Published in final edited form as:

Stat Med. 2020 December 30; 39(30): 4912–4921. doi:10.1002/sim.8760.

Optimal multiwave sampling for regression modelling in two-phase designs

Tong Chen, **Thomas Lumley**

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

Abstract

Two-phase designs involve measuring extra variables on a subset of the cohort where some variables are already measured. The goal of two-phase designs is to choose a subsample of individuals from the cohort and analyse that subsample efficiently. It is of interest to obtain an optimal design that gives the most efficient estimates of regression parameters. In this paper, we propose a multi-wave sampling design to approximate the optimal design for design-based estimators. Influence functions are used to compute the optimal sampling allocations. We propose to use informative priors on regression parameters to derive the wave-1 sampling probabilities because any pre-specified sampling probabilities may be far from optimal and decrease the design efficiency. The posterior distributions of the regression parameters derived from the current wave will then be used as priors for the next wave. Generalised raking is used in the final statistical analysis. We show that a two-wave sampling with reasonable informative priors will end up with a highly efficient estimation for the parameter of interest and be close to the underlying optimal design.

Keywords

influence function; Neyman allocation; prior; design-based estimators; optimal design

1 | INTRODUCTION

Large epidemiological studies often collect information on disease status and a large number of covariates for the entire cohort. However, variables of interest, such as risk factors or some expensive exposures, are cost-prohibitive to collect. It is only possible to measure these variables on a subsample of individuals under a fixed budget. Two-phase stratified sampling[1] can be useful in this situation. At phase 1, we collect relatively cheap information for the entire cohort, and at phase 2, we sample a small number of individuals from the strata defined by phase-1 data and measure the variables of interest. With considerate choices of stratification and phase-2 sampling probabilities, a two-phase design will result in efficient parameter estimations under a fixed budget constraint.[2]

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

The estimation methods of two-phase designs have been extensively studied, which can be classified into design-based estimation methods and model-based estimation methods. For design-based estimators, weighted likelihood is the most widely used method which weights each observation by the inverse of its sampling probability. Generalised raking[3,4] is a more efficient class of design-based estimators. It improves the efficiency by adjusting the sampling probability based on auxiliary variables. For model-based methods, efficiency gains can be achieved by making assumptions on the outcome model. The maximum likelihood estimators assume the outcome model is correctly specified, see Scott and Wild[5]; Breslow and Holubkov[6] for discrete phase-1 data and Tao et al.[7] for continuous or discrete phase-1 data. Typically, the maximum likelihood methods are the most efficient estimation methods but not robust to model misspecification.[8] We focus on the design-based estimation methods in this work.

Compared with estimation methods, the sampling design has not been widely studied. It is of interest to obtain the optimal design, which will include more informative individuals in the phase-2 sample. However, the optimal design will be different for different estimation methods. For the maximum likelihood estimators, as the outcome model is assumed to be correctly specified, sampling one individual can allow us to extrapolate information about other individuals in the population. For the design-based estimators, it cannot because we do not make any assumptions on the outcome model. Recently, Tao et al.[9] showed that the optimal design for the maximum likelihood estimators would sample from two extreme tails of the derivative of loglikelihood in each stratum when β_X is not a strong predictor. This design does not even allow consistent estimations with design-based methods. The optimal design of design-based estimators is Neyman allocation[10] applied to the influence functions, which samples relatively evenly across strata.

Optimal designs have been considered in some previous works. Reilly and Pepe[11] derived a closed-form expression of the optimal design for their mean-score estimator. Since the expression depends on phase-2 data which are not available at the design stage, Reilly[12] suggested to estimate the expression using data from a further pilot study. McIsaac and Cook[13] proposed to save this extra cost by using a multi-wave sampling. The idea is to sample wave 1 with some pre-specified sampling probabilities and then combine phase-1 and wave-1 data to estimate design components. The later waves can then be sampled adaptively.

In this work, we exploit an optimal multi-wave sampling approach for design-based estimators. In survey literature, the well-known Neyman allocation[10] is the optimal sampling strategy; it minimises the variance of population total for the variable of interest. The regression parameter can be written as the total of its influence functions,[14] so Neyman allocation can then be adopted for minimising the variance of the regression parameter. The influence functions also depend on phase-2 data so that a multi-wave sampling can be useful.

However, the wave-1 sampling probability and sampling size of each stratum turn out to be important. If the wave-1 sampling probabilities are far from optimal, we may oversample individuals from some less interesting strata. Moreover, a small wave-1 sample size may lead to the influence functions to be poorly estimated. In this paper, we show that informative priors on model parameters can improve both two-phase designs and estimations, even for a non-Bayesian final analysis. An efficient wave-1 sampling can be derived with the help of reasonable informative priors. For the following waves, the posterior distributions obtained from the previous wave will be used as priors for the current wave. The priors both improve the design efficiency and regularise the analysis for each wave.

The rest of this paper is organised as follows. In Section 2, we define notations and introduce Neyman allocation. In Section 3, the proposed multi-wave sampling and generalised raking are discussed in details. We report the results of simulation studies in Section 4. The performance of the proposed sampling method is illustrated using the National Wilms' Tumor Study (NWTS)[15,16] dataset example in Section 5. Code for all the simulation studies is available from https://github.com/T0ngChen/multiwave. Remarks are made in Section 6.

2 | TWO-PHASE DESIGNS AND NEYMAN ALLOCATION

2.1 | Notation

Consider a two-phase sampling design, let N and n be the phase-1 and phase-2 sample size respectively. Let Y denote an outcome variable, Z denote inexpensive covariates and A denote auxiliary variables. We have variables Z, A and Y measured for everyone in the cohort at phase 1. Let X be a variable of primary interest and X is only measured on the phase-2 subsample. Let R_i be an indicator variable, if $R_i = 1$, individual i is in the phase-2 sample, otherwise $R_i = 0$. The probability for individual i selected in the phase-2 sample is $(R_i|Z,A,Y) = \pi_i$. The sampling weight for ith observation can then be defined as $w_i = 1/\pi_i$.

We refer to $P(Y|X,Z;\beta)$ as the outcome model and $P(X|Z,A;\alpha)$ as the imputation model, so that Z are the components of phase-1 information that we want to put in the outcome model and A are auxiliary variables that are not in the outcome model, but can be used for stratification and imputation.

In two-phase designs, we assume that the missingness on X only depends on phase-1 data ($P(R|X,Y,Z,A) = P(R|Y,Z,A)$), so that the phase-2 data are missing at random.[17] We use the generalised raking estimator as described in Section 3.3 in the statistical analysis, and our goal is to minimise the variance of $\hat{\beta}_x$ by utilising the optimal multi-wave sampling design.

2.2 | Neyman allocation

Suppose a cohort is divided into H strata, and the unbiased estimator of the population total for the outcome variable Y can be written as

$$T_Y = \sum_{h=1}^{H} N_h \bar{y}_h, \qquad (1)$$

where $\bar{y}_h$ is the sample mean for stratum h. Neyman[10] derived the optimal sampling allocation to minimise the sampling variance of an estimator of a total with respect to the constraint $n_1 + n_2 + \cdots + n_H = n$. It can be expressed as

$$n_i = \frac{nN_i\sigma_i}{\sum_{h=1}^{H} N_h\sigma_h}, \tag{2}$$

where σ_i is the population standard deviation for stratum i, n_i and N_i are the phase-2 and phase-1 sample size for stratum i respectively.

However, Neyman allocation treats n as a continuous variable, so the value of n_i calculated from Equation (2) is not an integer in general. The usual practice is to round to the nearest integer and adjust iteratively if needed, but this does not always lead to the optimal allocation. Recently, Wright[18] developed an algorithm to find an exact optimal allocation; this algorithm is related to the Huntington-Hill method used to assign US Congress seats to states.

3 | MULTI-WAVE SAMPLING FOR DESIGN-BASED ESTIMATORS

3.1 | Neyman allocation for β_X

We are interested in improving the efficiency for the regression parameter β_X, so we need to write β_X as a total. Breslow et al.[14] noted that an estimator of the regression parameter can be written as a total of its influence functions, so we have

$$\sqrt{N}(\hat{\beta} - \beta_0) = \sum_{i=1}^{N} \mathbf{h}_i(\beta) + o_p\left(N^{-1/2}\right), \tag{3}$$

where $\mathbf{h}_i(\beta)$ is the influence function for observation i in the cohort. It can be approximated by the delta-beta which is the change in $\hat{\beta}$ when observation i is deleted. According to Equation (3), a weighted estimator β_w can be written as

$$\sqrt{n}(\hat{\beta}_w - \beta_0) = \sum_{i=1}^{n} w_i\mathbf{h}_i(\beta) + o_p\left(n^{-1/2}\right). \tag{4}$$

According to the proposed sampling design in Section 3.2, within each wave, the missing values of X will be imputed, and the influence functions for all individuals can then be estimated. Replacing σ_i with the standard deviation of influence functions $\mathrm{Var}(\mathbf{h}_i(\beta))^{1/2}$ in Equation (2), we have the optimal continuous allocation

$$n_i = \frac{nN_i\mathrm{Var}(\mathbf{h}_i(\beta))^{1/2}}{\sum_{h=1}^{H} N_h\mathrm{Var}(\mathbf{h}_h(\beta))^{1/2}}. \tag{5}$$

This is the same formula as McIsaac and Cook[13] who derived by directly minimising the estimated variance. In this work, we use the integer-valued algorithm[18] to find a global optimal allocation which is slightly more efficient than simply rounding off to the nearest integer.

3.2 | Multi-wave sampling with priors

The influence functions depend on the primary variable of interest X, and we do not have any information about it at the design stage. McIsaac and Cook[13] showed that a multi-wave sampling was helpful. Based on their ideas, the wave 1 can be sampled with pre-specified sampling probabilities, and the influence functions can then be estimated.

The efficiency gains can be realized by finding a better choice of wave 1. On the one hand, any pre-specified sampling probabilities may be far from optimal, so bad decisions of wave 1 will oversample some less informative individuals. On the other hand, as we want the influence functions $\mathbf{h}(\beta)$ but end up with having the estimated influence functions $\mathbf{h}(\hat{\beta})$, a relatively small sample size may lead to the influence functions to be poorly estimated.

If we have reasonable informative priors on the parameters in both outcome and imputation model, the influence functions can be derived by combining phase-1 data and priors. An efficient wave-1 sampling allocation can be estimated by Equation (5). After wave 1, the efficiency gains can be realized by using the posterior distribution obtained from the previous wave as the priors for the current wave. Based on this idea, we propose the following optimal multi-wave sampling design:

1. Combine priors, phase-1 data, outcome model and imputation model to compute posterior distributions for α, β, and X.

2. Impute X for all the cohort subjects and estimate the influence functions.

3. Derive the optimal wave-1 sampling allocations using integer-valued Neyman allocation[18] and sample wave 1.

4. The posterior distributions of α and β obtained from wave 1 are used as priors at wave 2. Repeat steps 1–3 to sample wave 2. Note that during the wave-2 sampling process, we need to put in wave-1 data when computing the posterior distributions.

The later waves can be sampled adaptively if needed. We also add a constraint $n_i \geq 2$ for $i = 1, 2, \ldots, H$ in the wave-1 sampling process to ensure a valid variance estimation for each stratum.

3.3 | Generalised raking

Generalised raking is a more efficient class of design-based estimators. Suppose the objective is to estimate the population total $T_x = \sum_{i=1}^{N} X_i$ and the population totals of a vector of auxiliary variables S_i are known. The idea of generalised raking is to adjust weights so that the estimated population totals of S_i equal the true population totals. The generalised raking estimator is $T_{xr} = \sum_{i=1}^{N} R_i w_i^* X_i$, where $w_i^* = g_i/\pi_i$ are calibrated weights. The calibration constraints can be written as

$$\sum_{i=1}^{N} R_i w_i^* S_i = \sum_{i=1}^{N} S_i \,. \tag{6}$$

The calibrated weights w_i^* can be obtained by minimising the total weight change under a given distance measure

$$\sum_{i=1}^{N} R_i d\left(\frac{g_i}{\pi_i}, \frac{1}{\pi_i}\right),$$

while satisfying the calibration constraints.[3] We use distance function $d(a,b) = a\log(a/b) + (b - a)$ in this work.

In two-phase designs, we are interested in improving the efficiency of regression parameters in the outcome model. The variables in the outcome model cannot be directly used as generalised raking variables because the generalised raking variables should be linearly correlated with regression parameters $\hat{\beta}$, and X and Y are approximately uncorrelated with $\mathbf{h}_i(\beta)$.[4] According to Equation (3) and (4), influence functions can be used as generalised raking variables. The efficient design-based estimators use $[\mathbf{h}_i(\beta_0)|A, Y, Z]$ as the auxiliary variable. This is the same class as AIPW[19] estimators.[4]

Kulich and Lin[20] derived an efficient doubly weighted estimator and proposed a “plug-in” method to approximate the optimal choice of auxiliary variables. Breslow et al.[14,21] adopted the “plug-in” method to conduct imputation generalised raking for case-cohort studies; Rivera and Lumley[22] used the same method in the analysis of counter-matched samples. We use the same technique to get the generalised raking variables in our final statistical analysis. The procedures are described as follows:

1. Fit imputation models using phase-2 data to impute the partially missing variables X for all individuals.
2. Fit the outcome model using phase-1 data and the imputed values of X, and then compute the estimated influence functions $\mathbf{h}_i(\hat{\beta})$ for all individuals.
3. Using the estimated influence functions $\mathbf{h}_i(\hat{\beta})$ as auxiliary variables in generalised raking, and estimate the parameter of interest β by weighted likelihood using the calibrated weights.

It is worth to note that, priors are not used in the statistical analysis because currently it is not standard in these fields to do a Bayesian analysis. Arguably even a better option would be to do the Bayesian analysis, but we show that we can still gain from prior information in the design even if we cannot do that. The generalised raking is only used in the final statistical analysis and not used during the sampling process.

4 | SIMULATION STUDY

We conducted extensive simulation studies to evaluate the efficiency of our proposed sampling design. We examined the situation that the exposure of interest X is cost-prohibitive, but there exists an inexpensive surrogate variable for it.

1000 phase-1 samples of size 1000 were simulated. A binary variable of interest X was generated with 15% exposure, so $X \sim \text{Bern}(0.15)$. A surrogate variable A was simulated with pre-specified sensitivity and specificity. We also simulated a continuous covariate $Z_1 \sim U(0,1)$ and a binary covariate $Z_2 \sim \text{Bern}(0.6)$. A binary outcome variable Y was simulated using the outcome model

$$P(Y \mid X, Z_1, Z_2) = \text{expit}(\beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2),$$

where $\text{expit}(X) = \exp(X)/(1 + \exp(X))$, $\beta_0 = -2$ and $\beta_2 = \beta_3 = 1$. The imputation model was $(X|A,Z_1,Z_2)$.

The data were divided into 8 strata based on Z_2, A and Y. We were interested in priors centered either close to or far from the truth with small or moderate variances. Typically, two-wave sampling designs were considered because they were widely used and relatively easy to implement. We implemented the following designs in our simulation studies:

1. A single-wave proportional stratified sampling design where phase-2 strata sizes were proportional to phase-1 strata sizes.
2. A single-wave balanced stratified sampling design where phase-2 strata sizes were the same.
3. An optimal sampling design where the sampling allocations were derived using the whole data (X also available for individuals not in phase-2 sample). This cannot be done in practice.
4. Two-wave sampling designs where balanced or proportional stratified sampling was used at wave 1.
5. Our proposed two-wave sampling designs (Section 3.2) where informative normal priors on the parameters of the outcome model and imputation model were used. 4 different normal priors were considered, specifically,
   - prior 1 represented a well-calibrated tight prior with $\beta_i \sim N\left(\beta_i - \sqrt{0.1}/2, 0.1\right)$, $\alpha_j \sim N\left(\alpha_j - \sqrt{0.1}/2, 0.1\right)$;
   - prior 2 represented a well-calibrated flat prior with $\beta_i \sim N\left(\beta_i - \sqrt{0.1}/2, 1\right)$, $\alpha_j \sim N\left(\alpha_j - \sqrt{0.1}/2, 1\right)$;
   - prior 3 represented a poorly-calibrated tight prior with $\beta_i \sim N(\beta_i - 1/2, 0.1)$, $\alpha_j \sim N(\alpha_j - 1/2, 0.1)$;
   - prior 4 represented a poorly-calibrated flat prior with $\beta_i \sim N(\beta_i - 1/2, 1)$, $\alpha_j \sim N(\alpha_j - 1/2, 1)$.

Let n_a be the number of individuals sampled at wave 1. For two-wave sampling designs, we considered 5 different choices for the proportion of phase-2 samples selected at wave 1 (n_a/n), which ranged from 1/6 to 5/6 in 1/6 increments. In the statistical analysis, generalised raking described in Section 3.3 was followed.

Results were presented in terms of Mean Squared Error (MSE) and Empirical Relative Efficiency (ERE) to the optimal design for the parameter of interest β_1. MSE is the average squared difference between the estimated values $\hat{\beta}_1$ and the actual value β_1. Larger values of MSE indicate lower efficiencies. ERE of design $\mathscr{D}$ is defined as the ratio of the empirical variance of $\hat{\beta}_1$ from the optimal design to the empirical variance of $\hat{\beta}_1$ from design $\mathscr{D}$.[13] Values of ERE smaller than 1 indicate a loss of efficiency compared with the optimal design.

Results were shown in Table 1. Two-wave sampling designs were slightly more efficient than optimal design in some settings, these were consistent with the simulation studies of McIsaac and Cook[13], because the underlying optimal designs are only optimal when sample size goes to infinity. Besides that, Neyman allocation is not the optimal design for the generalised raking estimator. However, the optimal design of the generalised raking estimators requires the true influence functions, which are not available in practice.

In our simulation studies, single-wave balanced stratified sampling designs were more efficient than single-wave proportional stratified sampling designs, but single-wave sampling designs did not achieve near optimality.

Two-wave sampling designs with pre-specified wave 1 generally performed better than single-wave sampling designs except in the situation that wave 1 sample size n_a was small, because influence functions tended to be poorly estimated with a small amount of phase-2 data. With the increase of wave-1 sample size, we got more precise estimates of influence functions which would lead to a better wave-2 sampling design, but the disadvantage was that we also had to sample fewer people at wave 2. When wave-1 sample size was around half ($n_a \approx n/2$), two-wave designs had a better performance and were more efficient than single-wave sampling designs. These findings were a confirmation of findings in McIsaac and Cook,[13] they showed there is a bias-variance trade-off when deciding the wave-1 sample size n_a for a two-wave design and recommended to sample nearly the same number of individuals at wave 1 and wave 2. In addition to the choice of n_a, the wave-1 sampling probabilities were also essential. In our simulation studies, single-wave balanced sampling was even more efficient than a two-wave design with proportional sampling at wave 1.

Table 1 showed that our proposed designs were very close to the optimal design for the priors centered either close to or far from the truth with small or moderate variances. This indicated that all the 4 priors resulted in good wave 1 allocations, so a small wave 1 sample size n_a did not lose efficiencies with reasonable informative priors.

Other simulation studies (not shown here) showed that when the normal priors centered exactly at the truth, tight priors did slightly better than flat priors, and wave-1 sample size n_a did not affect the design efficiency. However, when the normal priors centered at wrong values (α and β all centered at −1.5 for the simulation studies in this section), flat priors performed better than tight priors, and they were all worse than the single-wave balanced sampling design. A smaller wave-1 sample size n_a was also preferable under this circumstance. In practice, we will not have priors either centered at the truth or wrong values, but these simulations indicated tight priors had some risks. Furthermore, as weakly informative priors prevented us from getting extreme inference,[23] we recommended to use weakly informative priors.

5 | DATA EXAMPLE

We illustrated the performance of our proposed sampling design using data from the National Wilms' Tumor Study (NWTS).[15,16] The data consisted of $N = 3915$ observations. The variables available for all individuals included histology evaluated by the institution (favorable vs. unfavorable (instit)), histology evaluated by the central lab (favorable vs. unfavorable (histol)), stage of disease (I-IV (stage)), age at diagnosis (age), diameter of tumor (tumdiam), study (3 vs. 4 (study)) and indicator of relapse (relapse). We assumed central lab histology was only available at phase 2 and was the variable of primary interest. All the other variables were assumed to be available for the whole cohort. We fitted a similar outcome model[20,21,22]:

$$P(\text{relapse}|\text{histol}, \text{age}_1, \text{age}_2, \text{stage}_1, \text{tumdiam}) = \text{expit}(\beta_0 + \beta_1 \text{histol} + \beta_2 \text{age}_1 \\ + \beta_3 \text{age}_2 + \beta_4 \text{stage}_1 + \beta_5 \text{tumdiam} + \beta_6 \text{tumdiam} \times \text{stage}_1),$$

where age_1 and age_2 were a linear spline with separate slope for greater or less than 1 year old and stage_1 was a binary indicator (III–IV vs. I–II). We took institutional histology as central lab histology measured with error, so it was a good surrogate variable. Central lab histology was imputed using a logistic model with predictors institutional histology, age_3 (>10 years vs. <10 years), stage_2 (IV vs. I–III), study and the interaction between study and stage_2.

The data were divided into 8 strata based on institutional histology, relapse and study with strata sizes (1257, 1769, 107, 113, 223, 284, 84, 78). In the simulation study, 720 individuals were sampled at phase 2. Based on above outcome model and the whole cohort data, the optimal phase-2 sample sizes for each stratum were $n_{opt} = (156, 241, 38, 39, 75, 111, 36, 24)$.

1000 phase-2 samples were simulated. Similarly to previous simulation studies, we examined five choices of wave-1 sample size n_a and implemented a single-wave balanced sampling design, a single-wave optimal sampling design based on the full data, a two-wave sampling design with balanced sampling at wave 1 and our proposed sampling designs. We considered 4 different normal priors, specifically,

- prior 1 represented a well-calibrated tight prior with $\beta_i \sim N\left(\beta_i - \sqrt{0.1}/2, 0.1\right)$, $\alpha_j \sim N\left(\alpha_j - \sqrt{0.1}/2, 0.1\right)$;
- prior 2 represented a well-calibrated flat prior with $\beta_i \sim N\left(\beta_i - \sqrt{0.1}/2, 1\right)$, $\alpha_j \sim N\left(\alpha_j - \sqrt{0.1}/2, 1\right)$;
- prior 3 represented a poorly-calibrated tight prior with $\beta_i \sim N(\beta_i - 1/2, 0.1)$, $\alpha_j \sim N(\alpha_j - 1/2, 0.1)$;
- prior 4 represented a poorly-calibrated flat prior with $\beta_i \sim N(\beta_i - 1/2, 1)$, $\alpha_j \sim N(\alpha_j - 1/2, 1)$.

Proportional stratified sampling was not considered because the sampling probabilities for each stratum were (0.32, 0.45, 0.03, 0.03, 0.06, 0.07, 0.02, 0.02). If 100 individuals were sampled at wave 1, it would only sample 2 individuals from seventh and eighth strata, so the influence functions would be very poorly estimated and the variance of influence functions for these strata might not exist.

Results were presented in Table 2. Single-wave balanced stratified sampling design was not close to optimal. Two-wave sampling with balanced sampling at wave 1 performed slightly better but still was not close to optimal for all the choices of wave-1 sample size. Our proposed designs were still very close to the optimal design for all the 4 priors. As the NWTS data had rich phase-1 information, efficiency gains of our proposed design were also from using the rich phase-1 data at the design stage.

6 | DISCUSSION

We describe a multi-wave adaptive sampling approach to approximate the optimal two-phase design for fitting a regression model using design-based estimators. The prior knowledge of parameters and phase-1 data combine to be usable to obtain an efficient wave 1, so we use the whole cohort information at the design stage even before phase-2 sampling. After wave 1, we propose to use the posterior distributions obtained from the previous wave as priors for the current wave. With reasonable well-calibrated informative priors, our proposed design is very close to the underlying optimal design.

There are two main advantages of using priors. It is obviously that priors help us put in available information. Based on analysing the rich and readily available medical data (e.g., electronic health record), genuine clinical knowledge and previous studies, it is reasonable to have useful prior knowledge. Moreover, even if we do not have much information, weakly informative priors are also found to be useful because they regularise extreme estimations in the analysis of each wave which occasionally happen if we are using completely non-informative priors or maximum likelihood.[23]

However, there is a bias-variance trade-off in the design process, since the true parameters are unknown. Over hypothetical repetitions of the design procedure, a stronger prior will lead to designs that are less variable, but are optimised for a parameter value that is influenced more strongly by the prior. If the prior is poorly-calibrated, these parameter values may be far from the true value. Conversely, a weaker prior leads to designs targeting parameter values that are more variable, but are less biased. Stronger priors are valuable when they are well-calibrated, but as we have shown, relatively weak priors can be valuable even if they are not well-calibrated. For this reason, we argue that the bias-variance trade-off leans towards relatively weak priors.

In addition to the choice of prior, the wave-1 sample size n_a needs to be decided. If the priors are well-calibrated, wave-1 sample size does not matter because the wave-1 design is efficient and close to optimal. If the priors are poorly-calibrated but strong, as discussed in Section 4, wave-1 design is far from optimal and this indicates a small n_a is preferable because it will give us more chance to learn from data. In practice, we are less likely to have priors either centered at the truth or values that are obviously wrong, so that a relatively small wave 1 is preferable in general. The ability to use a relatively small first wave is one of the benefits of using priors.

Our results confirm that single-wave sampling designs are not efficient in general. Balanced stratified sampling designs are more efficient than proportional stratified sampling designs, but still do not often achieve near optimality.[2] Like McIsaac and Cook,[13] we find that multi-wave sampling can improve over single-wave sampling. One contribution of our work is to show that McIsaac and Cook's optimal allocation is the same as Neyman allocation with influence functions.

We have treated the strata as prespecified constraints. There is potential for improving the design by optimising the choice of strata. For a single-wave two-phase sampling design, the design efficiency increases with the increase in the number of strata, and the optimal stratum size is two. The situation is more complicated for multi-wave designs, since a parameter must be estimated for each stratum to construct each wave of the design and very small strata are undesirable. Even so, optimisation of the stratum boundaries is likely to give improved designs. Prior information of the form we used here will be valuable in constructing strata, and we plan to consider choice of stratum boundaries in future work.

ACKNOWLEDGMENTS

The authors thank the reviewer for the helpful comments.

This work was supported in part by the University of Auckland doctoral scholarship, the Patient-Centered Outcomes Research Institute, Award Number: R-1609-36207 and U.S. National Institutes of Health, Award Number: R01-AI131771.

The authors wish to acknowledge the use of New Zealand eScience Infrastructure (NeSI) high performance computing facilities, consulting support and/or training services as part of this research. New Zealand's national facilities are provided by NeSI and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation & Employment's Research Infrastructure programme. URL https://www.nesi.org.nz.

References

1. Neyman J Contribution to the theory of sampling human populations. J Am Stat Assoc. 1938; 33(201): 101–116.
2. Breslow NE, Chatterjee N. Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. J R Stat Soc Ser C Appl Stat. 1999; 48(4): 457–468.
3. Deville JC, Särndal CE. Calibration estimators in survey sampling. J Am Stat Assoc. 1992; 87(418): 376–382.
4. 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. [PubMed: 23833390]
5. Scott AJ, Wild CJ. Fitting regression models to case-control data by maximum likelihood. Biometrika. 1997; 84(1): 57–71.
6. Breslow NE, Holubkov R. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. J R Stat Soc Series B Stat Methodol. 1997; 59(2): 447–461.
7. 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. [PubMed: 29479125]

8. Lumley T Robustness of semiparametric efficiency in nearly-true models for two-phase samples. ArXiv. 2017; arXiv:1707.05924.

9. Tao R, Zeng D, Lin DY. Optimal Designs of Two-Phase Studies. J Am Stat Assoc. 2019; 0(0): 1–14.

10. 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.

11. Reilly M, Pepe MS. A mean score method for missing and auxiliary covariate data in regression models. Biometrika. 1995; 82(2): 299–314.

12. Reilly M Optimal sampling strategies for two-stage studies. Am J Epidemiol. 1996; 143(1): 92–100. [PubMed: 8533752]

13. McIsaac MA, Cook RJ. Adaptive sampling in two-phase designs: a biomarker study for progression in arthritis. Stat Med. 2015; 34(21): 2899–2912. [PubMed: 25951124]

14. 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. [PubMed: 20174455]

15. 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. [PubMed: 2544249]

16. 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. [PubMed: 9440748]

17. Rubin DB. Inference and missing data. Biometrika. 1976; 63(3): 581–592.

18. Wright T Exact optimal sample allocation: More efficient than Neyman. Stat Probab Lett. 2017; 129: 50–57.

19. 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.

20. Kulich M, Lin D. Improving the efficiency of relative-risk estimation in case-cohort studies. J Am Stat Assoc. 2004; 99(467): 832–844.

21. 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. [PubMed: 19357328]

22. Rivera C, Lumley T. Using the whole cohort in the analysis of countermatched samples. Biometrics. 2016; 72(2): 382–391. [PubMed: 26393818]

23. Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. Ann Appl Stat. 2008; 2(4): 1360–1383.

TABLE 1

Mean squared error (MSE) and empirical relative efficiency (ERE) to the optimal design for β_1 based on 1000 Monte Carlo simulations.

		Opt	Prior 1		Prior 2		Prior 3		Prior 4		Two.prop		Two.bal		Single.prop		Single.bal	
(β_1, *se*, *sp*)	***n_a/n***	**MSE***	**MSE***	**ERE**	**MSE***	**ERE**	**MSE***	**ERE**	**MSE***	**ERE**	**MSE***	**ERE**	**MSE***	**ERE**	**MSE***	**ERE**	**MSE***	**ERE**
(0.5, 0.8, 0.8)	1/6	0.74	0.79	0.94	0.80	0.93	0.82	0.91	0.77	0.96	2.32	0.32	1.54	0.49	1.07	0.70	0.88	0.85
	2/6	0.74	0.87	0.86	0.83	0.90	0.77	0.96	0.78	0.95	1.09	0.68	0.95	0.79	1.07	0.70	0.88	0.85
	3/6	0.74	0.74	1.00	0.77	0.97	0.86	0.87	0.82	0.91	1.11	0.67	0.89	0.84	1.07	0.70	0.88	0.85
	4/6	0.74	0.81	0.92	0.77	0.96	0.84	0.88	0.87	0.85	0.92	0.80	0.82	0.90	1.07	0.70	0.88	0.85
	5/6	0.74	0.81	0.92	0.76	0.98	0.84	0.88	0.84	0.88	1.00	0.74	0.84	0.89	1.07	0.70	0.88	0.85
(0.5, 0.9, 0.9)	1/6	0.54	0.55	0.99	0.59	0.92	0.59	0.92	0.55	0.99	1.14	0.48	0.67	0.81	0.84	0.65	0.62	0.88
	2/6	0.54	0.57	0.94	0.57	0.96	0.59	0.93	0.58	0.94	0.65	0.84	0.63	0.86	0.84	0.65	0.62	0.88
	3/6	0.54	0.58	0.94	0.58	0.94	0.54	1.00	0.56	0.96	0.65	0.83	0.59	0.93	0.84	0.65	0.62	0.88
	4/6	0.54	0.56	0.96	0.60	0.91	0.62	0.88	0.55	0.99	0.70	0.78	0.60	0.92	0.84	0.65	0.62	0.88
	5/6	0.54	0.58	0.93	0.56	0.97	0.62	0.87	0.60	0.90	0.73	0.74	0.58	0.94	0.84	0.65	0.62	0.88
(1, 0.8, 0.8)	1/6	0.78	0.79	0.99	0.78	1.00	0.82	0.94	0.83	0.94	2.48	0.33	1.65	0.50	1.04	0.74	0.90	0.87
	2/6	0.78	0.81	0.96	0.77	1.01	0.89	0.87	0.77	1.01	1.19	0.66	0.98	0.82	1.04	0.74	0.90	0.87
	3/6	0.78	0.74	1.05	0.82	0.95	0.85	0.91	0.82	0.94	1.02	0.76	0.91	0.89	1.04	0.74	0.90	0.87
	4/6	0.78	0.80	0.97	0.82	0.94	0.83	0.94	0.81	0.96	0.96	0.80	0.90	0.88	1.04	0.74	0.90	0.87
	5/6	0.78	0.83	0.94	0.80	0.97	0.78	0.99	0.84	0.92	1.04	0.75	0.89	0.87	1.04	0.74	0.90	0.87
(1, 0.9, 0.9)	1/6	0.57	0.58	0.98	0.62	0.91	0.58	0.98	0.59	0.96	4.58	0.13	0.75	0.77	0.81	0.70	0.61	0.94
	2/6	0.57	0.59	0.96	0.55	1.03	0.59	0.96	0.58	0.98	0.78	0.75	0.61	0.93	0.81	0.70	0.61	0.94
	3/6	0.57	0.57	1.00	0.56	1.02	0.61	0.93	0.58	0.98	0.63	0.92	0.58	1.00	0.81	0.70	0.61	0.94
	4/6	0.57	0.58	0.98	0.59	0.96	0.60	0.96	0.60	0.95	0.74	0.77	0.54	1.07	0.81	0.70	0.61	0.94
	5/6	0.57	0.60	0.94	0.58	0.98	0.61	0.93	0.59	0.97	0.76	0.75	0.62	0.92	0.81	0.70	0.61	0.94
(1.5, 0.8, 0.8)	1/6	0.88	0.93	0.95	0.89	0.98	0.98	0.90	0.92	0.95	9.69	0.10	4.87	0.19	1.20	0.73	1.05	0.84
	2/6	0.88	0.92	0.96	0.90	0.97	0.91	0.97	0.92	0.96	1.79	0.54	1.32	0.72	1.20	0.73	1.05	0.84
	3/6	0.88	0.91	0.97	0.92	0.96	0.89	0.98	0.84	1.04	1.19	0.76	1.06	0.91	1.20	0.73	1.05	0.84
	4/6	0.88	0.85	1.03	0.94	0.93	0.95	0.92	0.88	0.99	1.15	0.77	0.99	0.93	1.20	0.73	1.05	0.84
	5/6	0.88	0.90	0.98	0.93	0.95	1.02	0.86	0.93	0.95	1.07	0.82	1.02	0.88	1.20	0.73	1.05	0.84
(1.5, 0.9, 0.9)	1/6	0.59	0.58	1.00	0.59	1.00	0.60	0.98	0.59	0.99	8.53	0.07	0.77	0.79	0.91	0.64	0.63	0.92

		Opt	Prior 1		Prior 2		Prior 3		Prior 4		Two.prop		Two.bal		Single.prop		Single.bal	
(*β_1*, *se*, *sp*)	***n_a*/ *n***	**MSE***	**MSE***	**ERE**	**MSE***	**ERE**	**MSE***	**ERE**	**MSE***	**ERE**	**MSE***	**ERE**	**MSE***	**ERE**	**MSE***	**ERE**	**MSE***	**ERE**
	2/6	0.59	0.63	0.93	0.59	0.99	0.60	0.97	0.60	0.97	1.09	0.58	0.61	0.98	0.91	0.64	0.63	0.92
	3/6	0.59	0.58	1.02	0.56	1.04	0.60	0.98	0.58	1.02	0.75	0.80	0.58	1.04	0.91	0.64	0.63	0.92
	4/6	0.59	0.59	0.99	0.61	0.96	0.64	0.91	0.59	1.00	0.67	0.88	0.60	0.99	0.91	0.64	0.63	0.92
	5/6	0.59	0.62	0.94	0.63	0.93	0.61	0.97	0.64	0.91	0.82	0.71	0.59	0.99	0.91	0.64	0.63	0.92

MSE*: MSE×10; Se, sensitivity used to generate auxiliary variable A; Sp, specificity used to generate auxiliary variable A; Opt, optimal design based on the full data; Two.prop, a two-wave design with proportional stratified sampling at wave 1; Two.bal, a two-wave design with balanced stratified sampling at wave 1; Single.prop, a single-wave proportional stratified sampling design; Single.bal, a single-wave balanced stratified sampling design.

TABLE 2

Mean squared error (MSE) and empirical relative efficiency (ERE) to the optimal design for β_1 based on 1000 Monte Carlo simulations.

	Opt	Prior 1		Prior 2		Prior 3		Prior 4		Two.bal		Single.bal	
n_α/n	MSE*	MSE*	ERE	MSE*	ERE	MSE*	ERE	MSE*	ERE	MSE*	ERE	MSE*	ERE
1/6	0.16	0.15	1.03	0.17	0.92	0.17	0.92	0.16	0.96	0.29	0.56	0.26	0.60
2/6	0.16	0.17	0.91	0.17	0.93	0.17	0.91	0.16	0.98	0.26	0.66	0.26	0.60
3/6	0.16	0.17	0.92	0.17	0.94	0.16	1.00	0.16	0.96	0.25	0.67	0.26	0.60
4/6	0.16	0.17	0.94	0.16	0.95	0.17	0.96	0.15	1.04	0.26	0.62	0.26	0.60
5/6	0.16	0.18	0.89	0.17	0.94	0.17	0.90	0.18	0.90	0.26	0.62	0.26	0.60

MSE*: MSE ×10; Opt, optimal design based on the full data; Two.bal, a two-wave design with balanced stratified sampling at wave 1; Single.bal, a single-wave balanced stratified sampling design.