adjusting weights based on the auxiliary information. We derive a closed-form solution of the optimal design for estimating regression coefficients from generalized raking estimators. We compare it with the optimal design for analysis via the IPW estimator and other two-phase designs in measurement-error settings. We consider general two-phase designs where the outcome variable and variables of interest can be continuous or discrete. Our results show that the optimal designs for analyses by the two classes of design-based estimators can be very different. The optimal design for analysis via the IPW estimator is optimal for IPW estimation and typically gives near-optimal efficiency for generalized raking estimation, though we show there is potential improvement in some settings.

KEYWORDS
generalized raking, influence function, model-assisted sampling, Neyman allocation, optimal design, residual, two-phase sampling

1 INTRODUCTION

In modern public health studies, routinely collected large databases, such as electronic health records (EHR), are increasingly used to study research questions of interest. However, variables within these databases can be error-prone. Without validation, directly analyzing EHR data may lead to invalid statistical inference. For these large databases, it will be prohibitively expensive to validate variables of interest for every record. A cost-effective strategy is to use a two-phase sampling design. At phase one, the outcome variable and several covariates (e.g., age, gender, and ethnicity) are collected or available for every individual in the cohort. The EHR databases can be used as the phase-1 sample. At phase two, variables of interest are only collected or validated for individuals selected in the phase-2 subsample.

Design-based methods produce robust estimations for fitting regression models to two-phase stratified sampling designs. Directly solving inverse-probability weighted (IPW) or Horvitz-Thompson type of likelihood functions leads to the IPW estimator. For analyzing EHR databases, the IPW estimator is not efficient as most of the information in the phase-1 sample has been ignored. Generalized raking is more efficient than the IPW estimator as it incorporates...
the whole cohort information in the analysis. The efficiency gains are achieved by adjusting sampling weights of the IPW estimator based on auxiliary information. Generalized raking estimators are closely connected with the augmented inverse-probability weighted (AIPW) estimators of Robins et al. Typically, generalized raking estimators are combined with imputations to leverage the whole cohort information in estimating regression coefficients under two-phase sampling designs. See Breslow et al for single imputation, and Oh et al and Han et al for multiple imputation.

An alternative estimation method is based on semiparametric maximum likelihood, which can be more efficient than design-based estimators if the model is correctly specified. However, Han et al showed generalized raking estimators can be more efficient than the semiparametric maximum likelihood estimator even under mild model misspecification.

The optimal sampling theory that is closely connected with design-based estimators has been studied in some previous literature. Reilly and Pepe derived a closed-form solution of the optimal phase-two sampling probabilities for the mean-score estimator. McIsaac and Cook extended the work using a multiwave sampling framework for binary and survival outcomes respectively. The optimal design for estimating regression coefficients from the IPW estimator is Neyman allocation applied to influence functions. For binary data and stratified sampling, the optimal design for analysis by the mean-score estimator is asymptotically identical to those for analysis by the IPW estimator. McIsaac and Cook also suggested the optimal design for analysis by the AIPW estimator can be derived numerically. If variables are all discrete, it is asymptotically equivalent to the optimal designs for analyses by the mean-score and IPW estimator. However, a closed-form solution of the optimal design for analysis via generalized raking is not known.

In this article, we derive a closed-form solution of the optimal design for analysis via generalized raking estimators. We then compare it with the optimal allocation for analysis by the IPW estimator and other commonly used sampling designs. It is noteworthy that sampling probabilities are assumed to be known in a two-phase design, so we do not consider doubly robustness. The rest of this article is organized as follows. Notations are defined in Section 2. The IPW estimator and generalized raking estimators are introduced in Section 3. In Section 4, we derive the optimal designs for analyses by the two classes of design-based estimators. Results of simulation studies are reported in Section 5. In Section 6, we further compare the proposed optimal design with other sampling strategies using the National Wilms’ Tumor Study (NWTS) dataset example. Remarks are made in Section 7. Code of numerical studies is available from https://github.com/T0ngChen/Opt_sampling_design_based. An interactive Shiny app, which compares the optimal designs for analyses by the two classes of design-based estimators, is available from https://tchen.shinyapps.io/raking.

## 2 Notation

Suppose we want to select \( n \) observations with stratified random sampling from a cohort of size \( N \) over \( K \) strata. Let \( N_j \) and \( n_j \) be the stratum size and phase-2 sample size for stratum \( j \) respectively. Let \( Y \) denote an outcome variable, \( Z \) denote phase-1 covariates, and \( A \) denote additional variables in the data. Variables \( Y, Z, \) and \( A \) are available for every individual in the phase-1 sample. Let \( X \) represent variables of interest which are only available for individuals selected in the phase-2 sample. Let \( R \) denote an indicator variable. If \( R_i = 1 \), subject \( i \) is selected in the phase-2 sample, otherwise \( R_i = 0 \). We assume the missingness of \( X \) only depends on phase-1 data, \( P(R|X,Y,A,Z) = P(R|Y,A,Z) \), so variables of interest \( X \) are missing at random. The phase-2 inclusion probability of individual \( i \) is \( E(R_i|Z_i,A_i,Y_i) = \pi_i \).

We describe \( P(X|A,Z,Y;\alpha) \) as the imputation model where \( \alpha \) are regression coefficients. \( \hat{X}_i \) is the imputed value of \( X \) for the \( i \)th observation. We further describe \( P(Y|X,Z;\beta) \) as the outcome model of interest where \( \beta \) are regression coefficients. \( h(\beta) \) are influence functions which are calculated based on the outcome model. Specifically, let \( \beta_i \) be the regression coefficient of \( X \) in the outcome model. Our target is to minimize the variance of \( \hat{\beta}_i \) by optimizing the design for design-based estimators.

In our notation, \( Y \) and \( Z \) are used to fit the outcome model of interest, whereas additional variables \( A \) are not included in the outcome model. There are various reasons that we do not include \( A \) in the outcome model. For example, \( A \) can be a surrogate variable of \( X \) in measurement-error settings; if \( X \) is in the outcome model, \( A \) will no longer be informative. Although additional variables \( A \) are not used to fit the outcome model, they still provide extra information about variables of interest \( X \) and the regression coefficients \( \beta \). The phase-1 variables \( Y, Z, \) and \( A \) can be used in improving the sampling design, generalized raking, and imputation.

We use the term “generalized raking” (instead of calibration) to avoid confusion, as there are many other techniques which use the term “calibration”. Let \( S \) be auxiliary raking variables that are used to adjust weights in generalized raking procedures and \( \beta^* \) be phase-1 estimates of \( \beta \). Specifically, \( \beta^* \) are estimated from the model \( P(Y|\hat{X},Z) \). In Section 3.2, we
discuss that if we want to estimate the total of $X$, auxiliary raking variables $S$ could include $Y$, $Z$, and $A$. If we want to estimate regression coefficients $\beta$, good choices of auxiliary raking variables $S$ are influence functions $h_i(\beta^*)$.

3  |  DESIGN-BASED ESTIMATORS

3.1  |  Inverse-probability weighted estimator

The IPW estimator\(^4\) can be derived by weighting each observation by the inverse of its sampling probability $\pi_i$. The IPW estimator for estimating regression coefficients $\beta$ can be obtained by solving the weighted score function

$$\sum_{i=1}^{N} R_i \frac{\partial}{\partial \beta} \log P(Y_i|X_i, Z_i; \beta) = 0. \quad (1)$$

The sampling probability $\pi_i$ in Equation (1) should be bounded away from zero so that every study subject should have a positive probability of being sampled at phase two. The IPW estimator is appealing because of its simplicity and robustness. However, it is not efficient as it ignores the phase-1 information.

3.2  |  Generalized raking estimator

Generalized raking estimators improve on the IPW estimator by adjusting sampling weights based on the auxiliary information. Suppose population totals of a vector of auxiliary raking variables $S$ are known in advance, and we want to estimate the population total $T_X = \sum_{i=1}^{N} X_i$. Generalized raking estimators are defined as $T_{S,r} = \sum_{i=1}^{N} R_i w_i X_i$, where $w_i$ are calibrated weights, and they depend on auxiliary raking variables $S$. The target is to minimize the total weight change $\sum_{i=1}^{N} R_i d(w_i, 1/\pi_i)$ with calibration constraints

$$\sum_{i=1}^{N} R_i w_i S_i = \sum_{i=1}^{N} S_i,$$

under a prespecified distance function $d(a, b)$. The optimization problem can be solved by Lagrange multipliers. Deville and Särndal\(^9\) provided a few example distance functions and showed all generalized raking estimators are asymptotically equivalent. Typically, distance function $d(a, b) = (a - b)^2 / 2b$ will lead to the generalized regression estimator (GREG).

In order to leverage generalized raking procedures, population totals or means of auxiliary raking variables should be known in advance. The efficiency of generalized raking estimators is influenced by auxiliary raking variables. Typically, the efficiency depends on the linear correlation between auxiliary raking variables and the estimator of the parameter of interest.\(^9,11,13\) We want to use generalized raking procedures to estimate regression coefficients so that good auxiliary raking variables should be both highly and linearly correlated with $\hat{\beta}$. An asymptotically linear estimator $\hat{\beta}$ satisfies

$$\sqrt{N}(\hat{\beta} - \beta) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} h_i(\beta) + o_p(1), \quad (2)$$

where $h_i(\beta)$ is a function of the $i$th observation. Breslow and Wellner\(^29\) derived a weighted version of Equation (2) for the IPW estimator. As the asymptotically linear estimator $\hat{\beta}$ can be approximated by the mean of influence functions, good choices of auxiliary raking variables should be highly correlated with influence functions.\(^30,31\) A generalized raking estimator will be asymptotically efficient among all design-based estimators if auxiliary raking variables are $E(h_i(\beta)|Y, Z, A)$.\(^11\)

Influence functions depend on unknown regression coefficients $\beta$, so optimal auxiliary raking variables are typically unavailable. Kulich and Lin\(^32\) proposed a “plug-in” method to approximate the conditional expectation $E(h_i(\beta)|Y, Z, A)$ where the missing values of $X$ are imputed using data from the phase-2 sample, and $h_i(\beta^*)$ are estimated using the imputed
X and phase-1 data. In this article, we follow the same procedures to conduct generalized raking estimations in the final analyses. The procedures are summarized as follows:

1. Fit a weighted regression model (imputation model) using phase-2 data, and impute X for all individuals.
2. Fit the outcome model with phase-1 data and imputed X, and then estimate the influence functions \( h_i(\beta^*) \). For linear regression, \( h_i(\beta^*) \) can be approximated using \texttt{dfbeta} in R. For other regression models, we can approximate \( h_i(\beta^*) \) from the score or jackknife estimators.
3. Estimate \( \beta \) using generalized raking with adjusted weights. We use the distance function \( d(a, b) = \log(a/b) - a + b \), which makes calibrated weights \( w_i \) non-negative.

The \texttt{survey} package in R is used to fit the weighted regression in step 1 and obtain generalized raking estimations in step 3. When fitting the outcome model in step 2, note that it is crucial to use imputed X for all observations.

4 | ON OPTIMAL DESIGNS OF DESIGN-BASED ESTIMATORS

4.1 | Neyman allocation

Neyman allocation minimizes the variance of an estimator of a population mean or total given a fixed sample size. Suppose we want to find the optimal allocation for the estimator of the population total of \( X \),

\[
T_X = \sum_{j=1}^{K} N_j \bar{X}_j,
\]

where \( \bar{X}_j \) is the mean of \( X \) for stratum \( j \). The objective can be written as

\[
\text{minimize } \text{var} (T_X) = \sum_{j=1}^{K} \frac{(N_j - n_j)N_j\sigma_j^2}{n_j} \quad \text{subject to } n_1 + n_2 + \cdots + n_K = n.
\]

Neyman showed the optimal allocation that minimizes \( \text{var} (T_X) \) subject to the constraint \( n_1 + n_2 + \cdots + n_K = n \) is

\[
n_j = n \frac{N_j\sigma_j}{\sum_{k=1}^{K} N_k\sigma_k}, \tag{3}
\]

where \( \sigma_j \) is the standard deviation of \( X \) for stratum \( j \). In Equation (3), both \( n \) and the denominator \( \sum_{k=1}^{K} N_k\sigma_k \) are constants, so Neyman allocation indicates that the optimal sample size for stratum \( j \) is proportional to \( N_j\sigma_j \). If \( N_j\sigma_j \) is fixed across strata, a balanced stratified sampling (\( n_1 = n_2 = \cdots = n_k \)) will be optimal. If \( \sigma_j \) is fixed across strata, a proportional stratified sampling (\( n_j \propto N_j \)) will be optimal.

The Equation (3) does not give integer solutions, and the usual practice is to round off to the nearest integer, but rounding does not necessarily end up with the optimal solution. Wright worked out an exact integer algorithm for Neyman allocation, which yields the minimum sampling variance.

4.2 | Optimal design for analysis via IPW estimator

We are interested in minimizing the variance of \( \hat{\beta}_1 \). According to Equation (2), the objective becomes

\[
\text{minimize } \text{var} \left( \sum_{i=1}^{N} h_i(\beta_1) \right) \quad \text{subject to } n_1 + n_2 + \cdots + n_K = n,
\]

and then the optimal design for analysis via the IPW estimator is to apply influence functions to Equation (3), which gives

\[
n_j \propto N_j \sqrt{\text{var} (h_i(\beta_1)|\text{stratum } j)}. \tag{4}
\]

McIsaac and Cook and Amorim et al derived similar results by minimizing the asymptotic variance of the mean-score and IPW estimator respectively. Chen and Lumley approximated the optimal design (Equation (4)) using a
multiwave sampling framework, and their final statistical analyses were conducted using generalized raking estimators. However, Equation (3) and (4) need not be the optimal allocation for analysis by generalized raking estimators. It is of interest to find the optimal design for analysis via generalized raking estimators as these are what will be used in analyses.

### 4.3 Optimal design for analysis via generalized raking estimators

Deville and Särndal\(^9\) showed all generalized raking estimators are asymptotically equivalent to the generalized regression estimator (GREG), so the optimal design for analysis via a particular generalized raking estimator is also optimal for others. We choose the GREG estimator because it makes arguments more straightforward. Suppose auxiliary raking variables are \(S\) and parameters in the regression estimator are \(\theta\). We want to estimate the population total of \(X\), and \(X\) can be decomposed as

\[
X = (X - S\theta) + S\theta.
\]

The GREG estimator can then be written as

\[
\hat{T} = \sum_{i=1}^{N} \frac{R_i}{\pi_i} (X_i - S_i\theta) + \sum_{i=1}^{N} S_i\theta,
\]

where the first term is the IPW estimator of the total of residuals from regressing \(X\) on \(S\). The second term involves the whole population. It has zero variance for any fixed \(\theta\) and has variance of smaller order than the first term for any estimator \(\theta\) that converges at \(\sqrt{n}\) rate. Therefore, to first order, we have

\[
\text{var} (\hat{T}) = \text{var} \left( \sum_{i=1}^{N} \frac{R_i}{\pi_i} (X_i - S_i\theta) \right).
\]

The variance of generalized raking estimators of the total is then the variance of the IPW estimator of the total of residuals \(X_i - S_i\theta\) from regressing \(X\) on \(S\). According to Section 4.1, the optimal design for analysis via generalized raking estimators is to apply Neyman allocation to residuals, which becomes

\[
n_j \propto N_j \sqrt{\text{var} ((X_i - S_i\theta) | \text{stratum } j)}.
\]

Expression (5) poses a problem for the influence-function approach discussed in Section 4.2. For the influence-function approach, \(X_i\) are the influence functions \(h_i(\beta_1)\) and \(S_i\) are the best estimates \(h_i(\beta^*_1)\) we have of them. Let \(\gamma\) be regression coefficients from regressing \(h_i(\beta_1)\) on \(h_i(\beta^*_1)\) and \(r_i\) be the residuals \(r_i = h_i(\beta_1) - h_i(\beta^*_1)\gamma\). The objective becomes

\[
\text{minimize } \text{var} \left( \sum_{i=1}^{N} r_i \right) \text{ subject to } n_1 + n_2 + \cdots + n_K = n.
\]

A plug-in estimator\(^32\) would estimate the residuals \(r_i\) as zero. Before sampling any phase-2 data, we cannot estimate the influence functions and residuals. In practice, it is still possible to estimate the residuals using a multiwave sampling framework. After wave 1, \(h_i(\beta_1)\) and \(h_i(\beta^*_1)\) can be estimated using data from the current and previous wave respectively. We can also simulate to see how optimal design varies, where \(h_i(\beta_1)\) and \(h_i(\beta^*_1)\) can be estimated from the full data and parametric models respectively.

### 5 Gain from optimizing the design

Our target is to improve the efficiency of design-based estimators by optimizing the design. However, for generalized raking estimators, if auxiliary raking variables are good, there is not much room for improvement (without model
assumption) because the variance of residuals is small. If auxiliary raking variables are bad, generalized raking will not improve on the IPW estimator. This is because the gain in generalized raking depends on the Pearson correlation between auxiliary raking variables and influence functions. If the correlation is low, generalized raking will not help. In this section, we examine the extent to which improvement is still possible.

### 5.1 Analytical results

In this subsection, we compare the optimal design for analysis by the IPW estimator with those for analysis by generalized raking estimators in the classical measurement-error setting where a classical measurement-error model is available and assumed to be correct. Let \( X = X + U \) be surrogate variables of \( X \), where \( U \) have the mean of zero and constant variance of \( \sigma_U^2 \). If \( X \) also have the mean of zero, at \( \beta = 0 \), the conditional variance of residuals \( r_i \) given stratum \( j \) can be written as

\[
\text{var}(r_i|\text{stratum } j) = \frac{\text{var}(U(Y - \mu)|\text{stratum } j)}{\text{var}(X(Y - \mu)|\text{stratum } j)\text{var}(h_i(0)|\text{stratum } j)}. \tag{6}
\]

The proof is provided in Appendix A. If we further stratify on \( Y \), Equation (6) can be simplified to

\[
\text{var}(r_i|\text{stratum } j) = \frac{\text{var}(U)}{\text{var}(X)}\text{var}(h_i(0)|\text{stratum } j). \tag{7}
\]

As \( \text{var}(U)/\text{var}(X) \) is a constant, the optimal design will be the same for the two classes of design-based estimators. An important example of outcome-dependent sampling is the case-control study. If we have a rare disease and small covariate effects, the optimal case-control design for analysis by the IPW estimator will sample the same number of cases and controls at \( \beta = 0 \), whose proof is provided in Appendix B. According to Equation (7), in the same setting, the optimal case-control design for analysis by generalized raking estimators will also take the same number of cases and controls.

We develop an interactive Shiny app to compare the optimal allocations for analyses by the two classes of design-based estimators, which is available from https://tchen.shinyapps.io/raking. If we stratify on \( Y \) and have small covariate effects, finite-sample numerical results obtained from the Shiny app are consistent with the analytical results obtained from Equation (7).

### 5.2 Simulation studies

In this subsection, we performed extensive simulation studies to compare the efficiencies of different two-phase sampling designs. We also examined the stratum-specific optimal sampling allocations for analyses by the two classes of design-based estimators.

In the first series of simulation studies, we assumed \( X \) followed a standard normal distribution, and \( Z \) was a binary variable generated from Bern(0.5). Let \( Y \) be the outcome which was generated from the linear model \( Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon \), where \( \epsilon \) followed a standard normal distribution. An auxiliary raking variable \( h_i(\beta^*_1) \), which correlated with the true influence functions \( h_i(\beta_1) \), was simulated with correlation \( \rho = \text{cor}(h_i(\beta^*_1), h_i(\beta_1)) \). We set phase-1 sample size \( N = 4000 \) and phase-2 sample size \( n = 600 \). The data were stratified into 2 strata based on \( h_i(\beta^*_1) \). Specifically, individual \( i \) was in stratum 1 if \( h_i(\beta^*_1) \) was in between its 35th and 65th percentile and was in stratum 2 otherwise. We implemented a simple random sampling (SRS), a balanced stratified sampling (BSS) (ie, \( n_1 = n_2 = \cdots = n_K \)), a proportional stratified sampling (PSS) (ie, \( n_j \propto N_j \)), an optimal design for analysis by the IPW estimator based on the true influence functions \( h_i(\beta_1) \) which were calculated using the (in practice, unavailable) full data (IF-IPW), and an optimal design for analysis by generalized raking estimators where \( h_i(\beta_1) \) and \( h_i(\beta^*_1) \) were used as influence functions and their best estimates respectively (IF-GR).

We also considered a simple random sampling with standard analysis of the phase-2 sample only (SRS-Ph2). We performed the IPW and generalized raking estimations. We used \( h_i(\beta^*_1) \) as auxiliary raking variables in generalized raking procedures. We set \( \beta_0 = \beta_1 = \beta_2 = 1 \) and performed 2000 Monte Carlo simulations.

For a linear model, influence functions can be written as

\[
h_i(\beta) = I^{-1}X^T(Y - \hat{Y}) \approx I^{-1}X^T\epsilon, \tag{8}
\]
where \( \hat{Y} = X \hat{\beta} \) and \( I \) was the per observation population information. According to Equation (8), for the same \( X \) and \( \epsilon \), varying \( \beta \) would not change the true influence functions \( h_i(\beta_1) \). If the stratification was independent of \( \beta \), then varying \( \beta \) would not change the IF-IPW design. In the first series of simulation studies, stratification and auxiliary raking variables were independent of regression coefficients \( \beta \), so varying \( \beta \) neither changed the IF-IPW design nor the \( \hat{\beta} \) estimated from generalized raking.

Results of the first series of simulation studies were given in Table 1 and Supplementary Figure S1 and Table S1. For analysis results of the IPW estimator, when \( h_i(\beta_1^*) \) were highly correlated with \( h_i(\beta_1) \), IF-IPW was more efficient than other sampling designs. The efficiency gain decreased as the correlation \( \rho \) decreased. The gain was not obvious when \( h_i(\beta_1^*) \) were not good approximations of \( h_i(\beta_1) \). BSS was the least efficient design as it sampled a lot more people from stratum 1 compared with the IF-IPW design. PPS was as efficient as SRS, and IF-GR was slightly more efficient than them.

Table 1 and Supplementary Table S1 showed that the median of estimated standard error was close to the simulation standard error so that estimators from all designs were valid. For all estimators we considered, simple random sampling with the IPW estimation was identical to simple random sampling with the standard analysis of the phase-2 sample. We also observed very similar results for all other simulation studies in this subsection, so we did not repeat these conclusions.

Table 1 showed IF-GR was also as efficient as SRS and PSS, which was partly due to the variance of residuals \( \text{var}(r_i) \) was roughly constant across strata, so these three designs were close. If \( \text{var}(r_i) \) was more different across strata, we expected

| \( \rho \) | SRS | BSS | PSS | IF-IPW | IF-GR | SRS-Ph2 |
|---|---|---|---|---|---|---|
| IPW | | | | | | |
| 0.99 | 1.65 | 0.041 | 2.31 | 0.048 | 1.66 | 0.041 | 1.36 | 0.037 | 1.63 | 0.040 |
| 0.90 | 1.67 | 0.041 | 2.27 | 0.048 | 1.64 | 0.041 | 1.43 | 0.038 | 1.54 | 0.039 |
| 0.80 | 1.67 | 0.041 | 2.37 | 0.049 | 1.67 | 0.041 | 1.52 | 0.039 | 1.63 | 0.040 |
| 0.70 | 1.77 | 0.042 | 2.25 | 0.047 | 1.64 | 0.041 | 1.58 | 0.040 | 1.62 | 0.040 |
| 0.60 | 1.60 | 0.040 | 2.29 | 0.048 | 1.72 | 0.042 | 1.65 | 0.041 | 1.55 | 0.039 |
| 0.50 | 1.66 | 0.041 | 2.04 | 0.045 | 1.66 | 0.041 | 1.71 | 0.041 | 1.67 | 0.041 |
| Generalized raking | | | | | | | Standard |
| 0.99 | 0.29 | 0.017 | 0.30 | 0.017 | 0.29 | 0.017 | 0.33 | 0.018 | 0.28 | 0.017 | 1.65 | 0.041 |
| 0.90 | 0.55 | 0.024 | 0.62 | 0.025 | 0.53 | 0.023 | 0.57 | 0.024 | 0.50 | 0.022 | 1.67 | 0.041 |
| 0.80 | 0.79 | 0.028 | 0.98 | 0.031 | 0.78 | 0.028 | 0.84 | 0.029 | 0.81 | 0.028 | 1.67 | 0.041 |
| 0.70 | 1.04 | 0.032 | 1.27 | 0.036 | 1.00 | 0.032 | 1.03 | 0.032 | 0.96 | 0.031 | 1.77 | 0.042 |
| 0.60 | 1.15 | 0.034 | 1.50 | 0.039 | 1.19 | 0.035 | 1.17 | 0.034 | 1.16 | 0.034 | 1.60 | 0.040 |
| 0.50 | 1.34 | 0.037 | 1.57 | 0.040 | 1.34 | 0.037 | 1.35 | 0.037 | 1.33 | 0.036 | 1.66 | 0.041 |

Note: We consider linear regression with continuous \( X \). MSE*: MSE \( \times 1000 \); SE: empirical standard error of \( \hat{\beta}_1 \); \( \rho \): \( \text{cov}(h_i(\beta_1^*), h_i(\beta_1)) \).

Abbreviations: BSS, balanced stratified sampling; IF-IPW, optimal design for analysis by the IPW estimator; IF-GR, optimal design for analysis by generalized raking estimators; PSS, proportional stratified sampling; SRS, simple random sampling; SRS-Ph2, SRS with standard analysis of the phase-2 sample.
stratum. Therefore, the optimal sample sizes were approximately the same for all strata. Based on the first series of

\[ Z \]

where

\[ U \]

followed a standard normal distribution. The error-prone variable we considered the situation that both the variable of interest

IF-GR design would first increase and then decrease. That, with the cut-off points moving toward extreme tails, the difference in sample size between the optimal IF-IPW and

Supplementary Figure S2. The cut-off points at the 20th and 80th percentiles were optimal for analysis via generalized raking and at the 14th and 86th percentiles were optimal for analysis via the IPW estimator. The efficiency gain was less

Supplementary Figure S2. The cut-off points at the 20th and 80th percentiles were optimal for analysis via generalized raking, and they were more efficient than the other three designs.

We then conducted simulation studies in measurement-error settings. In the second series of simulation studies, we chose 35th and 65th percentiles as cut-off points as we wanted to sample more people from stratum 2 (the tail stratum). However, these cut-off points were not optimal. For design-based estimators, given the number of strata, the optimal stratification would make Neyman allocation give the same

\[ N_j \sigma_j \]

in each stratum. Therefore, the optimal sample sizes were approximately the same for all strata. Based on the first series of simulation studies, we further considered different cut-off points and let \( \rho \) be 0.7. Results were presented in Table 3 and Supplementary Figure S2. The cut-off points at the 20th and 80th percentiles were optimal for analysis via generalized raking and at the 14th and 86th percentiles were optimal for analysis via the IPW estimator. The efficiency gain was less for generalized raking than for the IPW estimator by optimizing the strata boundary. Supplementary Figure S2 showed that, with the cut-off points moving toward extreme tails, the difference in sample size between the optimal IF-IPW and IF-GR design would first increase and then decrease.

We then conducted simulation studies in measurement-error settings. In the second series of simulation studies, we considered the situation that both the variable of interest \( X \) and outcome \( Y \) were continuous. We assumed \( X \) followed a standard normal distribution. The error-prone variable \( \tilde{X} \) was assumed to have additive error \( \tilde{X} = X + U \), where \( U \sim N(0, \sigma^2) \). The outcome of interest \( Y \) was generated from the linear model

\[ Y = \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + \epsilon \]

where \( Z_1 \sim Bern(0.5) \), \( Z_2 \) followed a standard uniform distribution, and \( \epsilon \) followed a standard normal distribution.

### TABLE 2  Mean squared error (MSE) and empirical standard error of \( \hat{\beta}_i \) estimated from generalized raking estimators for different designs with phase-1 sample size 4000, phase-2 sample size 600, and 2000 Monte Carlo simulations

| Cut-off points (percentile) | 35th & 65th | 20th & 80th | 14th & 86th | 5th & 95th |
|---------------------------|------------|------------|------------|-----------|
| IF-IPW MSE* SE            | IF-GR MSE* SE          | IF-IPW MSE* SE          | IF-GR MSE* SE          | IF-IPW MSE* SE          | IF-GR MSE* SE          |
| 1.54 0.039 1.63 0.040      | 1.41 0.038 1.47 0.038 | 1.36 0.037 1.43 0.038 | 1.36 0.037 1.43 0.038 |

Note: We consider linear regression with continuous \( X \) and let the correlation between influence functions and their best estimates be 0.7. The data were stratified into 2 strata based on \( h_2(\hat{\beta}) \), and we consider different cut-off points. MSE*: MSE \( \times 1000 \); SE: empirical standard error of \( \hat{\beta}_i \).

Abbreviations: BSS, balanced stratified sampling; IF-IPW, optimal design for analysis by the IPW estimator; IF-GR, optimal design for analysis by generalized raking estimators; PSS, proportional stratified sampling; SRS, simple random sampling; SRS-Ph2, SRS with standard analysis of the phase-2 sample.

### TABLE 3  Mean squared error (MSE) and empirical standard error of \( \hat{\beta}_i \) estimated from the IPW estimator and generalized raking estimators for IF-IPW and IF-GR with phase-1 sample size 4000, phase-2 sample size 600, and 2000 Monte Carlo simulations

| Cut-off points (percentile) | 35th & 65th | 20th & 80th | 14th & 86th | 5th & 95th |
|---------------------------|------------|------------|------------|-----------|
| IF-IPW MSE* SE            | IF-GR MSE* SE          | IF-IPW MSE* SE          | IF-GR MSE* SE          | IF-IPW MSE* SE          | IF-GR MSE* SE          |
| 0.95 0.031 0.98 0.031      | 0.97 0.031 0.92 0.030 | 0.98 0.031 0.97 0.031 | 1.01 0.032 1.00 0.032 |

Note: We consider linear regression with continuous \( X \) and control the correlation between influence functions and their best estimates around 0.7. MSE*: MSE \( \times 1000 \); SE: empirical standard error of \( \hat{\beta}_i \).

Abbreviations: BSS, balanced stratified sampling; IF-IPW, optimal design for analysis by the IPW estimator; IF-GR, optimal design for analysis by generalized raking estimators; PSS, proportional stratified sampling; SRS, simple random sampling; SRS-Ph2, SRS with standard analysis of the phase-2 sample.
Table 4 Mean squared error (MSE) and empirical standard error of $\hat{\beta}_1$ estimated from the IPW and generalized raking estimators for different designs with phase-1 sample size 4000, phase-2 sample size 600, and 2000 Monte Carlo simulations.

| $\tau$ | $\hat{\beta}_1$ | $\hat{\rho}$ | SRS | MSE* | SE | BSS | MSE* | SE | PSS | MSE* | SE | IF-IPW | MSE* | SE | IF-GR | MSE* | SE | SRS-Ph2 | MSE* | SE |
|--------|----------------|--------------|------|------|----|------|------|----|------|------|----|--------|------|----|--------|------|----|--------|------|----|
| 0.50   | 0.89           | 0.53         | 0.023| 0.68 | 0.026| 0.53 | 0.023| 0.58 | 0.024| 0.54 | 0.023| 1.57  | 0.040|    | 1.00   | 0.070| 1.68| 0.041  | 1.54 | 0.039|
| 1      | 0.85           | 0.75         | 0.027| 0.78 | 0.028| 0.71 | 0.027| 0.72 | 0.027| 0.74 | 0.027| 1.73  | 0.042|    | 0.75   | 0.78 | 1.76| 0.042  | 1.56 | 0.039|
| 2      | 0.73           | 1.03         | 0.032| 1.00 | 0.032| 0.98 | 0.031| 0.99 | 0.031| 0.98 | 0.031| 1.67  | 0.041|    |        | 0.65 | 1.26| 0.036  | 1.30 | 0.036|
| 1.00   | 0.70           | 0.96         | 0.031| 1.15 | 0.034| 0.95 | 0.031| 1.02 | 0.032| 0.97 | 0.031| 1.68  | 0.041|    |        | 0.74 | 1.27| 0.036  | 1.18 | 0.034|
|        | 0.61           | 1.46         | 0.038| 1.57 | 0.040| 1.52 | 0.039| 1.38 | 0.037| 1.44 | 0.038| 1.77  | 0.042|    |        | 0.61 | 1.46| 0.038  | 1.57 | 0.040|

Note: We consider linear regression with continuous $X$ and error-prone variable $\tilde{X} = X + U$, where $U \sim N(0, \tau^2)$. MSE*: MSE $\times 1000$; SE: empirical standard error of $\hat{\beta}_1$; $\hat{\rho}$: mean of the correlation between influence functions $h_i(\hat{\beta}_1)$ and their best estimates $h_i(\hat{\beta}_1^*)$ estimated from 2000 Monte Carlo simulations. Abbreviations: BSS, balanced stratified sampling; IF-IPW, optimal design for analysis by the IPW estimator; IF-GR, optimal design for analysis by generalized raking estimators; PSS, proportional stratified sampling; SRS, simple random sampling; SRS-Ph2, SRS with standard analysis of the phase-2 sample.

The data were divided into 3 strata based on $\tilde{X}$ ($\leq 20$th, $>20$th to $\leq 80$th, and $>80$th percentiles). We set $N = 4000$, $n = 600$, and $\beta_0 = \beta_2 = \beta_3 = \beta_4 = 1$. We considered the same sampling strategies as those implemented in the first series of simulation studies. The only difference was that, for the IF-GR design, the best estimates of influence functions were calculated using multiple imputation with 50 imputations. Specifically, imputations were calculated by Bayesian linear regression using the mice package in R. We performed the IPW and generalized raking estimations. The procedures described in Section 3.2 were adopted to calculate auxiliary raking variables in the final analysis.

Results of the second series of simulation studies were given in Table 4 and Supplementary Figure S3 to S5 and Table S3. The results were consistent with the previous simulation studies. For the analysis results of the IPW estimator, IF-IPW was more efficient than other designs. The efficiency gain decreased as $\tau$ increased. For the analysis results of generalized raking estimators, when $\tau = 0.5$, all five designs were equally efficient as $\tilde{X}$ was a good surrogate of $X$. When $\tau$ got larger, IF-GR was as efficient as IF-IPW, SRS, and PSS. BSS was the least efficient design compared with other designs. For all five designs, the efficiency decreased as $\tau$ and $\beta$ increased. Supplementary Figure S3 to S5 showed IF-GR tended to sample more observations from the middle stratum compared with IF-IPW, though it would be closer to IF-IPW as $\beta$ and $\tau$ increased.

In the third series of simulation studies, we let the outcome $Y$ be binary and the variable of interest $X$ be continuous. $X$, $Z_1$, and $\tilde{X}$ were generated exactly the same as the second series of simulation studies. The binary outcome was generated from $P(Y|X, Z_1) = \expit(\beta_0 + \beta_1X + \beta_2Z_1)$, where $\expit(x) = \exp(x)/(1 + \exp(x))$. The data were stratified into 6 strata base...
TABLE 5  Mean squared error (MSE) and empirical standard error of $\hat{\beta}_1$ estimated from the IPW and generalized raking estimators for different designs with phase-1 sample size 4000, phase-2 sample size 600, and 2000 Monte Carlo simulations

| $\tau$ | $\beta_1$ | $\rho$ | SRS | BSS | PSS | IF-IPW | IF-GR | SRS-Ph2 |
|-------|-----------|-------|-----|-----|-----|--------|-------|---------|
|       | MSE* | SE   | MSE* | SE   | MSE* | SE   | MSE* | SE   | MSE* | SE   | MSE* | SE   |
| 0.50  | 0.0  | 0.89 | 8.72 | 0.093 | 3.75 | 0.061 | 4.02 | 0.063 | 3.66 | 0.061 | 3.76 | 0.061 |
|       | 0.5  | 0.88 | 10.15 | 0.101 | 4.59 | 0.068 | 4.85 | 0.070 | 3.91 | 0.063 | 4.10 | 0.064 |
|       | 1.0  | 0.84 | 13.87 | 0.117 | 6.44 | 0.080 | 7.64 | 0.087 | 5.44 | 0.074 | 5.74 | 0.076 |
| 0.75  | 0.0  | 0.80 | 8.70 | 0.093 | 4.59 | 0.068 | 5.31 | 0.073 | 4.58 | 0.068 | 4.68 | 0.068 |
|       | 0.5  | 0.77 | 10.14 | 0.101 | 5.88 | 0.077 | 6.27 | 0.079 | 5.11 | 0.071 | 4.89 | 0.070 |
|       | 1.0  | 0.72 | 14.52 | 0.120 | 8.52 | 0.092 | 9.65 | 0.098 | 7.13 | 0.084 | 6.82 | 0.083 |
| 1.00  | 0.0  | 0.70 | 8.69 | 0.093 | 5.39 | 0.073 | 5.90 | 0.077 | 4.99 | 0.071 | 4.98 | 0.071 |
|       | 0.5  | 0.67 | 10.08 | 0.100 | 7.09 | 0.084 | 7.02 | 0.084 | 5.98 | 0.077 | 5.66 | 0.075 |
|       | 1.0  | 0.61 | 14.28 | 0.119 | 9.42 | 0.097 | 10.18 | 0.101 | 7.81 | 0.088 | 7.71 | 0.088 |

Generalized raking Standard

| $\tau$ | $\beta_1$ | $\rho$ | SRS | BSS | PSS | IF-IPW | IF-GR | SRS-Ph2 |
|-------|-----------|-------|-----|-----|-----|--------|-------|---------|
|       | MSE* | SE   | MSE* | SE   | MSE* | SE   | MSE* | SE   | MSE* | SE   | MSE* | SE   |
| 0.50  | 0.0  | 0.89 | 2.72 | 0.052 | 2.59 | 0.051 | 2.74 | 0.052 | 2.48 | 0.050 | 2.48 | 0.050 |
|       | 0.5  | 0.88 | 3.41 | 0.058 | 3.21 | 0.056 | 3.30 | 0.057 | 2.76 | 0.053 | 2.83 | 0.053 |
|       | 1.0  | 0.84 | 5.23 | 0.072 | 4.57 | 0.068 | 5.26 | 0.072 | 3.86 | 0.062 | 4.09 | 0.064 |
| 0.75  | 0.0  | 0.80 | 4.16 | 0.065 | 3.66 | 0.061 | 4.28 | 0.065 | 3.59 | 0.060 | 3.61 | 0.060 |
|       | 0.5  | 0.77 | 5.07 | 0.071 | 4.78 | 0.069 | 5.05 | 0.071 | 4.31 | 0.066 | 3.95 | 0.063 |
|       | 1.0  | 0.72 | 7.80 | 0.088 | 7.19 | 0.085 | 8.12 | 0.090 | 5.92 | 0.077 | 5.76 | 0.076 |
| 1.00  | 0.0  | 0.70 | 4.98 | 0.071 | 4.61 | 0.068 | 5.00 | 0.071 | 4.21 | 0.065 | 4.13 | 0.064 |
|       | 0.5  | 0.67 | 6.17 | 0.079 | 6.08 | 0.078 | 6.23 | 0.079 | 5.23 | 0.072 | 4.92 | 0.070 |
|       | 1.0  | 0.61 | 9.46 | 0.097 | 8.78 | 0.094 | 8.95 | 0.094 | 7.12 | 0.084 | 6.99 | 0.083 |

Note: We consider logistic regression with continuous $X$ and error-prone variable $\tilde{X} = X + U$, where $U \sim N(0, \tau^2)$. MSE*: MSE × 1000; SE: empirical standard error of $\hat{\beta}_1$; $\bar{\rho}$: mean of the correlation between influence functions $h_i(\hat{\beta}_1)$ and their best estimates $h_i(\hat{\beta}_1^*)$ estimated from 2000 Monte Carlo simulations.

Abbreviations: BSS, balanced stratified sampling; IF-IPW, optimal design for analysis by the IPW estimator; IF-GR, optimal design for analysis by generalized raking estimators; PSS, proportional stratified sampling; SRS, simple random sampling; SRS-Ph2, SRS with standard analysis of the phase-2 sample.

on $Y$ and the first and the third quartile of $\tilde{X}$. We set $\beta_0 = -1.5$, $\beta_2 = 1$, $N = 4000$, and $n = 600$. We considered the same sampling strategies as those implemented in the previous simulation studies.

Results of the third series of simulation studies were shown in Table 5 and Supplementary Figure S6 to S8 and Table S4. For the analysis results of the IPW estimator, IF-IPW was as efficient as IF-GR. When $\beta_1 = 0$, BSS was also as efficient as IF-IPW and IF-GR. The performance of BSS got worse as $\beta_1$ increased. Typically, PSS and SRS were less efficient compared with the other three designs. For analysis results of generalized raking estimators, we observed IF-IPW and IF-GR had similar performance and were more efficient than other designs. Supplementary Figure S6 to S8 showed that compared with a continuous outcome, IF-GR was closer to IF-IPW for a binary outcome, though IF-GR also sampled more people from the middle of the distribution of $\tilde{X}$ in each stratum defined by $Y$.

In the last series of simulation studies, we considered that the outcome $Y$, the variable of interest $X$, and the covariate $Z_1$ were all binary. Let $X \sim Bern(0.4)$ and $Z_1 \sim Bern(0.5)$. An error-prone variable $\tilde{X}$ was generated with prespecified sensitivity and specificity. The binary outcome was generated from $P(Y|X, Z_1) = \text{expit}(\beta_0 - \beta_1 X + \beta_2 Z_1)$. We considered a rare disease with $E(Y) = 0.05$ which was controlled by $\beta_0$. We set $\beta_2 = 1$ and $N = 10,000$. The data were divided into 4 strata based on $Y$ and $\tilde{X}$. Instead of balanced stratified sampling, we implemented a stratified case-control sampling (SCC) which sampled all cases and the same number of controls in each stratum defined by $\tilde{X}$. 
| Sen | Spe | \( \beta_1 \) | \( \hat{\rho} \) | SRS | SCC | PSS | IF-IPW | IF-GR | SRS-Ph2 |
|-----|-----|-----|-----|-----|-----|-----|------|------|-------|
|     |     |     |     | MSE* | SE  | MSE* | SE  | MSE* | SE  | MSE* | SE  |
| IPW |     |     |     | ----- |     | ----- |     | ----- |     | ----- |     |
| 0.95| 0.95| 0.0 | 0.89| 10.03| 0.317| 1.21  | 0.110| 2.52  | 0.159| 1.18  | 1.09 | 1.18 | 0.109|
|     | 0.5 | 0.88| 11.61| 0.340| 1.32  | 0.115| 3.09  | 0.176| 1.21  | 0.110| 1.24 | 0.111|
|     | 1.0 | 0.86| 14.60| 0.381| 1.67  | 0.129| 4.80  | 0.219| 1.53  | 0.124| 1.54 | 0.124|
| 0.90| 0.90| 0.0 | 0.79| 9.40  | 0.307| 1.35  | 0.116| 4.06  | 0.202| 1.37  | 0.117| 1.35 | 0.116|
|     | 0.5 | 0.77| 11.96| 0.345| 1.49  | 0.122| 5.02  | 0.224| 1.44  | 0.120| 1.47 | 0.121|
|     | 1.0 | 0.73| 15.44| 0.392| 1.95  | 0.139| 7.31  | 0.270| 1.67  | 0.129| 1.76 | 0.133|
| 0.85| 0.85| 0.0 | 0.69| 9.30  | 0.305| 1.49  | 0.122| 5.07  | 0.225| 1.44  | 0.120| 1.42 | 0.119|
|     | 0.5 | 0.66| 10.98| 0.331| 1.66  | 0.129| 6.57  | 0.256| 1.55  | 0.124| 1.53 | 0.124|
|     | 1.0 | 0.62| 14.60| 0.380| 2.07  | 0.144| 9.70  | 0.311| 1.83  | 0.135| 1.79 | 0.134|
| Generalized raking | Standard | MSE* | SE  | MSE* | SE  | MSE* | SE  | MSE* | SE  |
| 0.95| 0.95| 0.0 | 0.89| 2.65  | 0.163| 1.06  | 0.103| 2.51  | 0.158| 1.07  | 1.04 | 1.05 | 1.03 | 10.03 | 0.317|
|     | 0.5 | 0.88| 3.28  | 0.181| 1.16  | 0.108| 3.03  | 0.174| 1.14  | 0.107| 1.14 | 0.107 | 11.61 | 0.340|
|     | 1.0 | 0.86| 4.97  | 0.222| 1.51  | 0.123| 4.83  | 0.219| 1.43  | 0.119| 1.42 | 0.119 | 14.60 | 0.381|
| 0.90| 0.90| 0.0 | 0.79| 4.14  | 0.204| 1.24  | 0.111| 4.04  | 0.201| 1.29  | 0.114| 1.24 | 0.112 | 9.40  | 0.307|
|     | 0.5 | 0.77| 5.58  | 0.236| 1.39  | 0.118| 4.97  | 0.223| 1.33  | 0.115| 1.35 | 0.116 | 11.96 | 0.345|
|     | 1.0 | 0.73| 7.95  | 0.281| 1.81  | 0.135| 7.30  | 0.270| 1.60  | 0.126| 1.66 | 0.129 | 15.44 | 0.392|
| 0.85| 0.85| 0.0 | 0.69| 5.36  | 0.232| 1.41  | 0.119| 5.17  | 0.227| 1.36  | 0.117| 1.35 | 0.116 | 9.30  | 0.305|
|     | 0.5 | 0.66| 7.01  | 0.265| 1.58  | 0.126| 6.68  | 0.258| 1.46  | 0.121| 1.48 | 0.122 | 10.98 | 0.331|
|     | 1.0 | 0.62| 9.36  | 0.305| 1.96  | 0.140| 9.94  | 0.314| 1.77  | 0.133| 1.75 | 0.132 | 14.60 | 0.380|

Note: We consider a rare disease \( E(Y) = 0.05 \) and logistic regression with binary \( X \). The error-prone variable is generated with prespecified sensitivity and specificity. MSE*: MSE \times 100; SE: empirical standard error of \( \hat{\beta}_1 \); Sen: sensitivity; Spe: specificity; \( \bar{\rho} \): mean of the correlation between influence functions \( h_i(\hat{\beta}_1) \) and their best estimates \( h_i(\hat{\beta}_*1) \) estimated from 2000 Monte Carlo simulations.

Abbreviations: IF-IPW, optimal design for analysis by the IPW estimator; IF-GR, optimal design for analysis by generalized raking estimators; PSS, proportional stratified sampling; SRS, simple random sampling; SCC, stratified case-control sampling; SRS-Ph2, SRS with standard analysis of the phase-2 sample.

Results of the last series of simulation studies were shown in Table 6 and Supplementary Figure S9 to S11 and Table S5. For the analysis results of the IPW estimator, IF-IPW was as efficient as IF-GR. They were slightly more efficient than SCC and much more efficient than PSS and SRS. For the analysis results of generalized raking estimators, IF-IPW and IF-GR were also equally efficient. SCC was also close to them. These three designs were substantially more efficient than PSS and SRS. Supplementary Figure S9 to S11 showed IF-GR was quite close to IF-IPW when data were all discrete. This was consistent with the findings of McIsaac and Cook.21

### 6 DATA EXAMPLE: NATIONAL WILMS’ TUMOR STUDY

In this section, we compared different two-phase sampling strategies using the data example from the National Wilms’ Tumor Study.24,25 The cohort consisted of 3915 patients. Variables available for the whole cohort included histology evaluated by central lab (favorable vs unfavorable (histol)), histology evaluated by institution (favorable vs unfavorable (instit)), age at diagnosis (age), stage of disease (stage), study, diameter of tumor (tumdiam), and an indicator of relapse (relapse). We assumed histology evaluated by central lab (histol) was the phase-2 variable and histology evaluated by institution was the surrogate variable.
TABLE 7 A comparison of different sampling strategies where the strata were defined based on institutional histology and outcome for Wilm’s Tumor

| Stratum size       | SRS | SCC | PSS | IF-IPW | IF-GR |
|--------------------|-----|-----|-----|--------|-------|
| Controls & favourable | 3026 | 1034 | 507 | 1034   | 736   |
| Controls & unfavorable | 220  | 75  | 162 | 75     | 144   |
| Cases & favourable   | 507  | 173 | 507 | 173    | 345   |
| Cases & unfavorable  | 162  | 55  | 162 | 55     | 113   |

Abbreviations: IF-IPW, stratum-specific sample size of the optimal design for analysis by the IPW estimator; IF-GR, the mean stratum-specific sample size of the optimal sampling for analysis by generalized raking estimators based on 2000 Monte Carlo simulations; PSS, stratum-specific sample size of proportional stratified sampling; SRS, the mean stratum-specific sample size of simple random sampling based on 2000 Monte Carlo simulations; SCC, stratum-specific sample size of stratified case-control sampling.

TABLE 8 The National Wilms’ Tumor Study data example. Mean squared error (MSE) and empirical phase-2 standard error of $\hat{\beta}_{\text{histol}}$ estimated from the IPW and generalized raking estimators for different designs with phase-1 sample size 3915, phase-2 sample size 1338, and 2000 Monte Carlo simulations

| $\rho$ | SRS | SCC | PSS | IF-IPW | IF-GR | SRS-Ph2 |
|-------|-----|-----|-----|--------|-------|---------|
|       | MSE* | SE  | MSE* | SE     | MSE* | SE     | MSE    | SE     |
| IPW   | 0.76 | 2.49 | 0.158 | 0.79 | 0.089 | 1.04 | 0.102 | 0.69 | 0.083 | 0.66 | 0.081 |
|       | 0.76 | 0.97 | 0.099 | 0.77 | 0.088 | 0.93 | 0.097 | 0.65 | 0.081 | 0.63 | 0.079 | 2.49 | 0.158 |

Note: MSE*: MSE × 100; SE: empirical phase-2 standard error of $\hat{\beta}_{\text{histol}}$; $\rho$: mean of the correlation between influence functions $h_i(\hat{\beta}_{\text{histol}})$ and their best estimates $\hat{h}_i(\hat{\beta}_{\text{histol}}^\ast)$ estimated from 2000 Monte Carlo simulations.

Abbreviations: IF-IPW, optimal design for analysis by the IPW estimator; IF-GR, optimal design for analysis by generalized raking estimators; PSS, proportional stratified sampling; SCC, stratified case-control sampling; SRS, simple random sampling; SRS-Ph2, SRS with standard analysis of the phase-2 sample.

The data were stratified into 4 strata based on the indicator of relapse and histology evaluated by institution. The phase-1 sample sizes were (3026, 220, 507, 162). Following Kulich and Lin,32 Breslow et al,30 and Chen and Lumley,22 we fitted a logistic model with model terms histology (histol), a linear spline with the separate slope for greater or less than 1 year old, a binary indicator of stage (I-II vs III-IV), and interactions between stage and diameter. We performed the IPW and generalized raking estimations. We followed the procedures described in Section 3.2 and imputed central lab histology using a logistic model with model terms histology evaluated by institution, a binary indicator of age (>10 vs ≤10) and interaction between study and a binary indicator of stage (I-III vs IV).

We considered a stratified case-control sampling (SCC) which sampled all cases and the same number of controls in each histology stratum, a proportional stratified sampling (PSS), an optimal design for analysis by the IPW estimator (IF-IPW) based on the true influence functions which were calculated using the full data, and an optimal design for analysis by generalized raking estimators (IF-GR) where the best estimates of the influence functions were calculated by multiple imputation with 50 imputations. We also considered a simple random sampling with standard analysis of the phase-2 sample only (SRS-Ph2).

Detailed sampling results were shown in Table 7. SRS and PSS sampled too few relapsed cases. Compared with SRS and PSS, SCC sampled more relapsed cases and rare unfavorable histology controls. IF-IPW was close to IF-GR, and they sampled a few more favourable histology controls compared with SCC.

Results were shown in Table 8 and Supplementary Table S6. Generalized raking estimators were more efficient than the IPW estimator. As expected, IF-GR was as efficient as IF-IPW, and they were much more efficient than the other three designs. SCC was also more efficient than SRS and PSS.

In Supplementary Table S6, simulation standard errors were smaller than the median of estimated standard errors for estimators from all designs. We had fixed phase-1 data in this real data example so that simulation standard error did not incorporate the phase-1 uncertainty. We then used the Fisher information to estimate the phase-1 variance using the
whole data. In Supplementary Table S6, the median of estimated standard error was close to the square root of the phase-1 variance plus the simulation variance. Therefore, estimators from all designs were still valid.

7 | DISCUSSION

In this article, we explore the optimal sampling for design-based estimators of regression models in two-phase designs. The optimal design for analysis via the IPW estimator is to apply Neyman allocation to influence functions. We derive a closed-form solution of the optimal design for analysis by generalized raking estimators. We show that, the optimal design for analysis by generalized raking estimators is to apply Neyman allocation to residuals from regressing the influence functions $h_i(\beta_1)$ on their best estimates $h_i(\beta_1^*)$.

In practice, it is hard to approximate the optimal design for analysis by generalized raking estimators. In order to approximate the design, we need to estimate the influence functions and their best estimates, which cannot be done without any phase-2 data or prior information. However, it is still possible to approximate the design using a multiwave adaptive estimator. After wave 1, influence functions estimated from the current and the previous wave can be used as $h_i(\beta_1)$ and their best estimates $h_i(\beta_1^*)$ respectively.

In our simulation studies, proportional stratified sampling and simple random sampling are highly efficient for generalized raking estimation with a continuous outcome. The results are consistent with simulation studies of Amorim et al. If data are all discrete, the optimal design for analysis by the IPW estimator is close to those for analysis by generalized raking estimators, and they are more efficient than other designs. These findings are also consistent with the results of McIsaac and Cook.

For linear regression, our simulation results show that if using generalized raking for estimation, optimizing the design only gains very little even compared with simple random sampling. However, this is not true for logistic regression, where case-control sampling is known to give important efficiency gains, or for survival analysis. If influence functions $h_i(\beta_1)$ are highly correlated with their best estimates $h_i(\beta_1^*)$, residuals from regressing $h_i(\beta_1)$ on $h_i(\beta_1^*)$ will be roughly constant, so the optimal design for analysis via generalized raking will be close to proportional stratified sampling and simple random sampling. Future theoretical work is merited to compare the optimal design for analysis via generalized raking with simple random sampling for linear regression.

For all estimators that we considered in this article, simple random sampling with IPW estimation is identical to simple random sampling with the standard analysis of the phase-2 sample only. Their point estimates are the same, but estimated standard errors are not identical by construction. The IPW estimator estimates standard errors by the sandwich estimator, whereas the standard analysis is based on the Fisher information. However, in the Supplementary tables, we show that they are equal to at least three decimal places.

If auxiliary raking variables are good, $h_i(\beta_1^*)$ will be highly correlated with $h_i(\beta_1)$, so there is not much room for gains from optimizing the design. If auxiliary raking variables are bad, on the one hand, generalized raking estimators will not improve on the IPW estimator because the correlation between auxiliary raking variables and influence functions is low; on the other hand, the variance of residuals $\text{var}(r_i)$ will be close to the variance of influence functions $\text{var}(h_i(\beta_1))$, so that the optimal design for analysis by the IPW estimator is close to those for analysis by generalized raking estimators. If auxiliary raking variables are neither too good nor too bad, it may still be possible to gain from optimizing the design.

It may also be possible to gain from optimizing the stratum boundary. For design-based estimators, given the number of strata, the optimal stratification makes Neyman allocation give the same $N_j\sigma_j$ in each stratum. Our simulation studies indicate that the efficiency gain is less for generalized raking than for the IPW estimator by optimizing the stratum boundary.

We show that optimal designs for analyses by the IPW and generalized raking estimators are different but often have similar efficiency. Lack of improvement is desired in practice because it is hard to approximate the optimal design for generalized raking estimations.

ACKNOWLEDGEMENTS

We thank the two reviewers for their useful comments. This work was supported in part by the University of Auckland doctoral scholarship (to the first author), Patient Centered Outcomes Research Institute (PCORI) Award R-1609-36207 and U.S. National Institutes of Health (NIH) grant R01- AI131771. The statements in this manuscript are solely the responsibility of the authors and do not necessarily represent the views of PCORI or NIH.
DATA AVAILABILITY STATEMENT
Code of all simulation studies is available from https://github.com/T0ngChen/Opt_sampling_design_based.

ORCID
Tong Chen https://orcid.org/0000-0003-2751-2412

REFERENCES
1. Huang J, Duan R, Hubbard RA, et al. PIE: a prior knowledge guided integrated likelihood estimation method for bias reduction in association studies using electronic health records data. J Am Med Inform Assoc. 2017;25(3):345-352.
2. Shepherd BE, Shaw PA. Errors in multiple variables in human immunodeficiency virus (HIV) cohort and electronic health record data: statistical challenges and opportunities. Stat Commun Infect Dis. 2020;12(1):2019015.
3. Chatterjee N, Chen YH, Maas P, Carroll RJ. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. J Am Stat Assoc. 2016;111(513):107-117.
4. Yadav P, Steinbach M, Kumar V, Simon G. Mining electronic health records (EHRs): a survey. ACM Comput Surv. 2018;50(6):1-40.
5. Shepherd BE, Han K, Chen T, et al. Analysis of error-prone electronic health records with multi-wave validation sampling: association of maternal weight gain during pregnancy with childhood outcomes. ArXiv 2021. arXiv:2109.14001.
6. Giganti MJ, Shaw PA, Chen G, et al. Accounting for dependent errors in predictors and time-to-event outcomes using electronic health records, validation samples and multiple imputation. Ann Appl Stat. 2020;14(2):1045-1061.
7. Neyman J. Contribution to the theory of sampling human populations. J Am Stat Assoc. 1938;33(201):101-116.
8. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc. 1952;47(260):663-685.
9. Deville JC, Särndal CE. Calibration estimators in survey sampling. J Am Stat Assoc. 1992;87(418):376-382.
10. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc. 1994;89(427):846-866.
11. Lumley T, Shaw PA, Dai JY. Connections between survey calibration estimators and semiparametric models for incomplete data. Int Stat Rev. 2011;79(2):200-220.
12. Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Using the whole cohort in the analysis of case-cohort data. Am J Epidemiol. 2009;169(11):1398-1405.
13. Oh EJ, Shepherd BE, Lumley T, Shaw PA. Improved generalized raking estimators to address dependent covariate and failure-time outcome error. Biom J. 2021;63(5):1006-1027.
14. Han K, Shaw PA, Lumley T. Combining multiple imputation with raking of weights: an efficient and robust approach in the setting of nearly true models. Stat Med. 2021;1-15.
15. Scott AJ, Wild CI. Fitting regression models to case-control data by maximum likelihood. Biometrika. 1997;84(1):57-71.
16. Tao R, Zeng D, Lin DY. Efficient semiparametric inference under two-phase sampling, with applications to genetic association studies. J Am Stat Assoc. 2017;112(520):1468-1476.
17. Reilly M, Pepe MS. A mean score method for missing and auxiliary covariate data in regression models. Biometrika. 1995;82(2):299-314.
18. McIsaac MA, Cook RJ. Adaptive sampling in two-phase designs: a biomarker study for progression in arthritis. Stat Med. 2015;34(21):2899-2912.
19. Han K, Lumley T, Shepherd BE, Shaw PA. Two-phase analysis and study design for survival models with error-prone exposures. Stat Methods Med Res. 2021;30(3):857-874.
20. Neyman J. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. J R Stat Soc. 1934;97(4):558-625.
21. McIsaac MA, Cook RJ. Response-dependent two-phase sampling designs for biomarker studies. Can J Stat. 2014;42(2):268-284.
22. Chen T, Lumley T. Optimal multi-wave sampling for regression modeling in two-phase designs. Stat Med. 2020;39(30):4912-4921.
23. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005;61(4):962-973.
24. D’Angio GJ, Breslow N, Beckwith JB, et al. Treatment of Wilms’ tumor: results of the third national Wilms’ tumor study. Cancer. 1989;64(2):349-360.
25. Green DM, Breslow NE, Beckwith JB, et al. Comparison between single-dose and divided-dose administration of dactinomycin and doxorubicin for patients with Wilms’ tumor: a report from the national Wilms’ tumor study group. J Clin Oncol. 1998;16(1):237-245.
26. Chang W, Cheng J, Allaite J, et al. shiny: web application framework for R; 2021. R package version 1.6.0.
27. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581-592.
28. Prentice RL. Covariate measurement errors and parameter estimation in a failure time regression model. Biometrika. 1982;69(2):331-342.
29. Breslow NE, Wellner JA. A Z-theorem with estimated nuisance parameters and correction note for ‘weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression’. Scand J Stat. 2008;35(1):186-192.
30. Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Improved Horvitz–Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. Stat Biosci. 2009;1(1):32-49.
31. Rivera C, Lumley T. Using the whole cohort in the analysis of countermatched samples. Biometrics. 2016;72(2):382-391.
32. Kulich M, Lin DY. Improving the efficiency of relative-risk estimation in case-cohort studies. J Am Stat Assoc. 2004;99(467):832-844.
33. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2020.
APPENDIX A. TECHNICAL DETAIL

The goal is to minimize

\[
\text{var} \left( \sum_{i=1}^{N} (h_i(\beta) - h_i(\beta^*_i))y_i \right)
\]

subject to \( n_1 + n_2 + \cdots + n_K = n \).

Suppose \( \tilde{X} = X + U \) are surrogate variables of \( X \), where \( U \) have mean zero and constant variance \( \sigma_U^2 \). Regression calibration\(^{28} \) suggested we can regress \( Y \) on \( E(X|\tilde{X}) \) to get unbiased estimates of \( \beta \). If \( X \) have mean 0, we have

\[ E(X|\tilde{X}) = \lambda \tilde{X}, \]

where \( \lambda = \text{var}(X)/\text{var}(\tilde{X}) \). At \( \beta = 0 \), \( h_i(\beta) \) and \( h_i(\beta^*_i) \) can be estimated by \( X(Y - \mu) \) and \( \lambda \tilde{X}(Y - \mu) \) respectively, where \( \mu \) is a constant. Since \( X \) are independent of \( Y \) at \( \beta = 0 \), \( \hat{\gamma} \) can then be estimated as

\[ \hat{\gamma} = \frac{\text{cov}(\lambda \tilde{X}(Y - \mu), X(Y - \mu))}{\text{var}(\lambda \tilde{X}(Y - \mu))} = \frac{\text{var}(X(Y - \mu))}{\lambda \text{var}(\tilde{X}(Y - \mu))} = 1. \]

The residual becomes

\[ r_i = X(Y - \mu) - \tilde{X}(Y - \mu) = -U(Y - \mu), \]

Comparing \( \text{var}(r_i|\text{stratum \ j}) \) with \( \text{var}(h_i(0)|\text{stratum \ j}) \), we end up with

\[ \text{var}(r_i|\text{stratum \ j}) = \frac{\text{var}(U(Y - \mu)|\text{stratum \ j})}{\text{var}(\tilde{X}(Y - \mu)|\text{stratum \ j})} \times \frac{\text{var}(h_i(0)|\text{stratum \ j})}{\text{var}(X(Y - \mu)|\text{stratum \ j})} \]

APPENDIX B. CASE-CONTROL SAMPLING FOR IPW ESTIMATOR

Let us assume we have a rare disease \((E[Y] = p_0)\) and modest covariate effects, so \( p_i \ll 1 \) for all \( i \). In the case stratum, the influence functions \( h_i(\beta) = X_i(1 - p_i) \), so

\[ \text{var}(h_i(\beta)|Y = 1) \approx \text{var}(X_i|Y_i = 1) \approx \text{var}(X) \]

where the last approximate equality is exact if \( \beta = 0 \). In the control stratum, the influence functions \( h_i(\beta) = -X_i p_i \), so

\[ \text{var}(h_i(\beta)|Y = 0) \approx p_i^2 \text{var}(X). \]
This approximation is not as good as the case one, since the relative variation in $p_i$ is greater than that in $1 - p_i$ for a rare disease: typically the control variance will be larger than $p_0^2 \text{var}(X)$. Neyman allocation says we need to take the population stratum sizes $N_j$ and the population stratum standard deviations $\sigma_j$ and compute $N_j \sigma_j$ for each stratum $j$. Under our approximations, these come to $N p_0 \sqrt{\text{var}(X)}$ for cases and $N (1 - p_0) \sqrt{p_0^2 \text{var}(X)}$ for controls, which are approximately equal. We should take the same number of cases and controls when covariate effects are small.