HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2023 March 01.

Published in final edited form as:

Biometrics. 2022 March ; 78(1): 388–398. doi:10.1111/biom.13405.

Power and Sample Size for Observational Studies of Point Exposure Effects

Bonnie E. Shook-Sa[*], **Michael G. Hudgens**[**]

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, U.S.A.

Summary:

Inverse probability of treatment weights are commonly used to control for confounding when estimating causal effects of point exposures from observational data. When planning a study that will be analyzed with inverse probability of treatment weights, determining the required sample size for a given level of statistical power is challenging because of the effect of weighting on the variance of the estimated causal means. This paper considers the utility of the *design effect* to quantify the effect of weighting on the precision of causal estimates. The design effect is defined as the ratio of the variance of the causal mean estimator divided by the variance of a naïve estimator if, counter to fact, no confounding had been present and weights were not needed. A simple, closed-form approximation of the design effect is derived that is outcome invariant and can be estimated during the study design phase. Once the design effect is approximated for each treatment group, sample size calculations are conducted as for a randomized trial, but with variances inflated by the design effects to account for weighting. Simulations demonstrate the accuracy of the design effect approximation, and practical considerations are discussed.

Keywords

Causal inference; Design effect; Effective sample size; Hájek estimator; Inverse probability weighting; Marginal structural modeling

1. Introduction

Researchers often aim to estimate causal effects rather than just associations between variables. In settings where experimental designs are implausible, inference relies on observational data from which measured associations can be confounded. Marginal structural models (MSMs) are a commonly used method to estimate causal effects in the presence of confounding variables (Hernán et al., 2000; Robins et al., 2000; Cole and Hernán, 2008; Brumback et al., 2004). These models can be fit using numerous methods, but are commonly fit via weighted estimating equations, where the weights are the inverse of each participant's probability of the observed treatment (or exposure). For a binary

[*] bshooksa@live.unc.edu . [**] mhudgens@email.unc.edu .

Supporting Information

Web Appendices, Tables, and Figures are available with this paper at the Biometrics website on Wiley Online Library. The R code developed for the simulation studies in Sections 4 and 6 is also available on the Biometrics website.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

point exposure, the estimand of interest is often the average causal effect, the difference in counterfactual means for the two treatment levels. With the assumptions of causal consistency, conditional exchangeability, and positivity, the inverse probability of treatment weight (IPTW) estimators are consistent for the MSM parameters for the causal means and the average causal effect (Lunceford and Davidian, 2004). Variance estimates are computed using standard estimating equation theory (Stefanski and Boos, 2002), with the empirical sandwich variance estimator providing a consistent estimator for the asymptotic variance of the estimated average causal effect.

While IPTW estimators provide researchers with an analytic tool for estimating causal effects in the presence of confounding variables, these estimators pose challenges during study design. The use of weights in the analysis affects the variance of the average causal effect estimator, making it challenging to determine the number of participants needed to achieve sufficient statistical power to detect a difference in causal means. Sample sizes cannot be calculated using standard methods that ignore weighting as in a randomized controlled trial (RCT) (e.g. as in Chow et al., 2017), as this will tend to be anti-conservative. Numerous papers have examined the properties of IPTWs and have developed guidelines and diagnostics for specifying weight models and adjusting estimated weights (Austin, 2009; Austin and Stuart, 2015; Cole and Hernán, 2008; Lee et al., 2011). However, currently no methods exist for power and sample size calculations for studies that will be analyzed using MSMs fit with IPTWs.

Weighted estimators are common in survey sampling, and methods have been developed to quantify the effect of weighting on the precision of estimates. Kish (1965, page 257) introduced the *design effect* under the randomization-based inferential paradigm for survey sampling. The design effect is the ratio of the variance of an estimator under a complex sample design to the variance of the estimator under a simple random sample. When participants are selected directly from the finite population rather than from clusters of correlated observations, the design effect for a population mean estimator simplifies to the design effect due to weighting ($deff_w$), or the unequal weighting effect (Kish, 1992). Let n be the sample size and w_i represent the sampling weight for the i^{th} participant, i.e., the inverse of participant i's probability of selection. The design effect due to weighting is defined using either of the two equivalent forms:

$$deff_w = \frac{n\sum_{i=1}^{n} w_i^2}{\left(\sum_{i=1}^{n} w_i\right)^2} = 1 + \frac{S^2(w)}{\left(n^{-1}\sum_{i=1}^{n} w_i\right)^2} \qquad (1)$$

where $S^2(w)$ is the finite sample variance of the weights. The design effect is interpreted as an estimator's increase in variance due to differential weights across participants. This metric is commonly applied to all types of complex sample designs in which individuals in the finite population have different probabilities of selection (Valliant et al., 2013, page 375). Gabler et al. (1999) provided a justification for how Kish's design effect also applies to model-based estimators. In practice, the design effect is used to calculate the *effective sample size*, which is equal to the observed sample size divided by the design effect. The effective sample size can be interpreted as the sample size under simple random sampling

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

that that would have produced the same variance as the sample selected under the complex design (Valliant et al., 2013, page 5).

Advantages of $deff_w$ are that it is outcome invariant and allows the sample size under a complex design to be translated into a sample size under a simpler design with the same variance. The former implies that $deff_w$ depends only on the participants' weights and is constant across outcomes. The latter means that once $deff_w$ is known or approximated, it can be used in power and sample size calculations along with the simpler assumptions needed to design a study without weights.

In this paper we consider design effects for planning observational studies to assess the effect of a treatment or exposure on an outcome of interest. In the analysis of observational data, McCaffrey et al. (2004, 2013) have used the effective sample size to quantify the loss of statistical precision following inference about causal effects using propensity score weighting. Here we describe the use of design effects for determining the sample size or power when designing an observational study to estimate point exposure effects. Section 2 introduces the design effect for causal inference and gives a large sample justification that the design effect can be approximated by Kish's $deff_w$. Section 3 demonstrates how the design effect can be used to determine the sample size or power of an observational study that will be analyzed using IPTWs. Section 4 examines the accuracy of the design effect approximation for various exposure and outcome types via simulations, and Section 5 provides practical considerations regarding the use of design effects. Section 6 considers estimating the design effect in the setting where pilot data are available, and Section 7 concludes with a discussion of the results and implications. The supporting information includes proofs of the propositions appearing in the main text along with additional simulations and supplemental tables and figures.

2. The Design Effect

2.1 Preliminaries

Suppose an observational study is being planned where n independent and identically distributed copies of (A_i, L_i, Y_i) will be observed, where A_i is the binary treatment (exposure) status for participant i such that $A_i = 1$ if participant i received treatment and $A_i = 0$ otherwise, L_i is a vector of baseline covariates measured prior to A_i or unaffected by treatment A_i, and Y_i is the observed outcome for participant i. The outcome Y_i may be continuous or discrete.

The aim of the observational study will be to estimate the effect of treatment A on outcome Y. Specifically, let Y_{1i} denote the potential outcome if an individual i, possibly counter to fact, receives treatment. Similarly let Y_{0i} denote the potential outcome if individual i does not receive treatment, such that $Y_i = A_i Y_{1i} + (1 - A_i) Y_{0i}$. Inference from the observational study will focus on parameters of the MSM $E(Y_a) = \beta_0 + \beta_1 a$, with particular interest in the parameter β_1 which equals the average causal effect $ACE = E(Y_1) - E(Y_0) = \mu_1 - \mu_0$. Note the MSM is saturated and thus does not impose any restrictions on the assumed structure of the data.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Under certain assumptions, the parameters of the MSM can be consistently estimated using IPTW. In particular, assume conditional exchangeability holds, i.e., $Y_a \perp A \mid L$ for $a \in \{0, 1\}$. Also assume that positivity holds such that $Pr(A = a \mid L = l) > 0$ for all l such that $dF_L(l) > 0$ and $a \in \{0, 1\}$, where F_L is the cumulative distribution function of L. Estimating the average causal effect under the stated assumptions with the IPTW estimator first entails estimating the propensity score for each participant, defined as $p_i = Pr(A_i = 1 \mid L_i)$ (Rosenbaum and Rubin, 1983). A model is fit to obtain $\hat{p}_i$, each participant's estimated probability of treatment conditional on observed covariates L_i. The estimated IPTW is then equal to $\widehat{W}_i = I(A_i = 1)\hat{p}_i^{-1} + I(A_i = 0)(1 - \hat{p}_i)^{-1}$, where $I(A_i = a)$ is a $\{0, 1\}$ treatment indicator for participant i. The estimated average causal effect $\hat{\beta}_1$ is obtained by regressing the observed outcome Y on treatment A with weights $\widehat{W}$ using weighted least squares. The resulting IPTW estimator is a difference in Hájek estimators for the two causal means (Hernán and Robins, 2020; Lunceford and Davidian, 2004):

$$\widehat{ACE} = \hat{\mu}_1 - \hat{\mu}_0 = \frac{\sum_{i=1}^{n} \widehat{W}_i Y_i I(A_i = 1)}{\sum_{i=1}^{n} \widehat{W}_i I(A_i = 1)} - \frac{\sum_{i=1}^{n} \widehat{W}_i Y_i I(A_i = 0)}{\sum_{i=1}^{n} \widehat{W}_i I(A_i = 0)} \tag{2}$$

Augmented IPW estimators, which incorporate both outcome and treatment models, may be used instead of (2) to estimate the ACE. Such estimators are doubly robust and will be more efficient than (2) if both the treatment and outcome models are correctly specified (Robins et al., 1994; Lunceford and Davidian, 2004). Thus, the power and sample size calculations derived below, which are based on (2), will be conservative for studies analyzed with augmented IPW estimators when the outcome model is correctly specified.

2.2 The Design Effect for a Single Causal Mean

Define the design effect to equal the ratio of the (finite sample) variance of $\hat{\mu}_a$ divided by the variance of a naïve causal mean estimator if, counter to fact, no confounding was present and weighting was not needed. That is,

$$deff_w^a = \frac{Var(\hat{\mu}_a)}{Var(\tilde{\mu}_a)}$$

where $\tilde{\mu}_a = \left\{\sum_{i=1}^{n} Y_i I(A_i = a)\right\} / \left\{\sum_{i=1}^{n} I(A_i = a)\right\}$. The derivation of the design effect estimator relies on the following proposition. The proposition assumes that the weights are known and are denoted by $W_a = P(A = a \mid L)^{-1}$ for $a \in \{0, 1\}$ with $W = AW_1 + (1-A)W_0$. Let $\sigma_a^2 = Var(Y_a)$ for $a \in \{0, 1\}$.

Proposition 1:

$$\sqrt{n}(\hat{\mu}_a - \mu_a) \xrightarrow{d} N(0, \Sigma_a)$$

where

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\Sigma_a = \sigma_a^2\left(\frac{E\left\{W^2 I(A = a)\right\}}{[E\{W I(A = a)\}]^2}\right) + R(L, Y_a)$$

and

$$R(L, Y_a) = E\left[\{W_a - E(W_a)\}(Y_a - \mu_a)^2\right]$$

with

$$|R(L, Y_a)| \leqslant \sqrt{Var(W_a) Var\left\{Y_a^2 - 2\mu_a Y_a\right\}}$$

for $a \in \{0, 1\}$

Proposition 1 implies that for large n the variance of $\hat{\mu}_a$ can be approximated as:

$$Var(\hat{\mu}_a) \approx \frac{\sigma_a^2}{n}\left(\frac{E\left\{W^2 I(A = a)\right\}}{[E\{W I(A = a)\}]^2}\right) + n^{-1} R(L, Y_a)$$

Because, for large n, $Var(\tilde{\mu}_a) \approx \sigma_a^2 / \left\{nP(A = a)\right\}$, it follows that

$$deff_w^a \approx \frac{P(A = a) E\left\{W^2 I(A = a)\right\}}{[E\{W I(A = a)\}]^2} + Er_a \tag{3}$$

where $Er_a = \left\{P(A = a)/\sigma_a^2\right\} R(L, Y_a)$, which by Proposition 1 is bounded by:

$$|Er_a| \leqslant \left\{P(A = a)/\sigma_a^2\right\}\sqrt{Var(W_a) Var\left(Y_a^2 - 2\mu_a Y_a\right)}$$

An approximation of (3) that does not depend on the potential outcome Y_a omits the remainder term Er_a:

$$\widetilde{deff}_w^a = \frac{P(A = a) E\left\{W^2 I(A = a)\right\}}{[E\{W I(A = a)\}]^2} \tag{4}$$

When planning an observational study, prior or pilot study data may be available to estimate (4). In particular, suppose based on a pilot study n_p copies of (L_i, A_i) are observed. Then replacing $P(A = a)$ with N_a / n_p where $N_a = \sum_{i = 1}^{n_p} I(A_i = a)$, $E\left\{W^2 I(A = a)\right\}$ with $n_p^{-1}\sum_{i = 1}^{n_p} \widehat{W}_i^2 I(A_i = a)$, and $E\{WI(A = a)\}$ with $n_p^{-1}\sum_{i = 1}^{n_p} \widehat{W}_i I(A_i = a)$, a consistent estimator of (4) is:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\widehat{deff}_w^a = \frac{N_a \sum_{i=1}^{n_p} \widehat{W}_i^2 I(A_i = a)}{\left\{\sum_{i=1}^{n_p} \widehat{W}_i I(A_i = a)\right\}^2} \quad (5)$$

This estimator has the same form as Kish's design effect (1), applied to treatment group $A = a$. When prior data are not available, the design effect can be approximated using (4) based on an assumed distribution for $A \mid L$ and the marginal distribution of L. The bias of (4) or (5) as an approximation to (3) in a given application depends on the value of Er_a. Bias of (4) and (5) for varying outcome types and confounding structures is evaluated empirically in simulation studies presented in Section 4. A modified design effect estimator that accounts for the remainder is considered in Section 6.

3. Sample Size Calculations using the Design Effects

When ACE is the focus of inference for the observational study being planned, the large sample distribution of $\widehat{ACE}$ can be used for power or sample size calculations. As $n \to \infty$, $\widehat{ACE}$ is consistent and asymptotically normal. In particular, from (A.1) in Web Appendix A and the delta method it follows that

$$\sqrt{n}(\widehat{ACE} - ACE) \xrightarrow{d} N(0, \Sigma^*), \quad (6)$$

where $\Sigma^* = \Sigma_1 + \Sigma_0$. Treating the weights as fixed or known leads to a larger asymptotic variance for $\widehat{ACE}$ compared to appropriately treating the weights as estimated, i.e., Σ^* is at least as large as the true asymptotic variance of $\widehat{ACE}$ (Lunceford and Davidian, 2004). Therefore, sample size formulae derived based on Σ^* would in general be expected to be conservative.

The results in Proposition 1 and (6) allow for sample size calculations for studies that will be analyzed using MSM with IPTW. Suppose the sample size for the observational study being planned is to be determined on the basis of the power to test $H_0 : ACE = 0$ versus $H_a : ACE \neq 0$ or equivalently $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$. Define the test statistic $t = \widehat{ACE}\left\{Var_n(\widehat{ACE})\right\}^{-1/2}$, where $Var_n(\widehat{ACE}) = \{nP(A = 1)\}^{-1}\sigma_{1, adj}^2 + \{nP(A = 0)\}^{-1}\sigma_{0, adj}^2$, with $\sigma_{a, adj}^2 = \sigma_a^2 deff_w^a$ for $a \in \{0, 1\}$. For large n, under the null t is approximately standard normal, thus H_0 is rejected when $|t| > z_{1-\alpha/2}$, where α is the type I error rate and z_q is the q^{th} quantile of the standard normal distribution.

Proposition 2:

The sample size required to achieve power $1-\beta$ for effect size $ACE = \delta$ and type I error rate α is approximately:

$$n_{deff} = \frac{(1 + k)\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2(\sigma_{1, adj}^2/k + \sigma_{0, adj}^2)}{\delta^2} \quad (7)$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

where $k = P(A = 1)/P(A = 0)$ is the marginal odds of treatment in the population. Note n_{deff} is the total required sample size across the two treatments.

The sample size formula (7) is the standard sample size equation commonly used to design RCTs, but with σ_a^2 replaced by $\sigma_{a,adj}^2$ (Chow et al., 2017, page 48). Thus, Proposition 2 simplifies power and sample size calculations for observational studies by allowing researchers to design studies as if they were designing an RCT, but inflating the assumed variances by the approximated design effects. The researcher first assumes that no confounding is present, specifies the desired α and $1-\beta$, and makes assumptions about σ_0^2, σ_1^2, δ and k. The design effect is then approximated. When data from a pilot or prior study are available, $deff_w^1$ and $deff_w^0$ can be approximated based on (5) for each treatment group. When no prior study data are available, the distribution of the anticipated weights can be estimated based on assumptions about the distribution of L and $A \mid L$ and the design effect can be calculated based on (4). While these assumptions may not be easy to make, this approach notably requires no assumptions about the potential outcomes Y_0 and Y_1 and their associations with A and L. Further discussion about these practical considerations is included in Section 5. Once the design effects are approximated by $\widetilde{deff}_w^a$ or $\widehat{deff}_w^a$, adjusted variances $\sigma_{a,adj}^2$ can be estimated by $\tilde{\sigma}_{a,adj}^2 = \sigma_a^2 \widetilde{deff}_w^a$ or $\hat{\sigma}_{a,adj}^2 = \sigma_a^2 \widehat{deff}_w^a$, respectively, for $a \in \{0, 1\}$.

4. Simulation Study

4.1 Simulation Scenarios

Simulation studies were conducted to demonstrate use of the design effect in study design under a variety of confounding structures and outcome types. Data were simulated according to Scenarios 1–4 shown in Table 1 and also Scenario 5 described below. For all scenarios, $\alpha = 0.05$ and $1 - \beta = 80\%$ were chosen.

4.2 Sample Size Calculation

Two general approaches can be used to design a study with the design effect approximation: when prior study data are not available, as in Scenarios 1–4, and when prior study data are available, as in Scenario 5. One example from each general approach is presented in detail.

4.2.1 Example 1: No prior study data (Scenario 1b).—

Suppose no prior study data are available to design the study of interest. Then, the researcher must make the same assumptions and design choices as when designing an RCT, namely by specifying α, $1 - \beta$, σ_0^2, σ_1^2, δ, and k. In general, σ_1^2 can be determined by deriving the marginal distribution of Y_1 based on the assumed distributions of $Y_1 \mid L$ and L. For Scenario 1b, $P(Y_1 = 1) = \sum_{l=0}^{1} P(Y_1 = 1 \mid L = l)P(L = l) = 0.58$, and thus $\sigma_1^2 = 0.2436$. Similarly, $\sigma_0^2 = 0.1971$. Here, the average causal effect is assumed to be $\delta = -0.15$. The proportion of the population receiving treatment can be derived by integrating the distribution of $A \mid L$ over L. For Scenario 1b, $P(A = 1) = \sum_{l=0}^{1} P(A = 1 \mid L = l)P(L = l) = 0.65$, and thus k

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

≈ 1.857. When prior study data are not available, the distribution of the IPTWs must be assumed at the design phase. Based on the assumptions in Table 1, four possible values of W exist. These assumed values of W, along with the joint distribution of A and L, allow for the computation of the design effects using (4). This leads to $\widetilde{deff}_w^0 = 1.12$ and $\widetilde{deff}_w^1 = 1.04$, with approximated adjusted variances of $\tilde{\sigma}_{0,adj}^2 = 0.2208$ and $\tilde{\sigma}_{1,adj}^2 = 0.2533$.

Under the assumptions outlined in Table 1 for Scenario 1b, to achieve 80% power to detect an average causal effect of −0.15 at the $\alpha = 0.05$ level, a sample size of approximately $n_{deff} = 356$ is required based on Proposition 2. The design effects and required sample sizes for Scenarios 1a, 1c, and Scenarios 2–4 can be determined similarly and are presented in Table 2. Note Scenarios 1 and 3 have the same design effects because in both instances the joint distribution of A and L is the same. Likewise, Scenarios 2 and 4 have the same design effects.

4.2.2 Example 2: Prior study data (Scenario 5).—

Prior study or pilot data may allow for better informed assumptions about σ_0^2, σ_1^2, δ, and k. Because $\sigma_a^2 = E\left(Y_a^2\right) - \{E(Y_a)\}^2$, σ_a^2 can be estimated by obtaining $\hat{E}\left(Y_a^2\right)$ and $\hat{E}(Y_a)$ from fitted MSMs based on the prior study data. The estimate $\widehat{ACE}$ and prevalence of the exposure or treatment in the prior study can inform assumptions about δ and k.

As an example, consider designing a new study based on the National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study (NHEFS) example presented in Chapter 12 of Hernán and Robins (2020). Hernán and Robins use MSM with IPTWs to estimate the average causal effect of smoking cessation (A) on weight gain after approximately 10 years of follow-up (Y) based on the NHEFS sample of smokers (n = 1566), assuming conditional exchangeability based on nine baseline confounders L: sex, age, race, education, smoking intensity, duration of smoking, physical activity, exercise, and weight.

Making the same assumptions as Hernán and Robins (2020), Scenario 5 considers the design of a new study to estimate the average causal effect of smoking cessation on 10-year weight gain. Based on the NHEFS data, assume that $\sigma_0^2 = 56.1$ and $\sigma_1^2 = 74.0$, obtained by fitting MSMs with IPTWs to estimate $E\left(Y_a^2\right)$ and $E(Y_a)$. In the Hernán and Robins example, $\widehat{ACE} = 3.441kg$. The new study will be designed to provide approximately 80% power to detect a difference in weight gain of $\delta \in \{1.0, 2.0, 3.0\}kg$ for Scenarios 5a-5c, respectively. From the NHEFS sample, assume $k \approx 0.346$.

When prior study data are available, $deff_w^0$ and $deff_w^1$ can be estimated using (5). For the NHEFS data, $\widehat{deff}_w^0 = 1.03$ and $\widehat{deff_w}^1 = 1.24$. This leads to approximated adjusted variances of $\hat{\sigma}_{0,adj}^2 = 57.78$ and $\hat{\sigma}_{1,adj}^2 = 91.76$. Based on these assumptions, a sample size of $n_{deff} = 853$ is needed to achieve approximately 80% power to detect an average causal effect of $2.0kg$ at the $\alpha = 0.05$ level using MSM with IPTWs.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

4.2.3 Naïve Sample Size Calculations.

—As a comparison, sample sizes n_{rct} were calculated naively under the assumptions of an RCT, ignoring the effect of weighting on the variances of the estimates. In other words, sample sizes were calculated as demonstrated above, except using σ_a^2 instead of $\tilde{\sigma}_{a,\,adj}^2$ or $\hat{\sigma}_{a,\,adj}^2$ from Table 2. The sample size n_{rct} represents the total across the two treatments assuming that the marginal probability of treatment in the planned study is the same as in the population.

4.3 Evaluation

For Scenarios 1–4, empirical power based on samples of size n_{deff} was evaluated via simulation by following these steps:

i. Generate a superpopulation of size $N = 1{,}000{,}000$ based on distributions in Table 1.

ii. Select a sample of size n_{deff} without replacement from the superpopulation, where n_{deff} is specified in Table 2.

iii. Estimate $\widehat{W}_i$ for each member of the sample based on the predicted values from the logistic regression of A on L.

iv. Fit the MSM $E(Y_a) = \beta_0 + \beta_1 a$ using weighted least squares, treating the weights as estimated by stacking the estimating equations from the weight model with the estimating equations for the causal means and difference in causal means using the geex package in R (Saul and Hudgens, 2020).

v. Test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ using a Wald test, rejecting H_0 at the $\alpha = 0.05$ significance level.

vi. Repeat steps (ii)-(v) $R = 2000$ times and calculate empirical power as the proportion of simulated samples where H_0 was rejected.

For Scenario 5, empirical power based on a sample of size n_{deff} was evaluated via simulation by following these steps:

i. Estimate the propensity score for each of the 1566 NHEFS participants from a logistic regression model of A on L as $\hat{p}_i = \widehat{Pr}(A_i = 1 \mid L_i = l_i)$. As in Hernán and Robins (2020), the logistic regression model includes main effects for each of the nine baseline confounders and quadratic terms for the four continuous covariates.

ii. For each participant, calculate $\hat{Y}_{ai}$, $a \in \{0, 1\}$, as the predicted value $\hat{E}(Y_{ai} \mid L_i = l_i)$ from the following linear regression model, fit only on participants with $A = a$: $E(Y_{ai} \mid L_i = l_i) = l_i\beta$, where l_i is a vector for participant i that includes an intercept term, the 9 previously defined covariates, and the four quadratic terms corresponding to continuous covariates. Also compute $\widehat{Var}(Y_{ai} \mid L_i = l_i) = MSE_a$, where MSE_a is the mean squared error from the model for $E(Y_{ai})$.

iii. Add $3.441 - \delta$ to $\hat{Y}_{0i}$ for all participants, such that $ACE = \delta$ in the simulated population instead of 3.441 as in the NHEFS sample.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

iv. Select a sample of size n_{deff} with replacement from the NHEFS dataset, where n_{deff} is specified in Table 2.

v. Randomly sample A_i from $Bernoulli(\hat{p}_i)$.

vi. Let $Y_{ai} = \hat{Y}_{ai} + \epsilon_{ai}$, where $\epsilon_{ai} \sim N(0, \widehat{Var}(Y_{ai} \mid L_i = l_i))$.

vii. Follow steps (iii)-(v) from the above list for Scenarios 1–4.

viii. Repeat steps (iv)-(vii) $R = 2000$ times and estimate empirical power as the proportion of simulated samples where H_0 was rejected.

The estimated propensity score distributions from simulated data sets for each of Scenarios 1–5 are displayed in Web Figure 5.

For each scenario, the steps above were repeated to calculate empirical power based on the naïve sample sizes, replacing n_{deff} with n_{rct}.

The results of the simulation study are presented in Figure 1. For all simulation scenarios, when the sample size was calculated using the design effect, empirical power was close to or exceeded the nominal 80% level. That is, use of the design effects to calculate required sample sizes led to close to the intended level of statistical power (Figure 1A). On the other hand, ignoring the effect of weighting and basing sample sizes on the naïve assumptions of an RCT led to empirical power that was lower than the nominal 80% level for all but Scenarios 3a-3c (Figure 1B). These results demonstrate that ignoring the weights in power and sample size calculations can lead to significantly underpowered studies, particularly when there are strong confounders that lead to high variability in the weights.

4.4 Additional Simulations

The simulations summarized in Section 4.3 demonstrate the performance of the design effect approximation for Scenarios 1–4 when there is no pilot study data. Additional simulations were conducted to evaluate the design effect approximation in these scenarios when pilot study data are available to estimate k and $\sigma^2_{a,\,adj}$ for $a \in \{0, 1\}$. Results of these simulations are presented in Web Appendix B.1 and are similar to those presented in Section 4.3. Additional simulations were also conducted under the null hypothesis to empirically confirm type I error control; see Web Appendix B.2.

5. Practical Considerations

When prior study data are not available, specifying the design effects can be challenging. A few general guidelines are offered to help researchers determine reasonable assumptions to facilitate power and sample size calculations.

When only a few categorical covariates will be included in the weight model, researchers can use subject matter knowledge or prior study information to nonparametrically specify the joint distribution of A and L, or the marginal distribution of L and the conditional distribution of $A \mid L$ (as in Example 1). Based on these assumptions, the anticipated weights

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

can be calculated nonparametrically and the design effects for each treatment group can be approximated.

When specification of these distributions is not feasible, researchers can forgo approximating the values of the weights and instead consider more generally how much variation is expected in the weights. The lower bound for $deff_w^a$ is 1, which implies that the weights within both treatment groups are all equal and thus covariates are not predictive of the treatment. Design effects tend to increase when more covariates are added to the weight model. The presence of covariates that are strong predictors of treatment tends to increase the design effect. Care must be taken to identify the appropriate set of confounders to include in the weight model (Vansteelandt et al., 2012). Inclusion of instrumental variables, which are predictive of the exposure but which do not affect the outcome, inflate the variance of the ACE estimator without reducing bias (Rubin, 1997; Myers et al., 2011). The use of weight truncation will decrease the design effect.

Figure 2 depicts propensity score distributions for various values of the design effect and mean propensity score $E(p)$ to aid researchers in choosing a design effect consistent with the expected variation in the weights. The propensity score distributions were generated by N_a = 1000 random draws from beta distributions with shape parameters set to achieve the desired $E(p)$ and design effect. The corresponding weight distributions are included in Web Figure 6. As variation in the propensity scores and thus the weights increases, so does the design effect approximation.

6. Estimating the Remainder with Pilot Study Data

The design effect estimators (4) and (5) omit the approximation error Er_a from (3). In this section estimators of the design effect are considered which include an estimate of Er_a. First we revisit the simulation study in Section 4 to examine the magnitude of the approximation errors for the data generating processes considered. Then we consider design effect estimators which incorporate an estimate of Er_a when pilot data are available.

Approximation errors Er_a were calculated for each of the scenarios included in the simulation study. For Scenarios 1–4, Er_a were calculated using the known distributions in Table 1 and are presented in Web Figure 4. For Scenario 5, Er_a were estimated empirically as approximately 0.03 for $a = 0$ and −0.03 for $a = 1$. Approximation errors were small for most scenarios and were in opposite directions for the two treatment groups, which tends to offset the effects of the errors (Web Figure 4). Scenario 2 had large approximation errors ($Er_0 = 0.59$ and $Er_1 = -0.18$ for Scenario 2b), but empirical power still equaled the nominal level when the design effect estimators (4) and (5) were used to calculate the sample size (Figure 1 and Web Figure 1). Note Scenario 2 is an extreme example, as it includes only a single and very strong confounding variable and only two possible and extreme values for W. For the binary outcome, this resulted in large approximation errors.

When pilot or prior data are available, Er_a can be estimated empirically rather than ignored, and a modified estimator of the design effect can be used for power or sample size calculations: $\widehat{deff}_{w,rem}^{a} = \widehat{deff}_{w}^{a} + \widehat{Er}_a$, where

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\widehat{Er}_a = \left\{N_a / \left(n_p \hat{\sigma}_a^2\right)\right\} \frac{\sum_{i=1}^{n_p} \widehat{W}_i I(A_i = a)\left\{\widehat{W}_i - \hat{E}\left(\widehat{W}_a\right)\right\}(Y_i - \hat{\mu}_a)^2}{\sum_{i=1}^{n_p} \widehat{W}_i I(A_i = a)}$$

where $\hat{E}\left(\widehat{W}_a\right)$ is estimated empirically from the pilot sample and $\hat{\sigma}_a^2$ is the estimator of σ_a^2 described in Section 4.2.2. Adjusted variances $\sigma_{a,adj}^2$ can then be estimated by $\hat{\sigma}_{a,adj}^2 = \hat{\sigma}_a^2 \widehat{deff}_{w,rem}^a$ for $a \in \{0, 1\}$, and the sample size calculated using (7).

Additional simulations were conducted using $\widehat{deff}_{w,rem}^a$ to calculate required sample sizes. The simulations are described in Web Appendix B.3 with the results in Web Figure 3 and Web Figure 4. For Scenarios 1 and 3 (small design effect scenarios), estimates of Er_a were empirically unbiased, even for small pilot sample sizes (Web Figure 4). However, for Scenarios 2 and 4 (large design effect scenarios), estimates of Er_a demonstrate considerable empirical bias for small pilot samples, which was reduced as the pilot sample size increased. As shown in Web Figure 3, for all scenarios, empirical power was similar to the results in Figure 1. Thus, there may not be much benefit in using the modified design effect estimator $\widehat{deff}_{w,rem}^a$ instead of $\widehat{deff}_w^a$.

7. Discussion

The design effect approximation simplifies power and sample size calculations of observational studies. Using the design effect allows researchers to utilize standard power and sample size software (e.g., nQuery, SAS Proc Power) for randomized trials, but with variances inflated by the approximate design effects. An additional advantage of using the design effect approximation (4) that ignores the remainder term is that no assumptions are required about the relationship between the potential outcomes and either the treatment or the confounders. Empirical results demonstrate the design effect approximation can yield the nominal level of power over a range of confounding and outcome structures.

Approximating the design effect when planning an observational study may be challenging. In survey sampling, it is common practice to report estimated design effects in analytic reports for better understanding of the precision of the estimates and to assist other researchers who are designing similar studies (see, for example Center for Behavioral Health Statistics and Quality, 2019). Reporting the estimated design effects corresponding to treatment or exposure effect estimates in observational studies may assist researchers with future study designs. In time, as more studies analyzed with IPTW estimators start to report their design effects, rules of thumb and practical upper bounds for the design effects will likely emerge to aid in the design of future studies (see, for example, United Nations Statistical Division (2008, page 41), Daniel (2012, page 251), and Salganik (2006) from the survey sampling literature).

In the absence of knowledge of estimated design effects from prior studies, the design effect may be approximated either using (4) or, if pilot data are available, (5). In either case, the

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

remainder in Proposition 1 is ignored, which in principle may introduce bias. The remainder may be large when individuals with extreme weight values tend to have potential outcomes that are also extreme relative to the mean. Nonetheless, simulation studies in Section 4 demonstrate empirically that using either (4) or (5) for sample size determination tends to yield the desired power. Of course, there may be other scenarios where this is not the case, and thus care should be exercised in generalizing beyond the simulation scenarios considered in this paper. When pilot data are available, the approximation error Er_a can be estimated as in Section 6. Empirical power when approximating Er_a tended to be similar to power when the remainder was ignored. Thus the simpler estimator that ignores the remainder may be preferred in practice.

In conclusion, the design effect approximation can be a useful tool for the design of studies to estimate point exposure effects with IPTW estimators, as currently no power and sample size methods exist in this context. The design effect can also be used in precision calculations using approaches analogous to those described in this paper, i.e., basing calculations on the adjusted variances $\tilde{\sigma}^2_{a,\,adj}$ or $\hat{\sigma}^2_{a,\,adj}$ rather than σ^2_a.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank Stephen Cole, Noah Greifer, Bryan Blette, Kayla Kilpatrick, Shaina Alexandria, and Jaffer Zaidi for their helpful suggestions. Thank you also to the Co-Editor, the Associate Editor, and the two anonymous reviewers for their helpful comments that strengthened the paper. This work was supported by NIH grant R01 AI085073. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Data Availability Statement

The NHEFS data that support the findings in this paper are openly available on the Harvard University Website at https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/ (Hernán, 2020).

References

Austin PC (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statistics in Medicine 28, 3083–3107. [PubMed: 19757444]

Austin PC and Stuart EA (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Statistics in Medicine 34, 3661–3679. [PubMed: 26238958]

Brumback BA, Hernán MA, Haneuse SJ, and Robins JM (2004). Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. Statistics in Medicine 23, 749–767. [PubMed: 14981673]

Center for Behavioral Health Statistics and Quality (2019). 2017 National Survey on Drug Use and Health Methodological Resource Book, Section 11: Person-Level Sampling Weight Calibration. Technical report, Substance Abuse and Mental Health Services Administration.

Chow S-C, Shao J, Wang H, and Lokhnygina Y (2017). Sample Size Calculations in Clinical Research. Chapman and Hall/CRC.

Cole SR and Hernán MÁ (2008). Constructing inverse probability weights for marginal structural models. American Journal of Epidemiology 168, 656–664. [PubMed: 18682488]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Daniel J (2012). Sampling Essentials: Practical Guidelines for Making Sampling Choices. Sage Publications.

Gabler S, Häder S, and Lahiri P (1999). A model based justification of Kish's formula for design effects for weighting and clustering. Survey Methodology 25, 105–106.

Hernán M (2020). NHEFS data. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book.

Hernán MÁ, Brumback B, and Robins JM (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology pages 561–570. [PubMed: 10955409]

Hernán MÁ and Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.

Kish L (1965). Survey Sampling. John Wiley & Sons.

Kish L (1992). Weighting for unequal pi. Journal of Official Statistics 8, 183–200.

Lee BK, Lessler J, and Stuart EA (2011). Weight trimming and propensity score weighting. PLOS ONE 6, 1–6.

Lunceford JK and Davidian M (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine 23, 2937–2960. [PubMed: 15351954]

McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, and Burgette LF (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Statistics in Medicine 32, 3388–3414. [PubMed: 23508673]

McCaffrey DF, Ridgeway G, and Morral AR (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological Methods 9, 403–425. [PubMed: 15598095]

Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, Joffe MM, and Glynn RJ (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. American Journal of Epidemiology 174, 1213–1222. [PubMed: 22025356]

Robins JM, Hernán MÁ, and Brumback B (2000). Marginal structural models and causal inference in Epidemiology. Epidemiology 11, 550–560. [PubMed: 10955408]

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

Rosenbaum PR and Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Rubin DB (1997). Estimating causal effects from large data sets using propensity scores. Annals of internal medicine 127, 757–763. [PubMed: 9382394]

Salganik MJ (2006). Variance estimation, design effects, and sample size calculations for respondent-driven sampling. Journal of Urban Health 83, 98–112.

Saul BC and Hudgens MG (2020). The calculus of M-estimation in R with geex. Journal of Statistical Software 92, 1–15.

Stefanski LA and Boos DD (2002). The calculus of M-estimation. The American Statistician 56, 29–38.

United Nations Statistical Division (2008). Designing Household Survey Samples: Practical Guidelines, volume 98. United Nations Publications.

Valliant R, Dever JA, and Kreuter F (2013). Practical Tools for Designing and Weighting Survey Samples. Springer.

Vansteelandt S, Bekaert M, and Claeskens G (2012). On model selection and model misspecification in causal inference. Statistical Methods in Medical Research 21, 7–30. [PubMed: 21075803]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

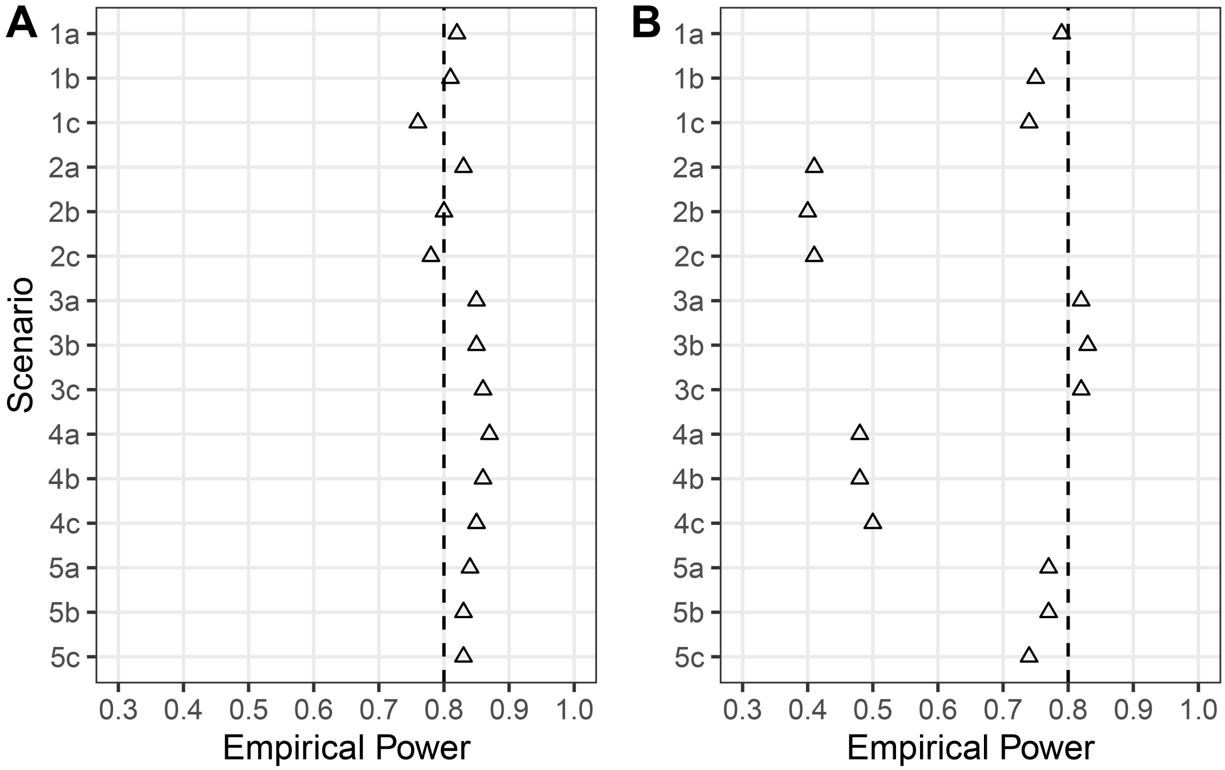

Figure 1.

Empirical power from the simulation study by scenario across $R = 2000$ samples based on sample sizes (A) n_{deff} and (B) n_{rct} from Table 2. Empirical power is the proportion of simulated samples in which the p-value for testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ was less than $\alpha = 0.05$ for the MSM $E(Y_a) = \beta_0 + \beta_1 a$. (Scenario 5c excludes 2 and 5 simulations for (A) and (B), respectively, in which the geex package failed to converge when estimating the standard error of $\widehat{ACE}$.)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

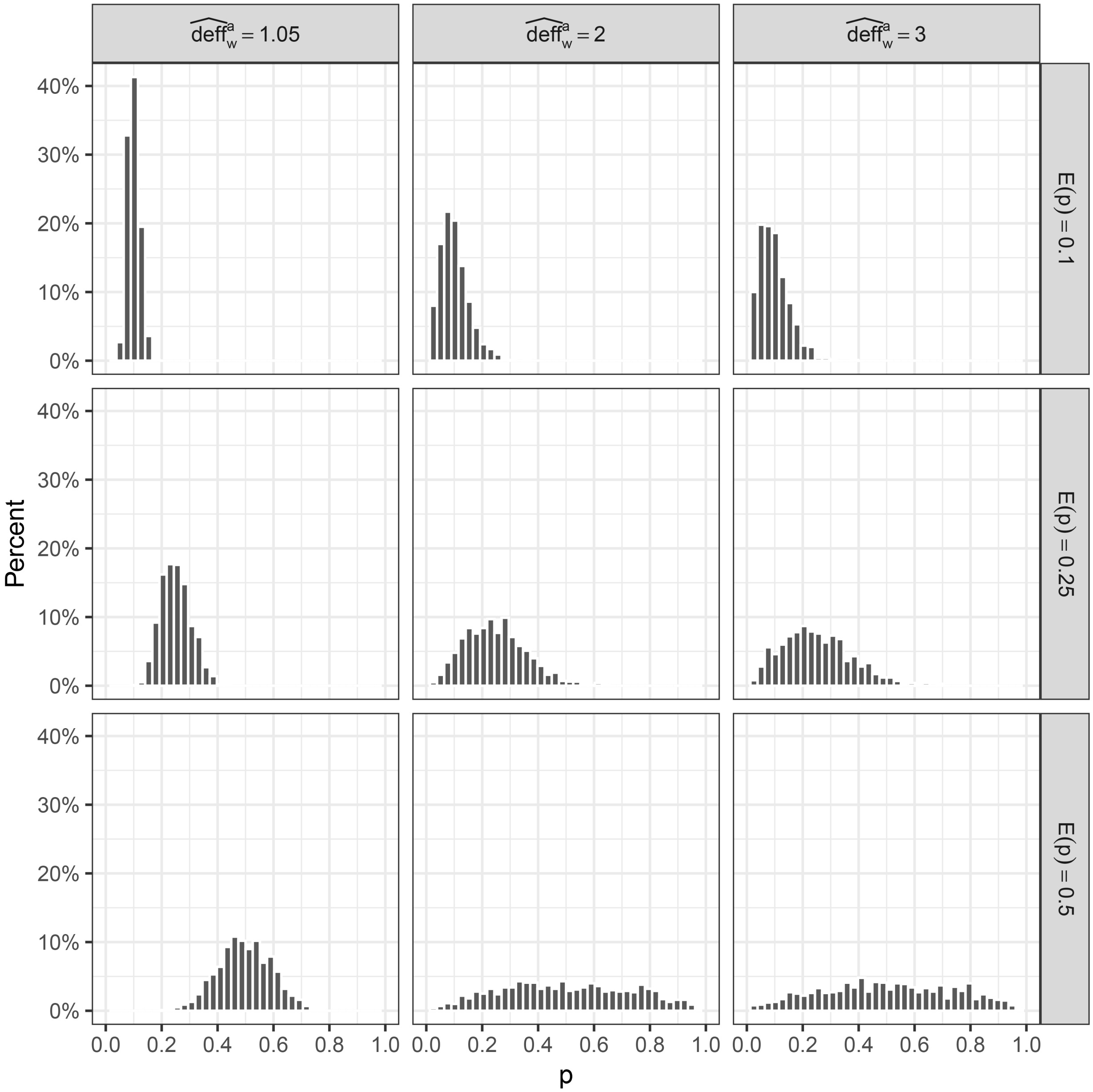

Figure 2.

Examples of propensity score distributions for various approximated design effects and mean propensity score $E(p)$. Distributions were generated from $N_a = 1000$ random draws from beta distributions with shape parameters set to achieve the desired $E(p)$ and design effect.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1:

Simulation scenarios where no prior study data are available. B(p) indicates Bernoulli distribution with mean p, and N(m, v) indicates Normal distribution with mean m and variance v.

	Scenario	Confounders (L)	Exposure ($A \mid L = l$)	Potential Outcomes ($Y_a \mid L = l$)	δ
1	binary Y, small $deff_w^a$	$B(0.6)$	$B(0.5 + 0.25l)$	$B(0.85 - 0.2l + \delta_a)$	(a) −0.10 (b) −0.15 (c) −0.20
2	binary Y, large $deff_w^a$	$B(0.5)$	$B(0.1 + 0.8l)$	$B(0.85 - 0.2l + \delta_a)$	(a) −0.10 (b) −0.15 (c) −0.20
3	continuous Y, small $deff_w^a$	$B(0.6)$	$B(0.5 + 0.25l)$	$N(20 - 10l + \delta_a, 144 + 112_a)$	(a) 2.5 (b) 5.0 (c) 7.5
4	continuous Y, large $deff_w^a$	$B(0.5)$	$B(0.1 + 0.8l)$	$N(20 - 10l + \delta_a, 144 + 112_a)$	(a) 2.5 (b) 5.0 (c) 7.5

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2:

Variances, approximated design effects, approximated adjusted variances, and required sample sizes for simulation scenarios by treatment.

	Scenario	a	σ_a^2	$\widetilde{deff}_w^a$ or $\widehat{deff}_w^a$	$\tilde{\sigma}^2_{a,adj}$ or $\hat{\sigma}^2_{a,adj}$	n_{deff}	n_{rct}
1	binary Y, small $deff_w^a$	0 1	0.1971 0.2436	1.12 1.04	0.2208 0.2533	(a) 801 (b) 356 (c) 201	(a) 736 (b) 327 (c) 184
2	binary Y, large $deff_w^a$	0 1	0.1875 0.2400	2.78 2.78	0.5208 0.6667	(a) 1862 (b) 828 (c) 466	(a) 671 (b) 298 (c) 168
3	continuous Y, small $deff_w^a$	0 1	168.0 280.0	1.12 1.04	188.2 291.2	(a) 1237 (b) 310 (c) 138	(a) 1143 (b) 286 (c) 127
4	continuous Y, large $deff_w^a$	0 1	169.0 281.0	2.78 2.78	469.4 780.6	(a) 3136 (b) 784 (c) 349	(a) 1129 (b) 283 (c) 126
5	prior study data, (NHEFS)	0 1	56.10 74.00	1.03 1.24	57.78 91.76	(a) 3409 (b) 853 (c) 379	(a) 2850 (b) 713 (c) 317