Optimal Auxiliary-Covariate Based Two-Phase Sampling Design for Semiparametric Efficient Estimation of a Mean or Mean Difference, with Application to Clinical Trials

Peter B. Gilbert[1,2,*], **Xuesong Yu**[1], **Andrea Rotnitzky**[3,4]

[1] Vaccine Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, 98109, U.S.A.

[2] Department of Biostatistics, University of Washington, Seattle, Washington, 98105, U.S.A.

[3] Department of Economics, Universidad Di Tella, Buenos Aires, Argentina

[4] Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, 02115, U.S.A.

[*] Fred Hutchinson Cancer Research Center
1100 Fairview Ave North, PO Box 19024
Seattle, WA 98109
Telephone: (206) 667 7299
Fax: (206) 667 4378
e-mail: pgilbert@scharp.org

Short title: Optimal two-phase design and analysis for clinical trials

Abstract: To address the objective in a clinical trial to estimate the mean or mean difference of an expensive endpoint Y, one approach employs a two-phase sampling design, wherein inexpensive auxiliary variables W predictive of Y are measured in everyone, Y is measured in a random sample, and the semi-parametric efficient estimator is applied. This approach is made efficient by specifying the phase-two selection probabilities as optimal functions of the auxiliary variables and measurement costs. While this approach is familiar to survey samplers, it apparently has seldom been used in clinical trials, and several novel results practicable for clinical trials are developed. Simulations are performed to identify settings where the optimal approach significantly improves efficiency compared to approaches in current practice. Proofs and R code are provided.

The optimality results are developed to design an HIV vaccine trial, with objective to compare the mean "importance-weighted" breadth (Y) of the T cell response between randomized vaccine groups. The trial collects an auxiliary response (W) highly predictive of Y, and measures Y in the optimal subset. We show that the optimal design-estimation approach can confer anywhere between absent and large efficiency gain (up to 24% in the examples) compared to the approach with the same efficient estimator but simple random sampling, where greater variability in the cost-standardized conditional variance of Y given W yields greater efficiency gains. Accurate estimation of $E[Y|W]$ is important for realizing the efficiency gain, which is aided by an ample phase-two sample and by using a robust fitting method.

KEY WORDS: Augmented inverse probability weighting; Efficient estimation; Efficient sampling; Missing data; Semiparametric model; Two-phase sampling.

1. Introduction

Consider a study with objective to estimate the mean of an expensive outcome Y based on a sample of individuals. Suppose inexpensive auxiliary covariates and/or response outcomes W predictive of Y are available. An efficient and robust approach to meeting the objective will measure the auxiliaries in everyone and measure the outcome in an optimally chosen sub-set, and then estimate the mean using a semi-parametric efficient approach that provides consistent estimation without parametric assumptions. Here we show how to optimally design a two-phase study using this approach, developing several novel results that account for costs of phase 1 and 2 measurements. In addition to addressing the one-sample problem, these results provide optimal two-phase designs for comparing the mean of an expensive outcome between two groups, which are of particular interest for clinical trials. In practice sub-optimal sampling designs and estimators are frequently used; our objective is to encourage use of the efficient trial design coupled with the efficient estimator for settings where it is advantageous.

The problem addressed here is different from the problem of "efficient two-stage clinical trial design;" such two-stage trials first assess the treatment effect on the primary endpoint in an initial cohort of individuals (stage one), and, based on the results, adaptively decide whether to enroll an additional cohort of individuals to increase the total sample size for assessing the treatment effect (e.g., [1]). Instead, our problem considers a clinical trial with fixed sample size, and the relevant clinical trials statistical literature is that of "two-phase designs" (cf., $[2-7]$), where the phase-one data are variables collected from all study participants, and the phase-two data are the expensive variable(s) collected in a judiciously chosen sub-set of participants. Whereas the articles cited above and others focus on more efficient estimation, very few have combined efficient two-phase sampling design with estimation, and none to our knowledge have tackled this problem for the case where the phase-two variable of interest is the primary endpoint [most of the literature is related to the case-control design (e.g., [8]) or to the case-cohort design originally proposed by Prentice (1986) [9], where the phase-two variables are expensive exposure covariates]. Outside of the clinical trials statistical literature, survey samplers have tackled this problem, and below we summarize how our work fits in that context.

This research is motivated by AIDS vaccine development. Development of an AIDS vaccine administered to HIV-free volunteers that prevents HIV infection is a global public health priority [10]. A central objective of clinical trials of current HIV vaccine candidates is detection and characterization of vaccine-induced T cells that react with HIV "epitopes"– short K-mer peptides of $K = 8-12$ contiguous amino acids (e.g., RLRPGGKKK). Two study endpoints of particular interest are the "breadth" and the "importance-weighted breadth;" breadth is the number of reactive HIV epitopes and weighted breadth is the sum of "importance weights" attached to the reactive epitopes, where importance reflects knowledge about the usefulness of the epitope for potentially contributing to protection. (The methods of $[11-13]$ are used to measure T cell reactions.) Clinical trials of HIV vaccine

candidates conducted by the U.S. NIH-funded HIV Vaccine Trials Network (HVTN) use breadth (W) and importance-weighted breadth (Y) as study endpoints, where Y is expensive to measure. While the high expense makes it cost-prohibitive to measure Y for every subject, the fact that W predicts Y provides an opportunity to effectively use a two-phase design. This case study uses these HVTN 054 data as pilot data for determining an optimal sampling design for HVTN protocol 083.

For two-phase clinical trials like HVTN 083 decribed later, Rotnitzky and Robin's ([6], henceforth RR) semiparametric efficient estimator of a group mean is asymptotically optimal, and we consider optimal sampling and estimation based on the RR estimator. Methods for optimizing the sampling design for estimation of a group mean have been developed in the survey sampling literature, for example $[14-17]$. However, this literature addresses a different goal, and uses a different perspective, than the goal and perspective of the statistical framework used in clinical trials. In particular, survey samplers are interested in finite-population estimands such as the average Y for a fixed population, whereas clinical trialists are concerned with "superpopulation" estimands such as the mean of Y, defined for a hypothetical infinite superpopulation from which the finite population is a sample. For clinical trials it is of little scientific interest to compare finite-population means between groups (which will virtually always differ); rather scientific interest centers on comparing superpopulation means that may be equal.

Many of the results developed in the finite-population framework can be translated to the superpopulation framework, and this is part of our objective. For estimation, Särndal, Swensson, and Wretman ([14], chapter 6) and predecessors (including $[18-20]$) developed a "regression estimator" of a population total that, in special cases for the form of the auxiliaries, is essentially equivalent to RR's estimator of a group mean, which is known to statisticians working in clinical trials. Moreover, for special cases of some of our results presented below, Särndal, Swensson, and Wretman's ([14], chapter 12) optimal sampling design for the regression estimator are equivalent to our optimal sampling designs for the RR estimator. In particular, we develop four results for the one-sample problem (Results 1–4), where Results 3 and 4 are direct translations of results in [14], Result 1 would be a direct extension of a result in [21] if our added constraint of a minimum sample size were dropped, and we are not aware of results in the survey sampling literature that directly extend to Result 2. We are also not aware of results in the survey sampling literature that directly extend to the four corresponding results that we develop for the two-sample problem (Results 1-two, 2-two, 3-two, 4-two). All of the results are practically novel for the clinical trials statistician given the major differences in language and perspective of the two literatures.

In Section 2 we summarize RR's method for estimating a group mean within a two-phase sampling design. In Section 3 we develop results for this objective on optimal sampling of subjects into phase two. In Section 4 we extend these results to the objective of comparing means between two groups. In Section 5 we describe the case study for optimally designing the HVTN 083 two-phase HIV vaccine trial, and

in Section 6 we compare the finite-sample efficiency of different implementations of Result 3 based on simulated vaccine trials. In Section 7 we conclude with discussion. The web-based supporting material provides proofs of the results, R code for implementing them, and additional results.

2. RR Semiparametric Estimation of a Group Mean Using Auxiliaries

2.1. Notation and Problem Set-up

For a two-phase design, let W be a vector of auxiliaries measured in everyone and Y be the outcome. The auxiliary W may have discrete and/or bounded quantitative components. The outcome Y may be discrete or quantitative, with results and estimation methods identical in each case; however, additional research would be needed to extend the results to handle an outcome Y subject to censoring (e.g., a failure time outcome). Let F be the joint cdf of W. Let R be the indicator of whether a subject is selected into phase two for measuring Y, and define $\lambda(W) \equiv \Pr(R = 1|W)$. Our motivating application has an unusual feature that there is a subgroup for which the value of W completely determines the value of Y ($W = 0$ implies $Y = 0$), such that Y is known without measurement. For such subgroups, we use the convention that $R = 1$ and $\lambda(W) = 1$ (Sections 3.1 and 3.7 provide further discussion). The observed data are n iid copies (W_i, R_i, R_iY_i), $i = 1, \cdots, n$. Let $\lambda_i \equiv \lambda(W_i)$. Our goal is estimation of the mean of Y, $\beta \equiv E[Y]$.

In our HIV vaccine trial example W and Y are measured at the same time-point after randomization, which is typically chosen a few weeks after the last immunization. For simplicity, throughout we assume no dropout before the measurement time-point. While missing at random dropout could be handled using the RR methodology it would be distracting to explicitly account for it.

2.2. Semiparametric Efficient Doubly Robust Estimation

To achieve consistent estimation of β with RR's method, the only assumption is that the λ_i's used in the estimator exceed zero for all i. Given any fixed positive λ_i's, the semiparametric efficient estimator of β, $\widehat{\beta}$, solves $\sum_{i=1}^{n} U_i(\beta, \lambda_i) = 0$, where U_i is RR's efficient influence function:

$$U_i(\beta, \lambda) = \frac{R_i}{\lambda_i}(Y_i - \beta) - \frac{(R_i - \lambda_i)}{\lambda_i}\left(E\left[Y_i|W_i\right] - \beta\right). \tag{1}$$

This efficient estimator is infeasible because it depends on the unknown $E[Y|W]$, and in Section 3.6 we discuss estimation in practice, which requires estimation of $E[Y|W]$ based on the phase-two sample. Our optimality results below are based on minimizing the asymptotic variance of $U_i(\beta, \lambda)$, such that they use the true $E[Y|W]$; thus the results are only assured to provide reliable guidance for trials with large sample sizes, and an important practical question is how the need to estimate $E[Y|W]$ in implementing the methods affects their utility for trials with moderate sample sizes. In Section 6 we evaluate this question via simulations by comparing the method in practice (using estimated $E[Y|W]$) with the unachievable benchmark version of the method that uses the true $E[Y|W]$.

The RR method is expected to perform well for HIV vaccine trials for two reasons. First, the occasion where an inverse probability weighted method performs poorly, wherein some estimated weights are large outliers (e.g., [22]), can be avoided because the investigator designs the weights λ_i. Second, the RR method performs best when there exist auxiliaries highly predictive of Y. The fact that the auxiliaries and outcome are measured using the same assay under identical experimental conditions on the same blood sample provides opportunity for high correlations. Consequently, Gilbert et al. [23] advocated use of the RR method for Phase I/II vaccine trials.

3. Optimal Design of the Phase Two Selection Probabilities λ_i

We develop optimality results for four different objectives that occur in practice (Results 1 through 4 below, with proofs in Supporting Materials Appendix A). All of these results determine the optimal function $\lambda(W)$ to use in RR's estimator, and the first two also determine the optimal sample size. The optimal function $\lambda(W)$ depends on the unknown distribution of the data, which will need to be estimated from pilot data. Once estimated, $\lambda(W)$ may be treated as known by design in the analysis, so that the resulting 'true' λ_i's are used in the RR estimator. However, estimation efficiency may be improved by using 'estimated' λ_i's in the RR estimator, as described in Section 3.6.

The optimality results are developed for Bernoulli sampling, wherein each subject i is selected based on a random Bernoulli variate with success probability λ_i. However, because the number of phase two subjects is random, some investigators may prefer without replacement sampling, which affords exact control of this number. In particular, if W has J levels $W_1, \cdots, W_J$, then the results may be implemented via simple random sampling without replacement within each level, such that exactly $\lambda(W_j)$ percent of subjects (after rounding) are selected at level j. This procedure is similar to the "equal aggregate σ rule" of ([14], Section 12.4), which is described in Supporting Materials Appendix C.

3.1. Minimizing Variance Given Expected Total Cost

The first optimality result minimizes $V(\lambda)/n$ over both λ and n given a fixed expected total budget B and n required to be greater than or equal to a fixed number n_0, where $V(\lambda)$ is the semiparametric efficient variance bound for β (n_0 may be 1, but practically one may require at least $n_0 = 10$ or 20 subjects). This bound is calculated as the variance of the efficient influence function (1), and equals

$$V(\lambda) = \mathrm{Var}\left(E\left[\varepsilon|W\right]\right) + E\left[\frac{\mathrm{Var}(\varepsilon|W)}{\lambda(W)}\right] \tag{2}$$

$$= \int E\left[\varepsilon|w\right]^2 dF(w) + \int \frac{\mathrm{Var}(\varepsilon|\mathrm{w})}{\lambda(w)} dF(w), \tag{3}$$

where $\varepsilon \equiv Y - \beta$.

Let C_0 be all initial trial costs that are independent of sample size n, and suppose all other costs scale linearly with n. Of the sample-size-dependent costs, let C_1 be

the per-person costs measured in everyone, and $C_2(W)$ be the per-person costs for measurement of Y for a subject with auxiliary W. With $B(\lambda) \equiv C_1 + E[C_2(W)\lambda(W)]$ the expected per-person sample-size dependent cost, the expected total cost is $B \equiv C_0 + nB(\lambda)$. Let $PVE \equiv \text{Var}(E[Y|W])/\text{Var}(Y)$ be the proportion of the variation in Y explained by W.

Result 1: The minimizer of $V(\lambda)/n$ among all possible designs indexed by (λ, n) that do not exceed the expected total cost B and with $n \geq n_0$ is achieved with

$$\lambda^*(W) = \sqrt{\frac{\text{Var}(Y|W)}{C_2(W)}}\sqrt{\frac{C_1}{\text{Var}(E[Y|W])}} \tag{4}$$

$$= \sqrt{\frac{C_1}{C_2(W)}}\sqrt{\frac{\text{Var}(Y|W)}{\text{Var}(Y)PVE}}, \tag{5}$$

$$n^* = \frac{(B - C_0)}{B(\lambda^*)} = \frac{(B - C_0)/\sqrt{C_1}}{\sqrt{C_1} + E\left[\sqrt{C_2(W)\text{Var}(Y|W)}\right]/\sqrt{\text{Var}(E[Y|W])}}, \tag{6}$$

provided that $\lambda^*(W) \leq 1$ for all W in the support of W and $n^* \geq n_0$. Otherwise, modified formulas described in the Supporting Materials Appendix A are used to achieve the optimal solution with $\lambda^*(W) \leq 1$ for all W and $n^* \geq n_0$. The above formulas assume $\text{Var}(E[Y|W]) > 0$; however, if W is a useless predictor, then $\text{Var}(E[Y|W]) = 0$. In this case, the above formulas are replaced with $\lambda^*(W) = 1$ and $n^* = (B - C_0)/E[\sqrt{C_2(W)}]^2$.

If the constraint $n \geq n_0$ were removed, then formulas (4) and (5) are Cochran's ([21], Section 12.3) result. While the proofs of the results assume W is discrete, they allow for the sample space Ω of W to be partitioned into an arbitrarily fine grid, with $\lambda^*(w)$ constant in each bin on the grid. Because in principal the bins can be made as small as desired (as long as Ω is bounded), the results for discrete W allow using continuous auxiliaries in practice. Moreover, in the Supporting Materials Appendix A we provide an extension of the results when W is continuous, and the simulation study supports veracity of the results for W continuous.

Note that Result 1 controls the expected total cost, but not the actual total cost, which is random. If it is important to control the actual cost, then subjects could be enrolled until the total cost $C_0 + n\left(C_1 + \sum_{i=1}^{n} R_i C_2(W_i)\right)$ meets the budget, in which case formula (6) is not needed and n is random.

Because $\lambda^*(W)$ depends on the unknown data distribution, the λ^*'s are computed by applying either formula (4) or (5), using estimates of the needed expressions from pilot data. The optimal $\lambda^*(W)$ is proportional to the input parameter $r(W) \equiv \sqrt{\text{Var}(Y|W)/C_2(W)}$, the cost-standardized conditional variance of Y given W. As such, the optimal design over-samples subjects with poorly predictive auxiliaries, for whom measurement of W provides relatively little information about Y (i.e.,

$\text{Var}(Y|W)$ is large), and for whom this direct measurement is relatively affordable (i.e., $C_2(W)$ is low). For these subjects directly measuring Y is most informative and cost-effective. In contrast, the optimal design under-samples subjects with highly predictive auxiliaries and/or with high cost for measuring Y; for these subjects less incremental knowledge would be gained by directly measuring Y. The parameter $r(W)$ is also the key input for the other optimality results.

In addition, $\lambda^*(W)$ depends on the cost ratio $C_1/C_2(W)$ and on the PVE. Theorem A.2 of Supporting Materials Appendix A implies that if the subject-specific costs in phase one dwarf those in phase two, then the optimal design will advance all subjects to phase two with probability $\lambda \equiv 1$, and the more costly the phase-two measurements, the more likely a sub-sampling design will be optimal. Furthermore, the larger the PVE, the lower the selection probabilities, again due to good recovery of information about Y from the auxiliaries.

The optimal sample size n^* increases with the expected budget $B - C_0$ left to run the study after the initial cost C_0 is taken into account, and increases with the PVE (better auxiliaries frees up budget for greater sample size). Moreover, n^* decreases with the per-person costs C_1 and $C_2(W)$. However, its dependence on $C_2(W)$ lessens if $E\left[\sqrt{C_2(W)\text{Var}(Y|W)}\right]/\sqrt{\text{Var}(E[Y|W])}$ is small because in this case the auxiliaries are good predictors, which will favor designs with lower selection probabilities.

Depending on the application, it may be more favorable to specify the inputs for (4) or (5). For either approach pilot data on F are needed for computing n^* in (6).

3.2. Minimizing Total Expected Cost Given Fixed Variance

For designing a study, it may be more scientifically relevant to solve the dual problem: Given $V(\lambda)/n$ fixed at a constant V, minimize the total expected cost B. For example, suppose our goal is to give the design power $1-\gamma$ to detect that β is Δ or more units smaller than a fixed value β_0, i.e., to reject $H_0 : \beta \geq \beta_0$ in favor of $H_1 : \beta < \beta_0$ when $\beta = \beta_0 - \Delta$. Based on the standard asymptotic approximation, a 1-sided $\alpha/2$ level test with $(1-\gamma)\%$ power is achieved with $V = \left[\Delta/(z_{1-\gamma} - z_{\alpha/2})\right]^2$, where z_c is the cth percentile of the standard normal distribution.

Result 2: Fix $V(\lambda)/n = V$ and require $n \geq n_0$ for n_0 a fixed counting number. Subject to these constraints the total expected cost B is minimized by $\lambda^*(W)$ the same as in (4) and (5), and

$$\begin{aligned} n^* &= \frac{E\left[\text{Var}\left(Y|W\right)/\lambda^*(W)\right] + \text{Var}\left(E\left[Y|W\right]\right)}{V} \\ &= \frac{\sqrt{\frac{\text{Var}(E[Y|W])}{C_1}}E\left[\sqrt{C_2(W)\text{Var}\left(Y|W\right)}\right] + \text{Var}\left(E\left[Y|W\right]\right)}{V}, \end{aligned} \tag{7}$$

provided that $\lambda^*(W) \leq 1$ for all W in the support of W and $n^* \geq n_0$. Otherwise, modified formulas described in the Supporting Materials Appendix A are used to achieve the optimal solution.

As for Result 1, n^* increases with the PVE, as better auxiliaries frees up budget for a greater sample size.

3.3. Minimizing Variance Given Fixed Expected Phase-Two Budget

Suppose phase one of the study has already been done, such that n is fixed and the goal is to pick the λ's that minimize $V(\lambda)$ subject to a fixed phase-two expected budget $B' = B - C_0 - nC_1 = nE[C_2(W)\lambda(W)]$. To do this, first the design with $\lambda(W) \equiv 1$ is considered, for which we check whether $nE[C_2(W)] \leq B'$. If so, then $\lambda(W) \equiv 1$ is trivially optimal (i.e., the study can afford to measure Y from all subjects). If not, then the following Result 3 is applied.

Result 3: The minimizer of $V(\lambda)$ given fixed n and fixed B' is

$$\lambda^*(W) = \sqrt{\frac{\mathrm{Var}(Y|W)}{C_2(W)}} \frac{B'}{nE\left[\sqrt{C_2(W)\mathrm{Var}(Y|W)}\right]}, \tag{8}$$

provided that $\lambda^*(W) \leq 1$ for all W in the support of W. Otherwise, a modified formula described in the Supporting Materials Appendix A is used to achieve the optimal solution.

3.4. Minimizing Expected Phase-Two Budget Given Fixed Variance

The dual problem of Result 3 is to minimize the phase-two expected budget $B' = nE[C_2(W)\lambda(W)]$ subject to $V(\lambda)/n$ fixed at constant V. Parallel to Result 3, we first consider $\lambda(W) \equiv 1$, and check whether $V(\lambda \equiv 1)/n = \mathrm{Var}(Y)/n \leq V$; if so, then $\lambda(W) \equiv 1$ is trivially optimal. If not, then Result 4 is applied.

Result 4: The minimizer of B' given fixed $V(\lambda)/n = V$ is

$$\lambda^*(W) = \sqrt{\frac{\mathrm{Var}(Y|W)}{C_2(W)}} \frac{E\left[\sqrt{C_2(W)\mathrm{Var}(Y|W)}\right]}{(nV - \mathrm{Var}(E[Y|W]))}. \tag{9}$$

provided that $\lambda^*(W) \leq 1$ for all W in the support of W. Otherwise, a modified formula described in the Supporting Materials Appendix A is used to achieve the optimal solution.

For Results 3 and 4, the larger the sample size n, the smaller the phase-two sample is needed to achieve the same power. Results (8)–(9) are the same as results (12.7.5) and (12.7.6) in [14], translated from the finite-population framework to the superpopulation framework.

3.5. Efficiency Relative to the Optimal Simple Random Sampling Design

To assess the amount of potential improvement conferred by the optimal design compared to the design that is optimal among designs restricting to simple random sampling of subjects into phase two, we computed the relative efficiencies (REs) of these approaches, for each of the results. For Result 1 the RE equals $V(\lambda^*)/n^*$ divided

by $V(\bar{\lambda})/n^*(\bar{\lambda})$, with the constant $\bar{\lambda} = \lambda_i$ and $n^*(\bar{\lambda})$ selected to minimize $V(\bar{\lambda})/n^*(\bar{\lambda})$ subject to the budget constraint; straightforward calculation yields

$$\bar{\lambda} = \sqrt{\frac{E\left[\mathrm{Var}\left(Y|W\right)\right]}{E\left[C_2(W)\right]}}\sqrt{\frac{C_1}{\mathrm{Var}\left(E\left[Y|W\right]\right)}} \tag{10}$$

and $n^*(\bar{\lambda}) = (B - C_0)/(C_1 + \bar{\lambda}E\left[C_2(W)\right])$. For Result 2 the RE equals the variance-constrained minimum B [computed with optimal (λ^*, n^*)] divided by the variance-constrained minimum B for $\bar{\lambda} = \lambda_i$ forced to be constant, with minimizer $\bar{\lambda}$ the same as for Result 1 and $n^*(\bar{\lambda}) = \{\mathrm{Var}(E\left[Y|W\right]) + E\left[\mathrm{Var}(Y|W)/\bar{\lambda}\right]\}/V$. For Results 3 and 4, the constraints defining the optimal simple random sampling designs are met by $\bar{\lambda} = B'/\left(nE\left[C_2(W)\right]\right)$ and $\bar{\lambda} = E\left[\mathrm{Var}\left(Y|W\right)\right]/\left\{nV - \mathrm{Var}\left(E\left[Y|W\right]\right)\right\}$, respectively.

Comparing (10) with (4) from Result 1, provides insight into the fundamental influence of the $r(W)$ function for determining whether and how much efficiency can be gained via optimal auxiliary-dependent sampling. In fact, this comparison shows that the *only* way to improve efficiency over optimal simple random sampling is via variation of $r(W)$ in W, and the example in Section 5 verifies greater efficiency gains via greater variation.

In addition, for each RE calculation, given fixed values of the input parameters $\mathrm{Var}(Y|W)$, $C_2(W)$, and $\mathrm{Var}(Y)$, the relative efficiency of the optimal auxiliary-dependent sampling design improves monotonically with smaller PVE. For example, the RE for Result 3 equals

$$RE = \frac{PVE * \mathrm{Var}(Y) + \frac{n}{B'}\left(E[\sqrt{C_2(W)\mathrm{Var}(Y|W)}]\right)^2}{PVE * \mathrm{Var}(Y) + \frac{n}{B'}E[\mathrm{Var}(Y|W)]E[C_2(W)]}, \tag{11}$$

and by the Cauchy-Schwarz inequality $RE \leq 1$ and RE gets closer to 1 with increasing PVE. This fact implies that if a very good auxiliary is available, then it may be less important to account for the auxiliary in the sampling design, compared to if only a 'good' auxiliary is available (e.g., $PVE = 0.8$ versus 0.5), as in the former case most of the efficiency can be achieved in the estimation alone. This result is observed in our simulation study in Section 6.

While everywhere else this article assumes W is measured in all subjects, it is also of interest to compare the optimal two-phase design to a one-phase design, wherein Y is measured in everyone and W is not measured, given that no prediction of Y is needed. This comparison addresses the question of when is it worth the full cost of measuring Y in everyone versus saving the cost of Y for some subjects but incurring the cost of measuring W? Supporting Materials Appendix B derives the condition determining which design is more efficient, under Results 1 and 2. While there does not appear to be a straightforward condition in general, when $C_2(W)$ and $\mathrm{Var}(Y|W)$ are constant, for both results the one-phase design is superior if and only if

$$\sqrt{PVE}/[1 + \sqrt{1 - PVE}] < \sqrt{C1/C2}.$$

Therefore, the one-phase design is more efficient when the auxiliary is a weak predictor, and the phase one and two costs both influence the tipping point.

3.6. Estimation and Optimization in Practice

For estimating β, it is not possible in practice to solve the efficient score function (1), because it depends on the unknown true regression $E[Y|W]$. Therefore, in practice estimation proceeds in two steps (see [24] for a helpful discussion). First, model selection is used to identify a good-fitting regression model $g(W;\gamma)$ for $E[Y|W]$, which only uses the elements of the full auxiliary vector helpful for the prediction; thus in practice W is restricted to 'useful auxiliaries' through an initial dimensionality reduction step. This model is fit on the phase-two sample (those with $R = 1$). Based on the selected model $g(W;\gamma)$ that depends on parameters γ, a predicted value $\widehat{g}_i \equiv \widehat{E}[Y|W_i] = g(W_i;\widehat{\gamma})$ is computed for each subject $i = 1, \cdots, n$. Second, the equation $\sum_{i=1}^{n} U_i^g(\beta, \lambda_i) = 0$ is solved for β, where

$$U_i^g(\beta, \lambda_i) = \frac{R_i}{\lambda_i}(Y_i - \beta) - \frac{(R_i - \lambda_i)}{\lambda_i}(\widehat{g}_i - \beta). \tag{12}$$

Solutions $\widehat{\beta}$ are consistent and asymptotically normal for any choice of the λ_i as long as they are bounded away from zero. Inferences for β can be based on the sandwich variance estimator or the bootstrap.

Successful implementation of an efficient two-phase study requires joint consideration of the needs of the optimization step and the estimation step. For any of the Results 1–4, the optimization step determines $\lambda^*(W)$ (and n^* for Results 1 and 2) by specifying all of the needed inputs, which include $\mathrm{Var}(Y|W)$ and $C_2(W)$ for all of the results. Subject matter knowledge and pilot data are used to maximize accuracy of the specifications. Success for this task depends crucially on the nature and dimensionality of the L-vector of available auxiliaries W. In particular, if the L components of W each take values on a fine grid then $\mathrm{Var}(Y|W)$ is approximately an L-variate surface, and if L is larger than 1 or 2, a very excellent pilot data-set may be required to adequately describe it. Assuming that all W's are measured in both the pilot and main studies, three implementation approaches are: (1) Base the design on a univariate auxiliary, which is chosen among the potential auxiliary variables as the one that yields the optimal $\lambda^*(W)$ that minimizes the estimated variance bound (estimated based on the pilot data); (2) Base the design on a univariate auxiliary defined as the linear combination of auxiliary variables that minimizes the estimated variance bound based on the pilot data; and (3) Base the design on a multivariate auxiliary, using a parametric model to estimate the input parameter $\mathrm{Var}(Y|W)$ from the pilot data.

For the analysis step, the RR estimator is used with the same W in $g(W;\gamma)$ as was used for specifying the inputs into the optimal design formulas. Thus under design approaches (1) or (2) with a univariate W, $g(W;\gamma)$ could be estimated nonparametrically or parameterically, whereas under design approach (3), the curse of dimensionality implies $g(W;\gamma)$ must be estimated parametrically. The survey-

sampling literature on two-phase design and analysis has focused almost exclusively on approach (1) with W a discrete categorical auxiliary, which is understandable because it is in this setting where pilot data are most often available for adequate specification of the inputs, and, moreover, in this setting $E[Y|W]$ is trivially modeled correctly, which helps ensure good-performance of the RR estimator. Nevertheless, in applications with good pilot data and/or understanding of the subject matter, efficiency may be improved by using a continuous or multivariate W. While our proofs assume a discrete W, as noted above, by allowing for an arbitrary number of levels of W, they may be implemented using continuous or multivariate W. Obviously, the more data available for guiding the inputs, and for checking the assumptions under which optimality is attained, the more likely the chosen optimal design-and-analysis approach will perform well in practice.

For the analysis step, thus far we have considered RR's estimator using the λ_i's chosen from the optimal design step treated as true/known constants. An alternative approach uses estimated λ_i's computed as fitted values from a maximum likelihood fit of the logistic regression model $\text{logit}(\lambda(W;\alpha)) = \text{logit}(\lambda^*(W)) + \alpha^T W$, where $\lambda^*(W)$ is the true $\lambda(\cdot)$, and the corresponding λ_i^*'s are entered into the model-fit as offsets. This approach assures consistent estimation of the λ_i's and can improve efficiency compared to using the true λ_i's. No efficiency is lost when true λ_i's are used if $E[Y|W]$ is correctly modeled. We recommend using estimated λ_i's in practice, as this approach generally provides at least as good efficiency as RR's estimator using true λ_i's (confirmed in the simulation study in Section 6), and sometimes provides improved efficiency. Whether true or estimated λ_i^*'s are used, in practice they should be plotted, to ensure no outliers near zero, which may lead to highly variable estimators.

3.7. Accommodating 'Perfect' Auxiliaries

Certain values of the auxiliaries may perfectly predict or constrain Y. If there exists a constant y_0 and a set $\mathcal{W}$ such that $\Pr(Y = y_0|W \in \mathcal{W}) = 1$, then $\text{Var}(Y|W) = 0$ for all $W \in \mathcal{W}$ such that the optimality formulas $\lambda*(W)$ from any of the Results 1–4 do not apply. This causes no problem, however, because when W perfectly predicts Y, there is no need to measure Y, such that $\lambda^*(W)$ is trivially equal to one. Moreover, the estimation based on (12) is implemented with $U_i^g(\beta,\lambda)$ set to $y_0 - \beta$ and λ_i and R_i set to 1 for subjects with $W_i \in \mathcal{W}$.

4. Optimal Design for the Two-Sample Problem

For the two-sample problem common in clinical trials, each parameter $\beta_l \equiv E[Y_l]$, ($l = 1$: group 1; $l = 2$: group 2) is estimated by RR's estimator, and the goal is estimation of the difference $\beta_1 - \beta_2$. Novel two-sample versions of Results 1–4 can be developed. Now the selection probabilities, sample sizes, and all of the input variables and parameters except $B - C_0$ and B' may differ for the two groups and hence are indexed by a subscript l. Let n_{l0} be the fixed minimal sample size for group l, for $l = 1, 2$. Let $V_l(\lambda_l)$ be the semiparametric efficient variance bound for β_l, and set $V_l' \equiv \text{Var}\left(E\left[Y_l|W_l\right]\right)$ and $E_l^* \equiv E\left[\sqrt{C_{2l}(W_l)\text{Var}(Y_l|W_l)}\right]$. The results involve the

same terms as the one-sample results, and we show the extension of Result 1 to the two-sample problem, relegating the remaining three results to Supporting Materials Appendix A.

Result 1-two: With

$$\lambda_l^*(W_l) = \sqrt{\frac{\mathrm{Var}(Y_l|W_l)}{C_{2l}(W_l)}}\sqrt{\frac{C_{1l}}{\mathrm{Var}(E[Y_l|W_l])}}, \quad (13)$$

$$n_l^* = \frac{\sqrt{V_l'/C_{1l}}(B-C_0)}{E_1^*+E_2^*+\sqrt{C_{11}V_1'}+\sqrt{C_{12}V_2'}}, \quad (14)$$

the design specified by $(\lambda_1^*, \lambda_2^*, n_1^*, n_2^*)$ is optimal in that, among all possible designs indexed by $(\lambda_1, \lambda_2, n_1, n_2)$ that do not exceed the expected total cost B, it minimizes $V_1(\lambda_1)/n_1 + V_2(\lambda_2)/n_2$, provided that $\lambda_l^*(W_l) \leq 1$ for all W_l in the support of W_l. Otherwise, the modified optimality formulas described in the Supporting Materials Appendix A are needed.

Parallel to the one-sample problem, the cost-adjusted conditional variances are the key input parameters. In addition, note that the ratio of optimal sample sizes n_1^*/n_2^* equals $\sqrt{V_1'/C_{11}}/\sqrt{V_2'/C_{12}}$, such that the phase-one cost adjusted relative variance determines the optimal randomization allocation.

5. Case Study: Optimal Design of a Two-Phase HIV Vaccine Trial

We use HVTN 054 as pilot data for developing optimal sampling designs for the Phase I HVTN 083 protocol. HVTN 083 randomizes study volunteers to one of five candidate prime-boost HIV envelope protein-based vaccine regimens, with $n = 30$ per group. The objectives are to assess vaccine-induced T cell responses for each regimen and to compare the responses among the pairs of regimens. Here we consider the optimal sampling design for estimating the mean importance-weighted breadth, $E[Y]$, for a single vaccine group.

In general (including for HVTN 054 and 083), HVTN vaccine trials evaluate T cell responses using a standard test panel of 1280 15-mer HIV peptides as a globally representative set of peptides for measuring reactive T cell epitopes. The trials measure the number of reactive 15-mers for each trial participant and define a subject's response breadth (W) as the number of reactive 15-mers. A variety of importance-weighted breadth endpoints Y are of interest, defined as the sum of importance-weights across a subject's reactive 15-mers, which may be expensive to determine. In particular, determination of the weight attached to a reactive 15-mer may require "fine epitope mapping," wherein many separate experiments are conducted on 9-mer peptides (and perhaps on 8-, 10-, 11- or 12-mer peptides) within the 15-mer to determine the actual reactive T cell epitope. The high expense implies it may be cost-prohibitive to measure Y in every subject. Fortunately, however, breadth is a highly predictive auxiliary.

Figure 1 shows data on W and Y from HVTN 054, where Y uses as importance-weight for a reactive 15-mer the magnitude of the T cell response to the optimized epitope inside the 15-mer. The breadth W is a highly predictive auxiliary variable, with $R^2 = 0.79$ in a simple linear regression model (Figure 1). The extreme outlier was removed for these calculations because we do not want the chosen optimal design to be sensitive to this unusual observation. Adding a quadratic term did not improve the R^2; therefore we use only the single auxiliary W to optimally design the HVTN 083 study.

We determined the costs C_1 and $C_2(W)$ by reviewing laboratory costs for HVTN vaccine trials. However, because the real costs cannot be published, we multiplied each real cost by a hidden proportionality constant. This yields $C_1 = \$13,500$, which includes all lab costs needed to define W for each subject. To estimate $C_2(W)$, we decompose the breadth auxiliary W into two pieces: $\tilde{W} = W_{kn} + W_{un}$, where W_{kn} (W_{un}) is the number of the subject's reactive 15-mers for which the optimal 9-mer inside the 15-mer is known (unknown). If the optimal 9-mer inside a 15-mer is known, then only the single 9-mer needs to be tested; whereas if the optimal 9-mer is unknown, then all seven 9-mers tiling the 15-mer need to be tested. Given that it costs approximately \$1000 to measure the ELISpot response to a single 9-mer, it follows that a linear cost function has the form $C_2(W) = 1000W_{kn} + 7000W_{un}$.

For the design exercise, throughout this section we index $C_2(W)$ by a constant $q \in [0, 1]$, $C_2(W, q) = [1000q + 7000(1 - q)]\, W$, where q is the proportion of reactive 15-mers that have a known optimal 9-mer. While q could be viewed as a random variable and hence is an additional auxiliary variable, for simplicity we conduct the design exercise using only the breadth auxiliary W, by assuming that everyone has the same fixed constant q. A range of fixed q's is considered as a form of sensitivity analysis to assess how the results depend on the fraction q. Result 1 requires fixing the expected total variable cost budget $B - C_0 = n(C_1 + E\left[C_2(W, q)\lambda(W)\right])$, which we determine by setting $n = 30, C_1 = \$13,500, q = 0.5$, and $\lambda(W) = 0.5$, which yields $B - C_0 = \$645,000$. For Results 2 and 4, V was chosen to give 90% power (1-sided 0.025-level test) to detect a one standard deviation (value 2.6 estimated by median absolute deviation) lower mean value than the null hypothesized value. To specify the key input parameter for determining the optimal $\lambda(W)$, $r(W, q) = \sqrt{\mathrm{Var}(Y|W)/C_2(W, q)}$, we estimate $\mathrm{Var}(Y|W)$ in two steps. First, we compute the sample variances of Y for four subgroups of subjects defined by those with $W = 0$ and the tertiles of those with $W > 0$. Second, we explored various model-fits to the four points, and found a commonly used parametric form ([14], page 449) to fit well, $\mathrm{Var}(Y|W) = cW^{\nu}$, where $c = 0.55$ and $\nu = 1.45$ gave the best maximum likelihood fit. For the two terms in the results that require computing an expectation over the distribution of W, $E\left[\sqrt{C_2(W, q)\mathrm{Var}(Y|W)}\right]$ and $\mathrm{Var}(E[Y|W])$, we use the empirical distribution from the pilot study HVTN 054. Specifically, $E\left[\sqrt{C_2(W, q)\mathrm{Var}(Y|W)}\right]$ is specified as the sample average of the $\sqrt{C_2(W_i, q)\mathrm{Var}(Y|W_i)}$ values and $\mathrm{Var}(E[Y|W])$

as the sample variance of the the fitted values $\widehat{E}[Y|W_i]$ based on a simple linear regression model (which fits well, Figure 1), for $i = 1, \cdots, n$.

Figure 2 shows the optimal $(\lambda^*(W), n^*)$ values for Results 1 and 2, for 11 cost functions $C_2(W, q)$ indexed by $q = 0, 0.1, 0.2, \cdots, 1.0$, and for a constant cost function $C_2(W, \text{const}) \equiv I(W > 0)1260$, computed as $C_2(W, \text{const}) = E\left[C_2(W, q = 0.5)\right]$ using the empirical distribution of W from HVTN 054. For subjects with auxiliary $W = 0$, Y is known to be 0 at no cost, such that in all cases the optimal design selects all subjects with $W = 0$ (indicated by the dots in the upper-left corners). For $W > 0$, $r(W, q)$ modestly increases with W for the designs indexed by q, moreso as q increases. The legend of Figure 2(a) shows that for Result 1 the optimal design is 4–9% more efficient than the simple random sampling optimal design, with greater efficiency gains achieved when $r(W, q)$ varies more in W. The optimal design has sample size n^* between 24 and 33, where designs with more expensive phase two costs have smaller sample sizes. The reverse pattern is found for Result 2 (Figure 2(b)), which is explained by the different outcomes being minimized. In addition, compared to Result 1, there are smaller ranges of optimal sample sizes ($n^* = 21 - 27$) and of efficiency gains compared to the simple random sampling optimal design (2–4%).

Figure 3 shows the optimal $\lambda^*(W)$ for Result 3, using three fixed phase-two budgets $B' = \$120,000, \$240,000$, or $\$600,000$. These budgets are chosen to represent vaccines with low, moderate, and high T-cell breadth (averages of $\bar{W} = 2$, 4, and 10 reactive 15-mers), calculated as $B'(\bar{W}) = 30 * (1000 * 0.5 + 7000[1 - 0.5])\,\bar{W} * \lambda$ with $\lambda = 0.5$. The pattern of REs is similar to those for Result 1. This demonstrates the fact that, if B' is made large enough, the optimal design will use complete phase-two sampling. Figure 4 shows the optimal $\lambda^*(W)$ values for Result 4. For the cost functions indexed by q, the optimal $\lambda^*(W)$ is constant in q because of canceling in formula (9). The smaller the treatment effect size the study is powered to detect, the more subjects are sampled into phase two. The cost-constant designs have larger efficiency gain than the cost-dependent designs ($RE = 0.76$ compared to $RE = 0.91$).

This case study illustrates that the optimal sampling design is only expected to confer material efficiency improvements over the simple random sampling optimal design if the cost-standardized conditional variance function $r(W, q)$ varies substantially in W. The $r(W, q)$ specified from HVTN 054 is moderately variable in W, yielding between a 2–24% efficiency gain. With more justification provided next, the optimal stratified random sampling design may be a prudent choice for HVTN 083.

6. Finite-Sample Efficiency of the Two-Phase Methods

Based on the well-characterized asymptotic properties of the RR estimator, for large studies with large amounts of pilot data, the above results should provide reliable guidance on the relative performance of different estimation approaches, where the 'estimation approach' refers to the combination of the sampling design (procedure to estimate the λ_i) and analysis (implementation of the RR estimator). Otherwise, simulation studies are useful for comparing finite-sample performance. We apply Result 3 to the problem of sampling design and estimation of $\beta = E[Y]$ given fixed n and

expected phase-two expected budget B', comparing the following three approaches:

1. **Naive:** Ignore auxiliaries in design and analysis
2. **Analysis Only:** Use auxiliaries in analysis (RR method) but not design
3. **Design and Analysis:** Use auxiliaries in both design and analysis, with $\lambda^*(W)$ determined by formula (8), and $\text{Var}(Y|W)$ estimated from pilot data by:
   a. Maximum likelihood estimation in a correctly specified variance model
   b. Maximum likelihood estimation in an incorrectly specified variance model
   c. The sample variance for the relevant quartile of W (stratified sampling)
   d. The true function (unobtainable gold standard reference)

Approach 1 is the 'base case' that is standard practice, which uses simple random sampling of Y and estimates β with the sample mean $\bar{Y}$. Approach 2 uses optimal simple random sampling of Y followed with the efficient RR estimator incorporating W. Approach 3 uses the same RR estimator as Approach 2 but also uses an estimated optimal sampling design following Result 3, based on three approaches to specifying $\text{Var}(Y|W)$ based on a pilot data-set.

We consider a continuous univariate W, which allows investigating our conjecture that the results hold for continuous W, and for comparison to approaches with discrete W. The simulations address the impact of the following factors on the RE of the estimation approaches: (i) Predictive accuracy of the auxiliary ($R^2 = 0.2$, 0.5, or 0.8); (ii) Sample size of the pilot study (small or large, $m = 50$ or $m = 200$); (iii) Sample size of the study (moderate or large, $n = 200$ or $n = 1000$); (iv) use of true or estimated λ_i's; and (v) Correct or incorrect specification of the conditional variance model (3a-3d). Bias was also compared, but all estimation approaches were unbiased (as predicted by theory) and is not reported on further. The simulations were also performed with $m = 1000$; results were nearly identical to those with $m = 200$, indicating that $m = 200$ represents a large pilot sample size scenario.

The RR estimator (described at (12)) for Approaches 2 and 3 is obtained by specifying a linear regression model $E[Y|W] = g(W;\gamma)$, and fitting it from study subjects with $R = 1$ either by ordinary least squares (OLS) or by the robust linear regression MM-estimator (ROB) of Yohai [25], implemented with the *lmrob* function in the R package robustbase. To evaluate the effect of needing to estimate $E[Y|W]$ we also implemented the unobtainable RR estimator with the true regression function $E[Y|W]$. In each case the methods were applied with 'true' $\lambda(W_i)$'s, specified based on the pilot data and formula (8), and also with estimated $\lambda(W_i)$'s, calculated as $\exp(\text{logit}(\lambda_i) + \widehat{\alpha}W_i)/\{1 + \exp(\text{logit}(\lambda_i) + \widehat{\alpha}W_i)\}$, where $\widehat{\alpha}$ is the maximum likelihood estimator in a logistic regression model, as described in Section 3.6.

Pairs of pilot and study data-sets were simulated as follows. To generate a pilot data-set, for each of the m subjects, first W was simulated as $N(3.3, 0.5)$. Second Y was simulated as $N(\alpha_0 + \alpha_1 W, \text{Var}(Y|W))$, with (α_0, α_1) set to (0.1,3) and $\text{Var}(Y|W) = \exp(c_0 + \nu_0 W + \nu_1 W^2)$. The true parameters (c_0, ν_0, ν_1) were set to

achieve weak ($R^2 = 0.2$), moderate ($R^2 = 0.5$), or strong ($R^2 = 0.8$) prediction of Y, and were set to reflect the case where $r(W) = \sqrt{\text{Var}(Y|W)/C_2(W)}$ is independent of W or depends strongly on W (approximately quadratic). We use the constant $C_2(W) = 4000$ such that $C_2(\cdot)$ and $r(\cdot)$ do not need to be indexed by q. The resulting six settings are $(c_0, \nu_0, \nu_1) = (2.890, 0, 0), (1.504, 0, 0), (0.118, 0, 0)$ for $r(W)$ independent of W and (-1.026,-0.2,0.3), (-2.413,-0.2,0.3), (-3.799,-0.2,0.3) for $r(W)$ dependent on W. To gauge the potential for efficiency gains, for each setting we computed the asymptotic variance bound $V(\lambda)$ (formula (3)) analytically using (8). For the settings with $r(W)$ independent of W, the $V(\lambda)$'s are equal for Approaches 2 and 3, implying that Approach 3 should not confer improvements over Approach 2. In contrast, for the settings with $r(W)$ dependent on W, the ratio of variance bounds for Approach 3a versus Approach 2 are 0.52, 0.55, 0.65 for $R^2 = 0.2, 0.5, 0.8$; and for Approach 3a versus 3c are 0.81, 0.83, 0.88, implying that the fully optimal Approach 3a should confer improvements over both optimal simple random sampling and optimal stratified sampling.

To generate a two-phase study data-set, for each of the n subjects, first W was simulated the same as for the pilot study. Second, the λ_i's for each approach were specified. We use $C_2(W) = 4000$ and $B' = n * 4000 * 0.10$, so that on average 10% of subjects are selected for phase two. For Approaches 1 and 2 that use simple random sampling, λ_i is taken to be $B'/(4000n) = 0.10$. For Approach 3c, the m pilot subjects and n study subjects are divided into their respective quartiles of W. For each quartile $j = 1, 2, 3, 4$, λ_j is determined by (8) with $C_2(j) = 4000$, $\widehat{\text{Var}}(Y|j)$ the sample variance of Y in the group of pilot subjects with W in the j^{th} quartile of W values of study subjects, $E\left[\sqrt{C_2(W)\text{Var}\left(Y|W\right)}\right] = \frac{1}{n}\sum_{i=1}^{n}\sqrt{C_2(W_i)\text{Var}(Y|W_i)}$, and $\text{Var}\left(E[Y|W]\right)$ the sample variance of the $\widehat{E}[Y|W_i]$'s, for $i = 1, \cdots, n$. Then, for W_i in quartile j of study subjects, λ_i is taken as λ_j. For Approach 3a, λ_i is determined by (8), with $C_2(W) = 4000$, $\text{Var}(Y|W)$ the maximum likelihood estimate in the correctly specified model $\text{Var}(Y|W) = \exp(c_0 + \nu_0 W + \nu_1 W^2)$ fit to the pilot data with heteroscedastistic linear regression model (with the *remlscore* function in the R library statmod), $E\left[\sqrt{C_2(W)\text{Var}\left(Y|W\right)}\right] = \frac{1}{n}\sum_{i=1}^{n}\sqrt{C_2(W_i)\text{Var}\left(Y|W_i\right)}$, and $\text{Var}\left(E[Y|W]\right)$ the sample variance of the $\widehat{E}[Y|W_i]$'s. For Approaches 3b and 3d, λ_i is determined similarly, where for 3b $\text{Var}(Y|W)$ is determined by fitting the misspecified model $\text{Var}(Y|W) = \exp(c_0 + \nu_0 W)$ and for 3d the true curve $\text{Var}(Y|W)$ is used. Once λ_i was fixed, the selection indicator R_i was generated according to a Bernoulli experiment with success probability λ_i, and for each subject with $R_i = 1$, Y_i was generated the same as for the pilot data.

Based on 10,000 generated pairs of pilot plus study data-sets, the sample variances of the estimates $\widehat{\beta}$ were computed. The RE of two estimators is measured by the ratio of these Monte Carlo variances. Figures 5 and 6 show results for studies with $n = 1000$ and $n = 200$ subjects, respectively, in each case at the two levels of pilot study sample sizes ($m = 50$ or $m = 200$), using estimated λ_i's. The pilot study size has almost no effect on the RE results for the large study ($n = 1000$), and a

minor effect for the smaller study ($n = 200$), with the RE differences between the pilot scenarios having no apparent systematic pattern. For the large-sample setting (Figure 5), the RR estimator (2, 3a-3d) is much more efficient than the complete-case estimator (1) irrespective of the sampling design, and, when $r(W)$ is constant (top panel), the efficiency is essentially unaffected by the sampling design. In addition, when $r(W)$ varies with W (bottom panel), using a more efficient sampling design (3a-3d) improves efficiency compared to optimal simple random sampling (2); using the (unknowable) true curve $r(W)$ does not improve efficiency compared to using the estimated curve (3d vs 3a); misspecifying $r(W)$ incurs a minor efficiency cost (3b vs 3a); under OLS fitting the methods 3a-3d are substantially more efficient when the true $E[Y|W]$ is used instead of the estimated $E[Y|W]$, whereas robust fitting removes this difference; and under both OLS and ROB fitting the optimal approach (3a) is always equally or more efficient than the quartilized stratified sampling optimal approach (3c). In addition, the efficiency improvement of methods 3a-3d over method 2 is largest for $R^2 = 0.20$ and smallest for $R^2 = 0.80$; this demonstrates the concept that if an excellent predictive auxiliary is available, then it may be less important to use the optimal sampling design, as a large amount of efficiency is gained solely from including the auxiliary in the analysis (as discussed in Section 3.5).

For the small-sample setting (Figure 6), we observe the same comparative performance of the different methods when the true $E[Y|W]$ is used, but different comparative performance when the estimated $E[Y|W]$ is used. In particular, under OLS fitting and $R^2 = 0.2$ or 0.5, the optimal approach (3a) is much less efficient than simple random sampling (2), and, for the weakest R^2, is even less efficient than the complete-case estimator. Moreover, under OLS fitting the optimal stratified sampling approach (3c) is much more efficient than the optimal approach (3a). This finding is explained by the fact that on average only 20 subjects have Y measured, making imprecise the estimation of $E[Y|W]$ for some simulated data-sets, and the Monte Carlo sample variance is sensitive to outlying simulation results. These problems are largely repaired by robust fitting (we studied the MM-estimator of Yohai [25]), under which the optimal approach (3a) is always more efficient than simple random sampling (2), and is generally at least as efficient as the stratified sampling optimal approach (3c). Furthermore, for the smaller study the need to estimate $E[Y|W]$ generally incurs a larger efficiency-cost than for the larger study. We infer that for studies with a small phase-two sample it is important to use a robust method for estimating $E[Y|W]$, and the optimal stratified sampling approach may be expected to perform as well as the fully optimal approach; achieving a material efficiency gain via the fully optimal approach may require a relatively large study.

The simulations presented in Figures 5 and 6 were repeated using true λ_i's instead of estimated λ_i's, yielding similar results (not shown), with the estimated approach providing slightly greater efficiency gains for Approach 3a versus Approach 2.

7. Discussion

For a two-phase sampling study where the goal is to estimate a mean or mean

difference of an expensive endpoint Y (e.g., a clinical trial), we developed optimal approaches to auxiliary-covariate based sampling design followed by semiparametric efficient estimation. We showed that combining optimal sampling with optimal estimation can improve efficiency in practice compared to the simpler approach that uses simple random sampling and optimizes only the estimation. In particular, the optimal combined approach offers superior performance when there are auxiliaries W available that at least weakly predict Y, and there is adequate knowledge and pilot data to characterize the cost-standardized conditional variance $\mathrm{Var}(Y|W)/C_2(W)$ reasonably well and to support that it varies substantially in W. We also found that, for relatively small studies, it is important to use a robust estimator of $E[Y|W]$ within the RR estimator to realize the efficiency gain. Moreover, for small studies we found that the simplified optimal combined approach that uses stratified two-phase sampling tends to provide at least as good as efficiency as the fully optimal approach that specifies $\mathrm{Var}(Y|W)$ more finely. In addition, while we developed the optimal combined approach for a discrete auxiliary W with an arbitrary number of levels, which allows the results to be used in practice with a continuous W and with multiple continuous W (through contructing arbitrarily fine grids), for most applications we expect it to provide a significant advantage only if W has a limited number of levels or is univariate continuous, for in these case it is most feasible to accurately estimate $\mathrm{Var}(Y|W)$ with pilot data and to accurately estimate $E[Y|W]$ with study data. It is rare in practice for clinical trials to use a combined optimal design and analysis approach, and we hope this article will encourage increased consideration of this approach.

We developed four optimality results for each of the one-sample and two-sample problems, which will fit different contexts for two-phase clinical trials. For trials where it is practical to design the whole trial (both phases) up front, such as when the expensive endpoint of interest and its predictive auxiliary endpoint are well-understood from the outset, Results 1 and 2 apply. Result 2 is ideal for study design when feasible, as it powers the study to detect a scientifically relevant alternative hypothesis; however it will not apply if the sponsor has a fixed budget, in which case Result 1 is appropriate. Results 3 and 4 fit the common scenario in practice where expensive biomarker studies are designed after the core study. By waiting until phase one is complete before designing phase two, Results 3 and 4 can benefit from richer, up-to-date information (for example, HIV vaccine trial designs rapidly adapt to the latest technologies for measuring immune responses). For such settings, Result 4 is ideal for two-phase design because it allows powering the trial for a relevant alternative hypothesis, and Result 3 will be needed when the sponsor fixes the phase-two budget.

While we focused on two-phase studies, the RR estimation method accommodates general Q-phase studies, where at phase one auxiliaries are measured in everyone, at phase two additional auxiliaries are measured in a random sample, and so on for subjects sampled through all successive phases, and the outcome Y is measured in subjects who are sampled in phase Q. Such designs find application in T cell HIV vaccine trials because three or four phases of laboratory testing of HIV peptide pools

may be used to determine the primary endpoint. The optimality results for two-phase studies extend to general Q-phase studies, but the formulas become increasingly complex with increasing Q; for example Supporting Materials Appendix D provides the lengthy formulas for Result 1 in three-phase studies. Furthermore, as Q increases it becomes increasingly difficult to specify the growing number of input parameters stably enough for the optimal strategy to improve efficiency compared to a simpler strategy; thus the methods described here may have their greatest practical application for two-phase studies.

Web-based Supporting Materials Materials

Title: Appendices Appendix A provides proofs of all of the optimality results for discrete W, including a statement of Results 1-two, 2-two, 3-two, and 4-two, and provides an extension of the results for continuous W. Appendix B provides the result and proof of when a one-phase versus two-phase design is more efficient (based on Results 1 and 2). Appendix C summarizes an approach to selecting the phase two sample under without replacement sampling, and Appendix D provides an optimality result for the three-phase sampling design. (pdf file)

Title: opt2phdesignanalysis R code for implementing optimal design Results 1–4 and for implementing the analysis with the RR method. (text file)

Acknowledgements The authors thank Nicole Frahm for her subject-matter collaboration as the Associate Director of the HIV Vaccine Trials Network (HVTN) laboratory program. This research was supported by an HVTN cooperative agreement with the National Institutes of Health Division of AIDS, grant 5 U01 AI068635-05, and by National Institutes of Health NIAID grants R01AI029168-20, 2 R37 AI054165-11, and 2 R37 AI032475-16A.

References

1. Liu Q, Li G, Anderson KM, Lim P. On efficient two-stage adaptive designs for clinical trials with sample size adjustment. *Journal of Biopharmaceutical Statistics* 2012; **22**:617–640. DOI 10.108010543406.2012.678226

2. Borgan L, Langholz B, Samuelson S, Pogoda J. Exposure stratified case-cohort designs. *Lifetime Data Analysis* 2000; **6**:39–58. DOI 10.1023a:1009661900674

3. Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Using the whole cohort in the analysis of case-cohort data. *American Journal of Epidemiology* 2009; **169**:1398–1405. DOI 10.1093ajekwp055

4. 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. *Statistical Biosciences* 2009; **1**:32–49. DOI 10.1007s12561-009-9001-6

5. Kulich M, Lin DY. Improving efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association* 2004; **99**:832–844. DOI 10.11980162145040000000584

6. Rotnitzky A, Robins JM. Semiparametric regression estimation in the presence of dependent censoring. *Biometrika* 1995; **82**:805–820. DOI 10.1093biomet82.4.805

7. Zhang M, Tsiatis AA, Davidian M. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* 2008; **64**:707–715. DOI 10.1111j.1541-0420.2007.00976.x

8. McNamee R. Optimal design and efficiency of two-phase case-control studies with error-prone and error-free exposure measures. *Biostatistics* 2005; **6**:590–603. DOI 10.1093biostatisticskxi029

9. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 1986; **73**:1-11. DOI 10.23072336266

10. Fauci AS, Johnston MI, Dieffenbach CW, Burton DR, Hammer SM, Hoxie JA, Martin M, Overbaugh J, Watkins DI, Mahmoud A, Greene W. HIV vaccine research: The way forward. *Science* 2008; **321**:530–532. DOI 10.1126science.1161000

11. Li F, Malhotra U, Gilbert PB, Hawkins N, Duerr A, McElrath MJ, Corey L, Self SG. Peptide selection for Human Immunodeficiency Virus Type 1 CTL-based vaccine evaluation. *Vaccine* 2006; **24**:6893–6904. DOI 10.1016j.vaccine.2006.06.009

12. Russell ND, Hudgens MG, Ha R, Havenar-Daughton C, McElrath MJ. Moving to Human Immunodeficiency Virus Type 1 vaccine efficacy trials: Defining T cell responses as potential correlates of immunity. *Journal of Infectious Diseases* 2003; **187**:226–242. DOI 10.1086367702

13. Moodie Z, Huang Y, Gu L, Hural J, Self SG. Statistical positivity criteria for the analysis of ELISpot assay data in HIV-1 vaccine trials. *Journal of Immunological Methods* 2006; **315**:121–132. DOI.10.1016j.jim.2006.07.015

14. Särndal C-E, Swensson B, Wretman J. *Model Assisted Survey Sampling.* Springer-Verlag: New York, 1992.

15. Royall RM. Robustness and optimal design under prediction models for finite populations. *Survey Methodology* 1992; **18**:179–185. DOI 10.108001621459.1973.10481440

16. Kozak M, Zieliński A, Singh S. Stratified two-stage sampling in domains: Sample allocation between domains, strata, and sampling stages. *Statistics and Probability Letters* 2008; **78**:970–974. DOI 10.1016j.spl.2007.09.057

17. Nedyalkova D, Tillé Y. Optimal sampling and estimation strategies under the linear model. *Biometrika* 2008; **95**:521–537. DOI 10.1093biometasn027

18. Särndal C-E. On π inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika* 1980; **67**:639–650. DOI 10.23072335134

19. Isaki CT, Fuller WA. Survey design under the regression superpopulation model. *Journal of the American Statistical Association* 2009; **77**:89–96. DOI 10.23072287773

20. Wright RL. Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association* 1983; **78**:879–884. DOI 10.23072288199

21. Cochran WG. *Sampling Techniques, 3rd Edition.* Wiley: New York, 1977.

22. Robins JM, Sued M, Lei-Gomez Q, Rotnitzky A. Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science* 2007; **22**:544–559. DOI 10.230727645860

23. Gilbert PB, Sato A, Sun X, Mehrotra DV. Efficient and robust method for comparing the immunogenicity of candidate vaccines in randomized clinical trials. *Vaccine* 2009; **27**:396–401. DOI 10.1016j.vaccine.2008.10.083

24. Tsiatis AA, Davidian M, Zhang M, Lu X. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine* 2008; **27**:4658–4677. DOI 10.1002sim.3113

25. Yohai VJ. High breakdown-point and high efficiency robust estimates for regression. *Journal of the American Statistical Association* 1987; **15**:642–656. DOI 10.1214aos1176350366

Figure Legends

Figure 1. For the HVTN 054 HIV vaccine trial: (a) Correlation of the auxiliary W (15-mer level breadth) with the study endpoint Y (importance-weighted breath). Panel (b) shows a boxplot of Y within categories defined by ranges of W.

Figure 2. (a) Optimal $\lambda^*(W)$ and n^* under Result 1; (b) Optimal $\lambda^*(W)$ and n^* under Result 2. The bottom line is for the largest phase-two cost ($q = 0$), and as the lines step up with increasing q the phase-two cost decreases.

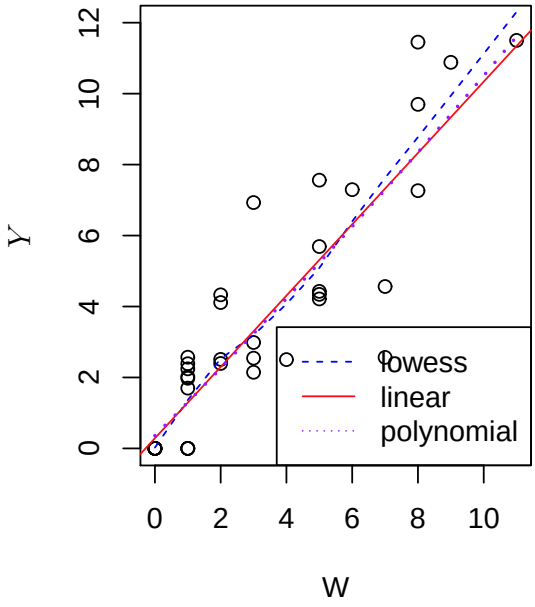
Figure 3. Optimal $\lambda^*(W)$ under Result 3 for (a) small, (b) moderate, and (c) large expected phase-two budgets.

Figure 4. Optimal $\lambda^*(W)$ under Result 4.

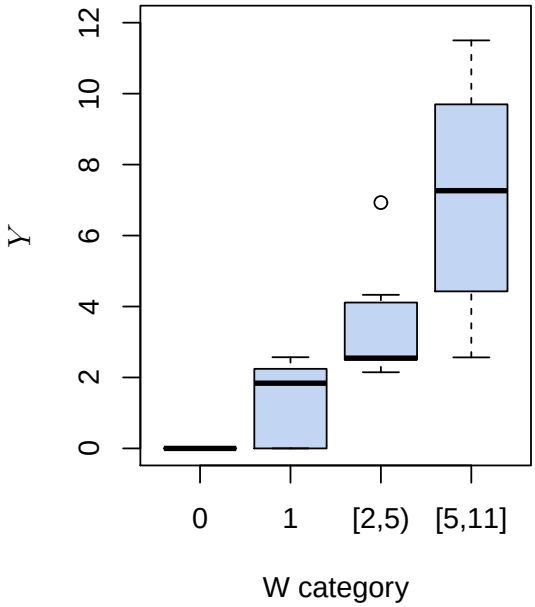
Figure 5. For studies with $n = 1000$ subjects, the figure shows Monte Carlo sample variances of β calculated from 10,000 generated pairs of pilot and study datasets for the design-analysis approaches (1) Naive, (2) Analysis Only, (3) Optimal design and analysis: (a) Correctly Specified, (b) Incorrectly Specified, (c) Quartilized Stratified, (d) Unobtainable Gold Standard. For the top panel the true $r(W) = \sqrt{\text{Var}(Y|W)/C_2(W)}$ is constant and for the bottom panel it is approximately quadratic. Solid (dotted) lines indicate the RR estimator with true (estimated) $E[Y|W]$. OLS indicates $E[Y|W]$ was estimated by ordinary least squares and ROB indicates $E[Y|W]$ was estimated by the robust estimator of Yohai [25]. Symbols 2, 5, and 8 indicate $R^2 = 0.2, 0.5, 0.8$. Estimated λ_i's were used.

Figure 6. For studies with $n = 200$ subjects, the figure shows the same information as in Figure 5.

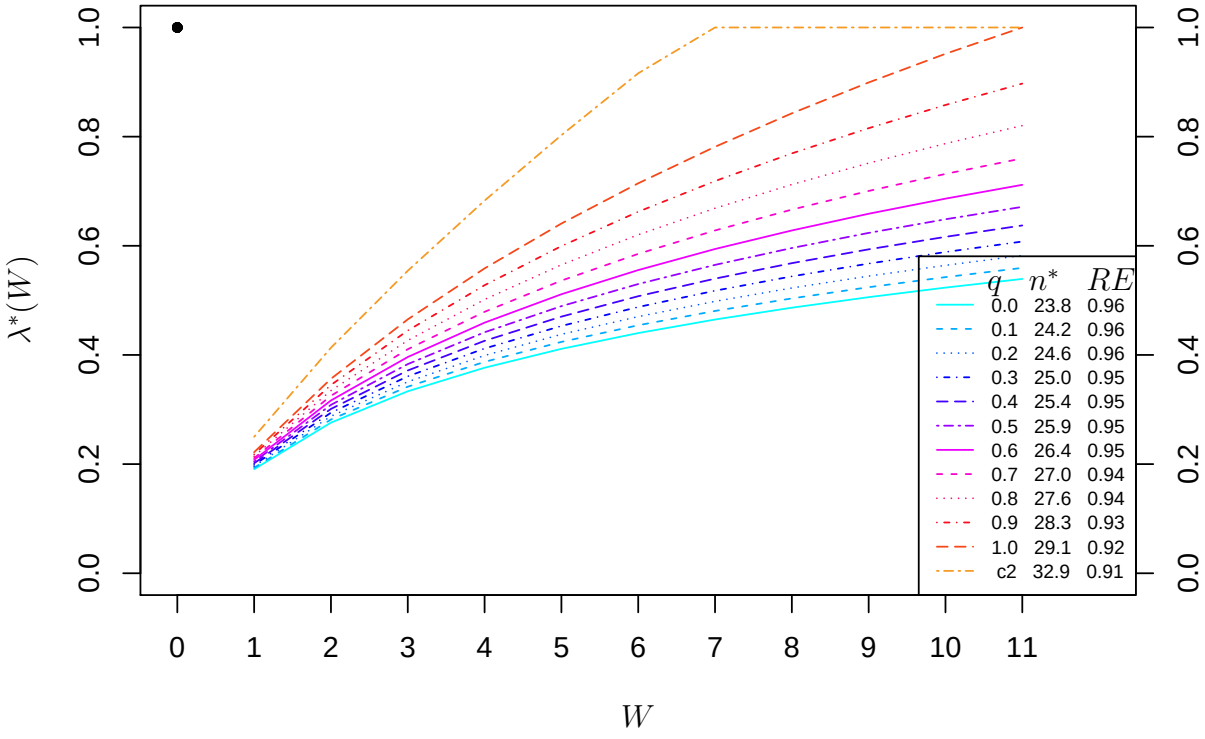
(a) Linear model $R^2 = 0.79$



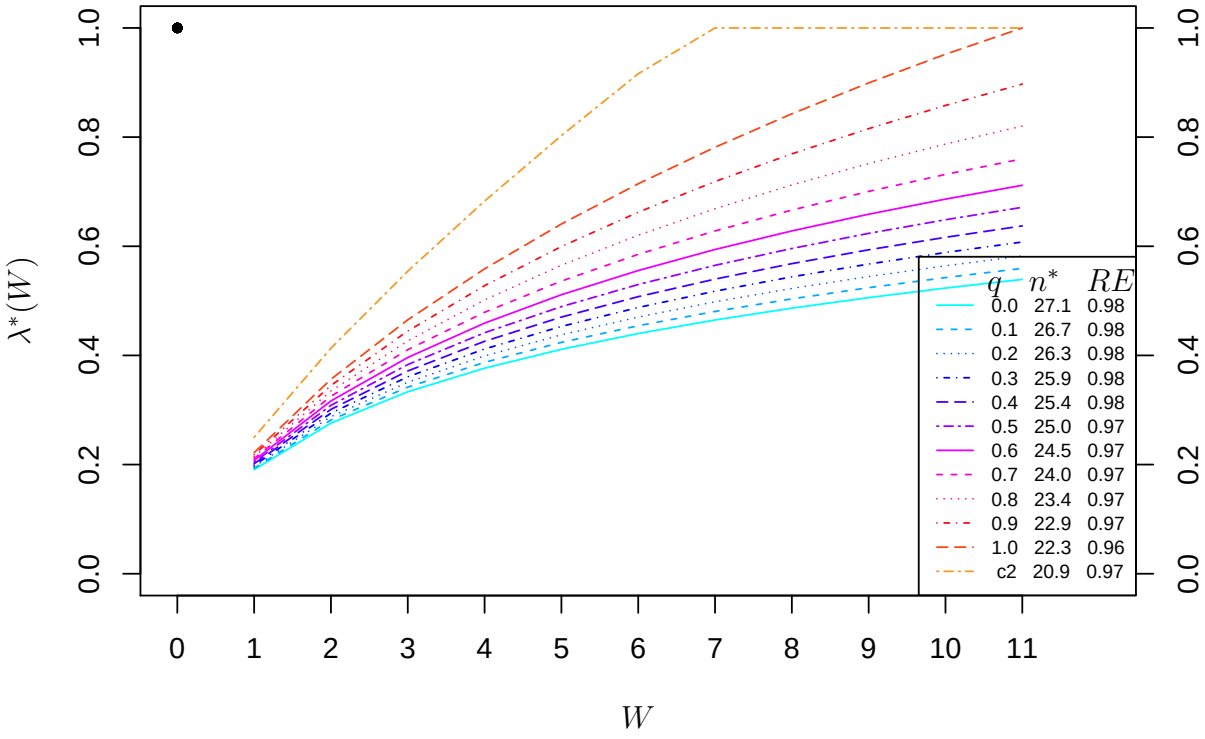
(b) Y (importance-weighted breadth)



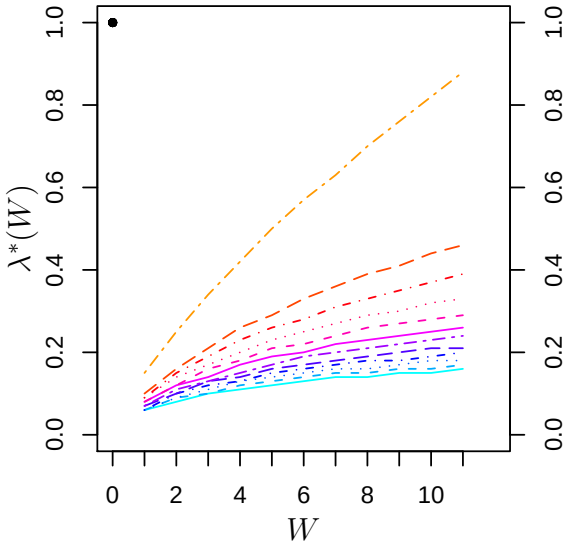
(a) Result 1



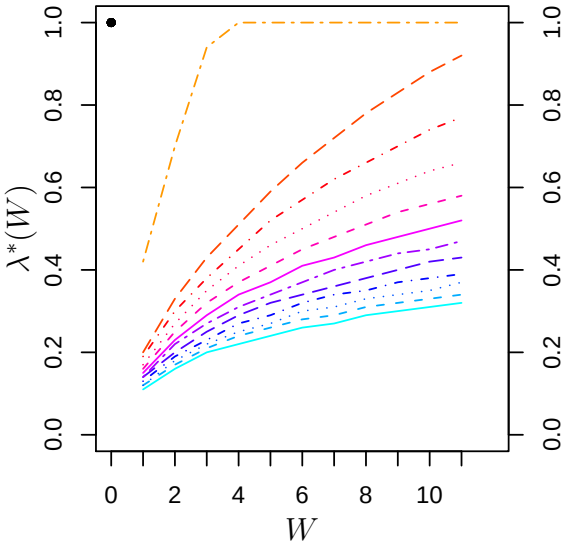
(b) Result 2: 90% Power to Detect Mean $\delta = 2.6$ ($\alpha = 0.05$)



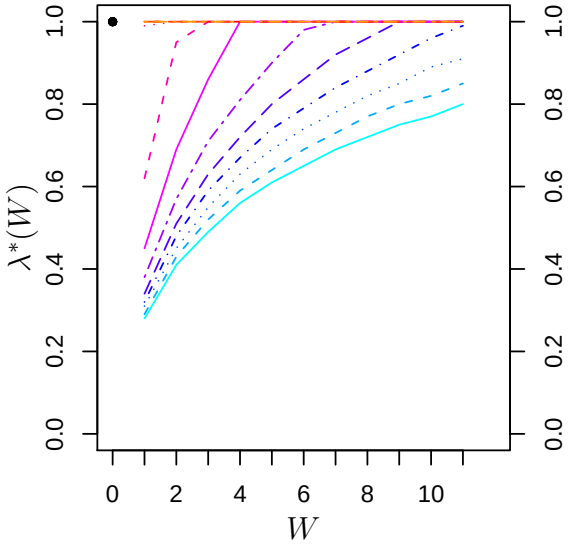
(a) Result 3: $X = 2$, $B' = \$120,000$



(b) Result 3: $X = 4$, $B' = \$240,000$

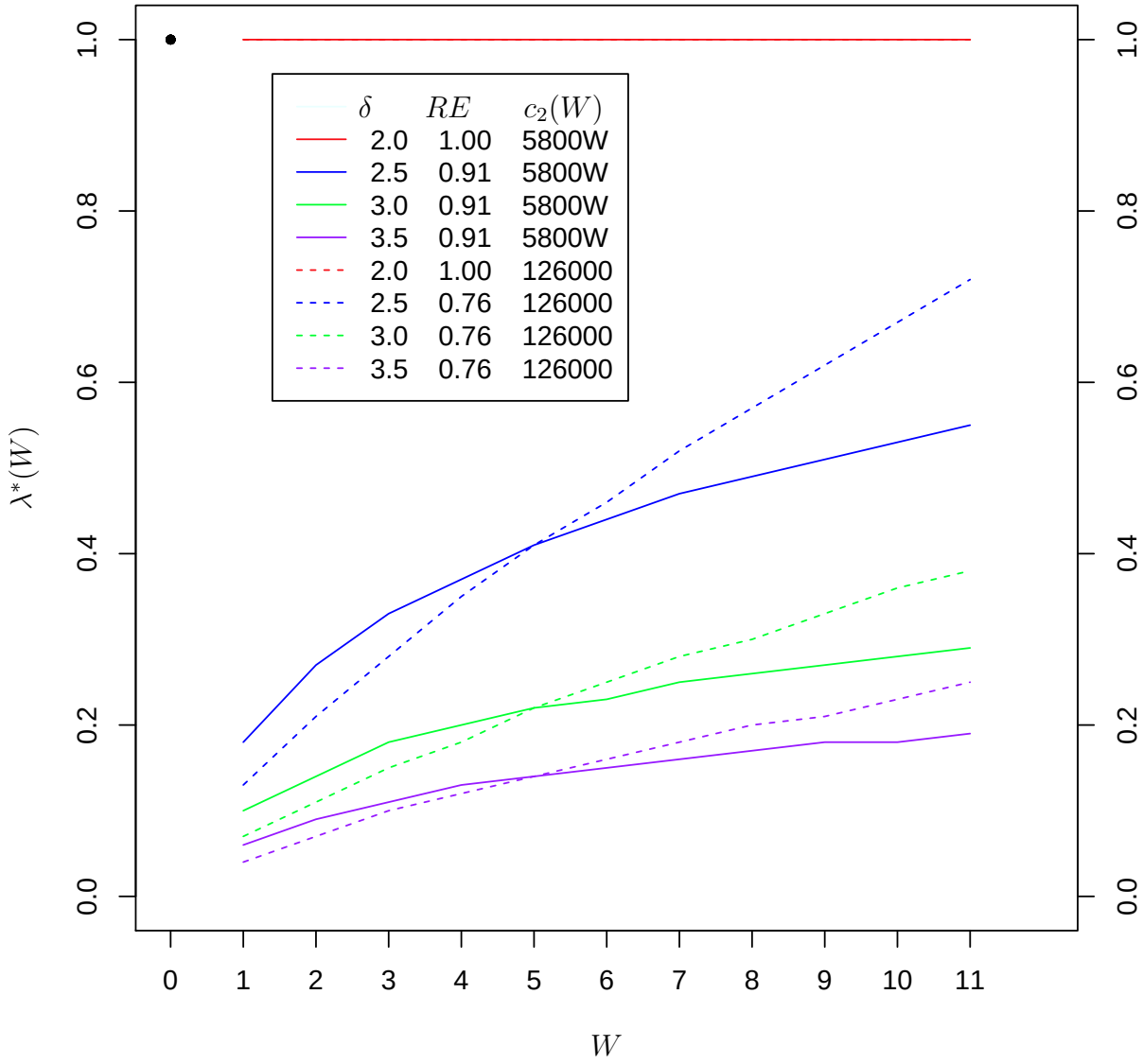


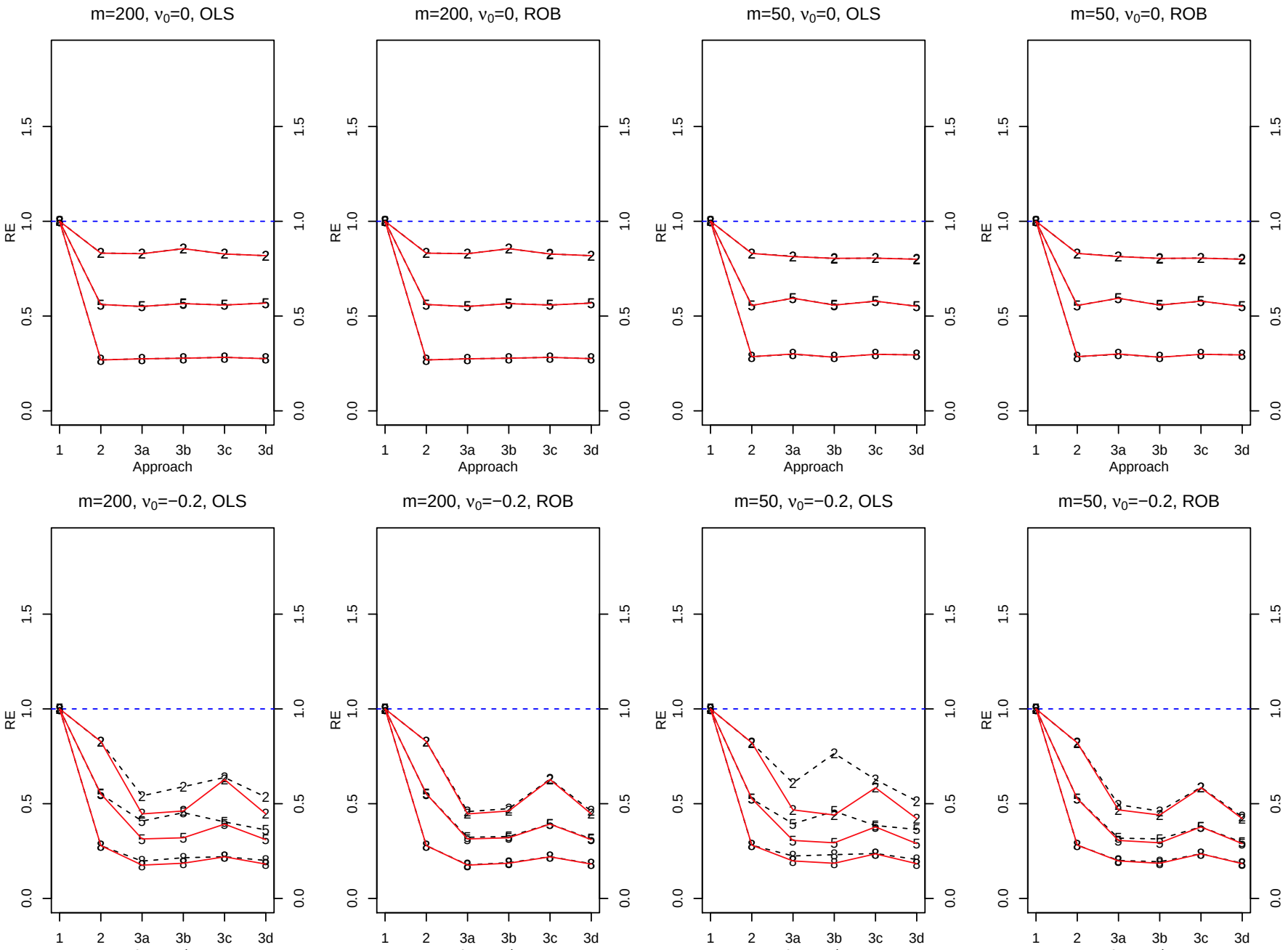
(c) Result 3: $X = 10$, $B' = \$600,000$

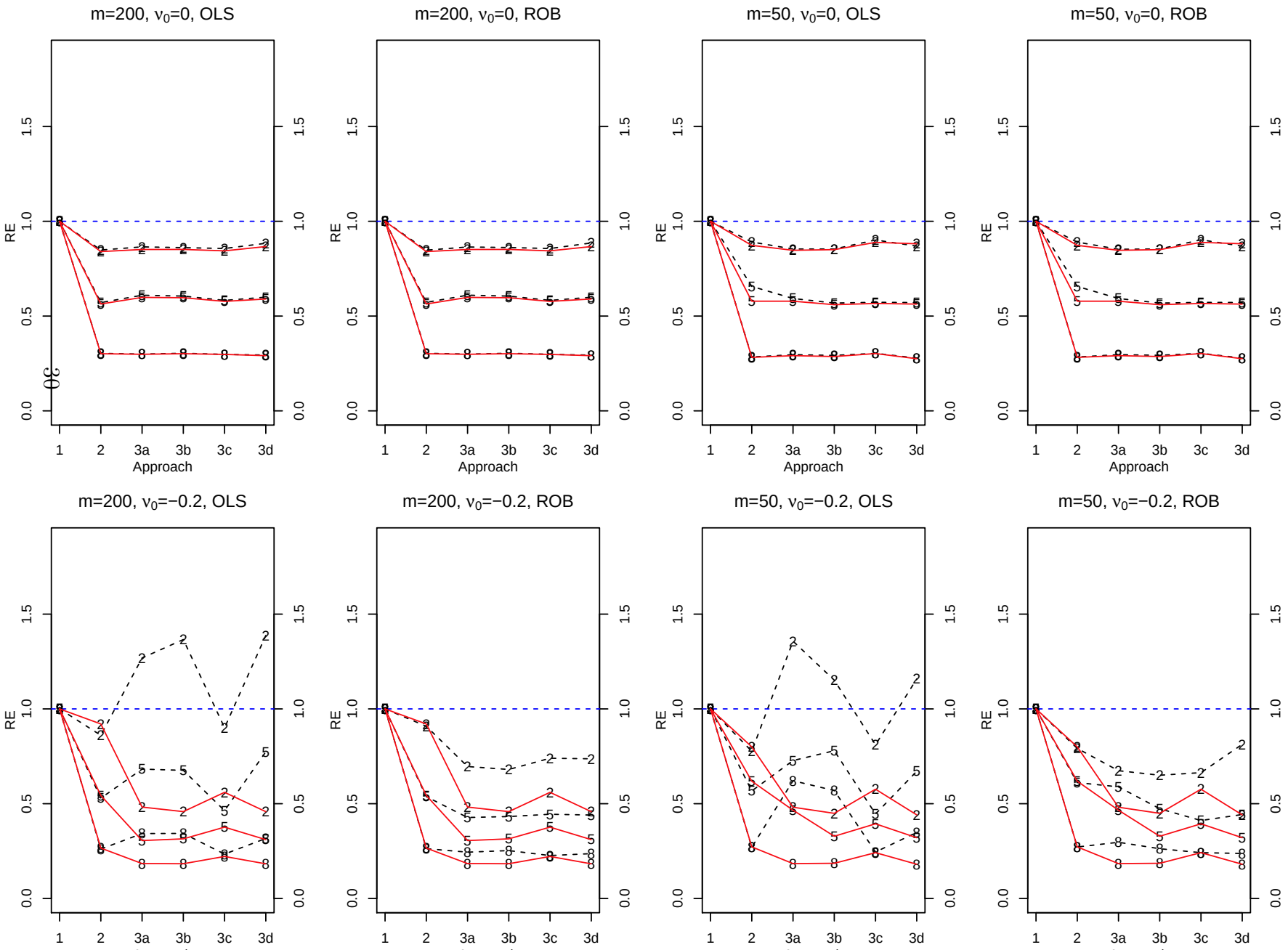


q	RE(a)	RE(b)	RE(c)
0.0	0.94	0.95	0.97
0.1	0.94	0.95	0.97
0.2	0.93	0.95	0.97
0.3	0.93	0.94	0.96
0.4	0.93	0.94	0.96
0.5	0.92	0.94	0.96
0.6	0.92	0.94	0.97
0.7	0.91	0.93	0.98
0.8	0.90	0.93	1.00
0.9	0.90	0.92	1.00
1.0	0.89	0.92	1.00
c2	0.88	0.94	1.00

Result 4: 90% Power to Detect Mean Difference δ ($\alpha = 0.05$, $n = 30$)




m=200, ν0=0, OLS
m=200, ν0=0, ROB
m=50, ν0=0, OLS
m=50, ν0=0, ROB
m=200, ν0=−0.2, OLS
m=200, ν0=−0.2, ROB
m=50, ν0=−0.2, OLS
m=50, ν0=−0.2, ROB
RE
Approach
1
2
3a
3b
3c
3d
0.0
0.5
1.0
1.5


m=200, ν0=0, OLS
m=200, ν0=0, ROB
m=50, ν0=0, OLS
m=50, ν0=0, ROB
m=200, ν0=−0.2, OLS
m=200, ν0=−0.2, ROB
m=50, ν0=−0.2, OLS
m=50, ν0=−0.2, ROB
RE
Approach
1
2
3a
3b
3c
3d
0.0
0.5
1.0
1.5