Designing efficient randomized trials: power and sample size calculation when using semiparametric efficient estimators

Alejandro Schuler*[1]
for the Critical Path for Alzheimer's Disease, † the Alzheimer's Disease Neuroimaging Initiative, ‡ and the Alzheimer's Disease Cooperative Study§

[1]Unlearn.AI, Inc., San Francisco, California

July 6, 2021

Abstract

Trials enroll a large number of subjects in order to attain power, making them expensive and time-consuming. Sample size calculations are often performed with the assumption of an unadjusted analysis, even if the trial analysis plan specifies a more efficient estimator (e.g. ANCOVA). This leads to conservative estimates of required sample sizes and an opportunity for savings. Here we show that a relatively simple formula can be used to estimate the power of any two-arm, single-timepoint trial analyzed with a semiparametric efficient estimator, regardless of the domain of the outcome or kind of treatment effect (e.g. odds ratio, mean difference). Since an efficient estimator attains the minimum possible asymptotic variance, this allows for the design of trials that are as small as possible while still attaining design power and control of type I error. The required sample size calculation is parsimonious and requires the analyst to provide only a small number of population parameters. We verify in simulation that the large-sample properties of trials designed this way attain their nominal values. Lastly, we demonstrate how to use this formula in the "design" (and subsequent reanalysis) of a real randomized trial and show that fewer subjects are required to attain the same design power when a semiparametric efficient estimator is accounted for at the design stage.

1 Introduction

Clinical research often aims to estimate the effect of a treatment on an outcome of interest [1]. The randomized trial is the gold standard for causal inference because randomization cancels out the effects of any unobserved confounders in expectation [2, 3, 4]. Randomization, however, does nothing to combat the natural variability that comes with finite samples of a larger population. Because of this, treatment effect estimates from a trial are always accompanied by a measure of uncertainty. The expected degree of uncertainty determines how likely the trial is to positively identify an effect of a certain size (the *power* of the

*aschuler@unlearn.ai

†Data used in the preparation of this article were obtained from the Critical Path Institute's Critical Path for Alzheimer's Disease (CPAD) consortium. As such, the investigators within CPAD contributed to the design and implementation of the CPAD database and/or provided data, but did not participate in the analysis of the data or the writing of this report.

‡Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

§Data used in preparation of this manuscript/publication/article were obtained from the University of California, San Diego Alzheimer's Disease Cooperative Study. Consequently, the ADCS Core Directors contributed to the design and implementation of the ADCS and/or provided data but did not participate in analysis or writing of this report.

trial). Since trials are expensive and time-consuming, it is standard to design for power of ≥80% or 90% to minimize the likelihood of failure.

Several factors affect power. In trials where subjects are assumed independent these factors separate into characteristics of the data-generating process (population, disease, etc.), the aggressiveness of the rule used to determine "success" (e.g. the p-value cutoff), the number of subjects enrolled into the trial (trial size), and the method of data analysis [5]. The population of interest is determined by clinical desiderata and is therefore outside of the statistician's control. The decision rule is tied to a desired type I error rate (false positive rate) and so is typically fixed. This leaves trial size and the analysis method as the primary determinants of power that may be modified.

For example, consider a trial of n subjects with 1:1 randomization. This study is to determine the effect that a new drug has on blood pressure for hypertensive patients after a year of treatment. The treatment effect here is the theoretical difference in blood pressure on treatment vs. placebo averaged across the entire population of hypertensive patients. To estimate the effect using our trial data we will take the difference of the mean blood pressure after one year in the control and treatment groups and test for the effect using an unpaired z-test (two-sided). This is a canonical "unadjusted" analysis. Assuming that the variability in the outcome is the same in each treatment arm (σ^2), it is known that the sampling variance of the effect estimated in this fashion will be near $4\sigma^2/n$ in large-enough samples (where n is the size of the trial) [6]. Basic statistical theory establishes that the power of this trial to detect an effect of size τ at significance level α is therefore $\Phi\left(\Phi^{-1}(\alpha/2) + \sqrt{n}\frac{\tau}{2\sigma}\right) + \Phi\left(\Phi^{-1}(\alpha/2) - \sqrt{n}\frac{\tau}{2\sigma}\right)$ where Φ is the standard normal cumulative distribution function.

In order to achieve a trial with power greater than, say, 80%, we must increase n until the result of the above formula exceeds 0.8. This determines the necessary enrollment for the trial. Formulas such as the one above (or approximations thereof) are routinely used to determine the size of a trial before it is begun [5, 7]. The form of the calculation depends on the estimator and test used [7]. We are using a simple unadjusted estimate and the corresponding z-test in our example so the formula turns out to be relatively simple. Note that the only population characteristic we need to assume (or estimate from prior data) is the outcome variability σ.

Given a target effect, the sampling variance of the estimated effect drives power, regardless of the specific formula associated with a given estimator. The smaller the sampling variance of an estimator, the more power a trial will have. This motivates the use of estimators that have less theoretical sampling variance (are more *efficient*). For a completed trial, using a more efficient estimator would typically result in smaller confidence intervals and smaller p-values. For example, adjusted linear regression (ANCOVA) is typically more efficient[1] than difference-in-means estimation and so is often used in the analysis of trial data even when the trial is powered under the assumption of a difference-in-means analysis [9]. Unfortunately, however, it is difficult to customize power calculations to adjusted estimators without specifying or assuming a large number of population parameters [10, 11, 12, 13].

There is a statistical limit to efficiency in the estimation of treatment effects. That is, without making unverifiable assumptions about the data-generating process, there is a sampling variance that no reasonable estimator can beat. But there are often feasible estimators that attain this limit. These estimators are called *semiparametric efficient*. By definition, no estimator is better than these semiparametric estimators in terms of asymptotic sampling variance [14]. The benefit of using a semiparametric efficient estimator is that the resulting confidence intervals will be as small as possible while still being valid in general settings (e.g. without assumptions of linearity, homoscedasticity, etc.). This in turn implies maximal power while maintaining control of type I error.

Semiparametric efficient estimators have been previously used in the (re)analysis of trial data and related simulations, where they have been empirically shown to be effective tools for increasing confidence (and thus power) while controlling type I error [15, 16, 17, 18, 19, 20, 21]. But to our knowledge it has not been demonstrated how to leverage their beneficial properties in order to benefit trial *design*. That is our task in this paper. In other words, instead of using more efficient estimators in order to shrink uncertainty for a trial with fixed design, we show how to use them to prospectively design smaller trials that maintain their power.

[1] And is *guaranteed* to be if treatment-covariate interactions are included in the adjustment or if randomization is 1:1. [8]

We do this without requiring the estimation or assumption of a large number of population parameters. The practical benefits of this are immense. Each patient in the trial contributes substantial cost and duration, so smaller trials mean cheaper and faster development cycles for treatments that patients need.

Our primary contribution is the development and exposition of a practical power formula that pertains to any semiparametric efficient estimator used in a trial, regardless of the outcome type and (marginal) treatment effect estimand. We verify in simulation that the large-sample properties of trials designed this way attain their nominal values. Lastly, we demonstrate how to use this formula in the "design" (and subsequent reanalysis) of a real randomized trial and show that fewer subjects are required to attain the same design power when a semiparametric efficient estimator is accounted for at the design stage.

2 Background

2.1 Setting and Notation

Our setting is a two-arm randomized controlled trial (RCT) with a single-timepoint outcome. Denote the outcome for subject i in the randomized trial with Y_i, their baseline covariates with X_i, and their treatment assignment with W_i. The trial dataset is a set of n tuples (X_i, W_i, Y_i), which we denote $(\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{Y}) \in \mathcal{X}^n \times \{0,1\}^n \times \mathbb{R}^n$ (we use boldface $\boldsymbol{A}$ to denote a vector of random variables, each associated with one observation in the dataset). Let Y_0 and Y_1 be the control and treatment potential outcomes of the subjects in the trial, respectively, and let $\boldsymbol{Y_W} = \boldsymbol{W}\boldsymbol{Y}_1 + (1-\boldsymbol{W})\boldsymbol{Y}_0$ [22]. Our structural assumption about the trial is,

$$P(\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{Y}, \boldsymbol{Y}_0, \boldsymbol{Y}_1) = \mathbf{1}(\boldsymbol{Y} = \boldsymbol{Y_W}) P(\boldsymbol{W}) \prod_i P(X_i, Y_{0,i}, Y_{1,i}) \,. \tag{1}$$

In other words, a) the observed outcomes are the potential outcomes corresponding to the assigned treatment, b) the treatment is assigned independently of everything (thus at random), and c) our counterfactual trial subjects are independent of each other. In addition to being independent, we also assume the subjects are identically distributed, i.e., $(X_i, Y_{0,i}, Y_{1,i}) \overset{\text{IID}}{\sim} P(X, Y_0, Y_1)$.

Denote the marginal mean outcomes under each treatment condition w as $\mu_w = \mathbb{E}[Y_w]$ and denote the conditional means $\mu_w(X) = \mathbb{E}[Y_w|X]$. Our structural assumptions about the trial ensure that $\mathbb{E}[Y|W=w] = \mathbb{E}[Y_w]$ and $\mathbb{E}[Y|X, W=w] = \mathbb{E}[Y_w|X]$, so the collected data can be used to make inferences about counterfactual quantities of interest. In general, a (marginal) "treatment effect" is any function of the marginal means, i.e. $\tau = r(\mu_0, \mu_1)$. Examples of this are the difference in means $\tau = \mu_1 - \mu_0$ and odds ratio $\tau = (\mu_1/(1-\mu_1))/(\mu_0/(1-\mu_0))$. For the purposes of this paper we also require that $\frac{\partial r}{\partial \mu_0} \leq 0$ and $\frac{\partial r}{\partial \mu_1} \geq 0$. This sensible condition is satisfied by most definitions of the treatment effect, including difference-in-means and odds ratio, and essentially means that the treatment effect is increasing in μ_1 and decreasing in μ_0 as would logically be expected from a definition of a treatment effect.

Finally, we denote the treatment indicators as $W_{1,i} = W_i$ and $W_{0,i} = 1 - W_i$ to allow for symmetric notation. In other words, $W_w = 1(W = w)$ for the random variable W and the constant w. Let $\pi_1 = P(W_1 = 1)$ and let $\pi_0 = P(W_0 = 1)$ be the probability that a subject is assigned to the treatment or control arm in the trial, respectively. In simple randomized experiments, these are constants that apply to all subjects.

In what follows, we abbreviate the usual empirical (sample) average of IID variables $A_1 \ldots A_n \sim A$ with the notation $\widehat{\mathbb{E}}[A] = \frac{1}{n}\sum A_i$ (or $\bar{A}$). Denote an empirical *conditional* average $\widehat{\mathbb{E}}[A|B=b] = \frac{1}{n_b}\sum_{B_i=b} A_i$ with n_b the number of observations where $B_i = b$. Let $\tilde{A} = A - \mathbb{E}[A]$ (or $\tilde{A} = A - \widehat{\mathbb{E}}[A]$) be centered (or empirically centered) versions of the random variable A, with usage clear from context or otherwise noted. Let $\mathbb{V}[A]$ denote the variance of A and $\mathbb{C}[A, B]$ denote the covariance between A and B. When we describe "asymptotic" properties of an estimator in all cases we are referring to the asymptote where the number of observations is increasing while other properties of the data-generating process remain fixed.

2.2 Semiparametric efficient estimators

No "reasonable"[2] estimator can possibly attain an asymptotic variance lower than that of a semiparametric efficient estimator [14]. The benefit of using a semiparametric efficient estimator is therefore that the resulting confidence intervals will be as small as possible while still being valid in general settings (e.g. without assumptions of linearity, homoscedasticity, etc.). This in turn implies maximal power while maintaining control of type I error.

A few semiparametric efficient estimators for marginal treatment effects have been described in the literature [23, 24, 25]. The differences between these estimators pertain to their small-sample properties or the technical assumptions they require to attain efficiency, neither of which are relevant to our current discussion. To fix ideas we will describe one semiparametric efficient estimator in detail here, but our subsequent discussion applies equally to any estimator that attains the semiparametric efficiency bound under the assumptions used for the trial analysis.

The estimator we will examine here is most commonly called augmented inverse propensity weighting (AIPW, which we will use throughout the paper), but it has also been referred to as double machine learning (DML) [25] or the efficient influence function (EIF) estimator [20]. This estimator is popular in the analysis of observational data and is also sometimes called the "doubly robust" estimator in reference to the fact that its estimates are consistent if either the propensity or outcome models are correctly specified (although other estimators also have that property). In the context of a randomized trial, the propensity score is in fact fixed and known, which means this estimator is *always* consistent. The estimator is given by:

$$\begin{aligned} \hat{\tau} &= r\left(\hat{\mu}_0, \hat{\mu}_1\right) \\ \hat{\mu}_w &= \widehat{\mathbb{E}}\left[\frac{W_w}{\pi_w}\left(Y - \hat{\mu}_w^{(-k)}(X)\right) + \hat{\mu}_w^{(-k)}(X)\right] \\ \hat{\mu}_w^{(-k)}(\cdot) &= \mathcal{M}\left(\boldsymbol{X}_w^{(-k)}, \boldsymbol{Y}_w^{(-k)}\right) \end{aligned} \tag{2}$$

To understand the mechanics of this estimator one should work bottom-up in eq. 2. Temporarily ignoring the $(-k)$ superscripts, the first step is to use the data and some form of machine learning algorithm $\mathcal{M}$ (e.g. random forest) to estimate the conditional mean functions $\hat{\mu}_w(X)$ in each treatment group. For this we use the data from all subjects in group w, denoted $(\boldsymbol{X}_w, \boldsymbol{Y}_w)$. The learned functions $\hat{\mu}_w(X)$ represent estimated versions of the true, unknown, conditional means $\mu_w(X) = \mathbb{E}\left[Y|X, W = w\right]$. With those estimates in hand, we plug our data into the middle equation in order to derive point estimates of the *marginal* means $\hat{\mu}_w$. The right-hand side of the middle expression is interpretable as the group-specific mean outcome "augmented" with the model predictions in such a way that eliminates possible bias from the models. Finally, the treatment effect is estimated by plugging the estimated means into the function r (top line).

Now we turn our attention to the $(-k)$ superscripts. For desirable asymptotic properties to hold without additional assumptions, it turns out that the conditional means must be *cross-estimated* from the trial data [26, 27]. Briefly, that means we must split our trial data into $k \in 1 \ldots K$ non-overlapping folds and fit K different models for $\hat{\mu}_w(X)$, each excluding the data from one of the folds. To notate this, let $\hat{\mu}_w^{-k}(X)$ denote the model trained without the kth fold of the data. We use the models trained *without* the kth fold to make predictions *for* the kth fold. In other words, when we need to get the prediction for a subject i, we use the model that omitted that subject from its training set. Intuitively, we do this because we do not want to "train on the test data" and unknowingly overfit the models. Statistically, the cross-fitting process allows us to make statements about the resulting AIPW estimator that are agnostic to the specifics of the machine learning method used to fit the model. Note that when we say "AIPW" throughout this article, we are always referring to AIPW with cross-fit estimates of the conditional means.

It is shown in the appendix (restating known arguments in the context of a randomized trial) that this estimator is semiparametric efficient as long as the predicted conditional means are estimated in a way that is mean-square consistent[3], i.e.

[2] regular and asymptotically linear- see appendix for details.

[3] When the data are observational, the conditions are more strict because the propensity score is unknown and must be

$$\mathbb{E}\left[\left(\hat{\mu}_w^{(-k)}(X) - \mu_w(X)\right)^2\right] \to 0 \tag{3}$$

This justifies the use of many popular machine learning methods $\mathcal{M}$ for learning the functions $\hat{\mu}_w(X)$ [28, 29, 30, 31].

Asymptotic Variance Recall that r is the function that defines the treatment effect from the true mean outcomes $\mu_w = \mathbb{E}[Y_w]$. Let $r'_w = \frac{\partial r}{\partial \mu_w}(\mu_0, \mu_1)$. Any semiparametric efficient estimator is asymptotically normal $\sqrt{n}(\hat{\tau} - \tau) \rightsquigarrow N(0, \nu^{*2})$ where ν^{*2} is the efficiency bound given by

$$\begin{aligned} \nu_*^2 &= \mathbb{V}[\phi] \\ \phi &= r'_0\phi_0 + r'_1\phi_1 \\ \phi_w &= \frac{W_w}{\pi_w}(Y - \mu_w(X)) + (\mu_w(X) - \mu_w) \end{aligned} \tag{4}$$

A consistent standard error for $\hat{\tau}$ can be calculated from an empirical plug-in estimator $\hat{\nu}_*^2 = \widehat{\mathbb{E}}[\phi^2]$, substituting $\hat{\mu}_w^{(-k(i))}(X_i)$ for $\mu_w(X_i)$ and $\hat{\mu}_w$ for μ_w in the expression for ϕ [27, 32, 20]. Having an estimate of the asymptotic variance allows us to presume that $\hat{\tau}$ is distributed approximately as $N(\tau, \hat{\nu}^2/n)$ which allows for the construction of confidence intervals $(\hat{\tau} \pm 1.96 \times \hat{\nu}/\sqrt{n})$ and p-values $(P(|T| > \hat{\tau})$ under the null $T \sim N(0, \hat{\nu}^2/n))$.

2.3 Power

Randomized trials are costly and slow, so sponsors must ensure a high chance of a statistically significant finding when the treatment truly has a clinically significant effect. This chance is assessed by a power calculation, which is performed in the design phase of the trial before any subjects have been enrolled [5]. The purpose of the power calculation is to help the analyst decide how many subjects would be required to ensure adequate power if the data were analyzed with a particular estimator.

Presuming that statistical significance of the result is assessed using a two-sided p-value cutoff $p < \alpha$, the probability of this event occurring when in fact the true effect is τ such that $\hat{\tau} \sim N(\tau, \nu^2/n)$ is

$$\text{Power} = \Phi\left(\Phi^{-1}(\alpha/2) + \sqrt{n}\frac{\tau}{\nu}\right) + \Phi\left(\Phi^{-1}(\alpha/2) - \sqrt{n}\frac{\tau}{\nu}\right) \tag{5}$$

where Φ denotes the CDF of the standard $N(0, 1)$ normal distribution.

Assuming a particular target effect τ, this formula allows the analyst to pick n and an estimator $\hat{\tau}$ with asymptotic variance ν^2 such that the resulting power is greater than a desired fraction, e.g. 80%. The trouble in doing this is that the asymptotic variance ν^2 of any given estimator ultimately depends on the specifics of the counterfactual data-generating (population) distribution $P(Y_0, Y_1, X)$ in the trial, which is unknown. For many estimators this requires estimating or guessing a large number of parameters or otherwise making strong assumptions [10, 11, 12, 13].

To get around this problem, it is common to perform the power analysis with an estimator that must always attain a larger sampling variance than the one that will ultimately be used in the analysis, but which allows for tractable estimation of the asymptotic sampling variance from a small number of interpretable population-level parameters. Since the variance of the simplified estimator is greater or equal to that of the

estimated. The product of the propensity and outcome model residuals must decay at a $\sqrt{n}$ rate to ensure that the resulting AIPW estimator is asymptotically normal as desired. Moreover, further assumptions are required to bound the estimated propensity scores away from 0 and 1 so that the difference of their inverses converges at a the same rate. These (rather strict) conditions are unnecessary when the treatment is randomized.

estimator used for analysis, the resulting power will be a conservative estimate of the true power (up to correct specification of the population parameters).

For example, consider a 1:1 randomized trial with a continuous primary outcome in which the data will be analyzed with linear regression (ANCOVA). It may be shown that the difference-in-means (or "unadjusted") estimator $\hat{\tau}_\Delta = \widehat{\mathbb{E}}[Y|W=1] - \widehat{\mathbb{E}}[Y|W=0]$ has asymptotic variance $\nu^2 = 2\mathbb{V}[Y_0] + 2\mathbb{V}[Y_1]$, which is always greater than or equal to the asymptotic variance of the linear regression estimator in this case. The benefit of using the unadjusted estimator for the power calculation is that the asymptotic variance only depends on the presumed marginal variances of the two potential outcomes: $\mathbb{V}[Y_w]$. Under a the simplifying assumption $\mathbb{V}[Y_0] = \mathbb{V}[Y_1]$ this reduces to a single number. In any case, these are interpretable parameters that could be estimated from existing data from prior trials, registries, or electronic health records; or even guessed at by a experienced domain expert. Fixing these parameters effectively fixes ν, which allows the analyst to vary n until the desired power is reached. This ensures that the true power using the linear regression analysis will actually exceed the design power by some unknown amount.

Use the ANCOVA estimator but use the difference-in-means variance estimator to calculate sample size

3 Approach

The obvious downside to the usual conservative approach to powering trials is that more subjects are enrolled than are actually necessary to attain the design power (assuming correct specification of the required population parameters). This is of no scientific concern since we may be confident that the analysis is well-powered, but it is of enormous practical concern because each additional subject adds significant cost, time, and complexity to the trial.

3.1 Accounting for Semiparametric Efficiency

Our aim here is therefore to examine the asymptotic variance of the semiparametric estimators described in section 2.2 and reduce it down to a small number of population-level parameters which are estimable from data and/or have natural interpretations. An expression of that kind would allow the analyst to proceed with the power calculation without substituting the anticipated (larger) variance from a simplified estimator. This, in turn, would remove undesired conservatism and allow for smaller trials that maintain their design power.

Theorem 1. *Let $\sigma_w^2 \equiv \mathbb{V}[Y_w]$ be the marginal outcome variances in each treatment arm and let $\kappa_w^2 \equiv \mathbb{E}[\mathbb{V}[Y_w|X]]$ be the corresponding average conditional variances. Moreover define $\gamma = Corr[\mu_0(X), \mu_1(X)]$, the correlation between the conditional means and recall $r'_w = \frac{\partial r}{\partial \mu_w}(\mu_0, \mu_1)$.*

Subject to mild regularity conditions on the data-generating distribution, the asymptotic variance of any semiparametric efficient estimator of the parameter $\tau = r(\mu_0, \mu_1)$ is

$$\nu_*^2 = r_0'^2\left(\frac{\pi_1}{\pi_0}\kappa_0^2 + \sigma_0^2\right) + r_1'^2\left(\frac{\pi_0}{\pi_1}\kappa_1^2 + \sigma_1^2\right) - 2|r'_0 r'_1|\gamma\sqrt{(\sigma_0^2 - \kappa_0^2)(\sigma_1^2 - \kappa_1^2)} \tag{6}$$

The proof of this, which requires nothing but algebra, is given in the appendix.

Special cases There are a few interesting special cases of this relationship. The first is when $\mu_w(X) = \mu_w$ are constants. In this case we have $\kappa_w^2 = \sigma_w^2$ (note $\kappa_w^2 \leq \sigma_w^2$ by the law of total variance) and the above reduces to $\nu_*^2 = r_0'^2\frac{\sigma_0^2}{\pi_0} + r_1'^2\frac{\sigma_1^2}{\pi_1}$, the variance of the unadjusted (difference in means) estimator. In other words, the unadjusted estimator is efficient when the conditional means are constant, as should be obvious because the covariates impart no exploitable information.

A special case that illuminates the role of γ is when $\pi_0 = \pi_1$, $\sigma_0 = \sigma_1 = \sigma$, and $\kappa_0^2 = \kappa_1^2 = \kappa^2$. Presume the estimand of interest is $\tau = \mu_1 - \mu_0$ such that $r'_0 = -1$ and $r'_0 = 1$. In this case the above reduces to

$\nu_*^2 = 2[(1-\gamma)\sigma^2+(1+\gamma)\kappa^2]$. Compare this to the asymptotic variance of the unadjusted estimator under these conditions, which is $\nu_\Delta^2 = 4\sigma^2$. Indeed, in the worst case scenario (when $\gamma = -1$) our semiparametric efficient estimator gives the same asymptotic variance as the unadjusted estimator, meaning that no efficiency gain is possible from covariate adjustment. In the best-case scenario ($\gamma = 1$) the asymptotic variance is $\nu_*^2 = 4\kappa^2$. Note that this case obtains when there is a constant treatment effect across the population. When $\gamma = 0$ we get the intermediate $\nu_*^2 = 2\sigma^2 + 2\kappa^2$.

The takeaway is that γ mediates the the extent to which the asymptotic variance depends on the *marginal variance* (worst case) vs. the *average conditional variance* (best case) of the potential outcomes in each treatment arm. This again should be intuitive. The law of total variance states that the average conditional variance is the component of the marginal variance that is left over when the variance of the conditional mean is subtracted off. By performing a nonlinear adjustment for the covariates, we come close to recovering the true conditional mean in large samples and we can "explain away" exactly that component of the variance.

3.2 Prospective estimation

Recall that our goal is to estimate the asymptotic variance *before* running the trial so that we may use it in conjunction with eq. 5 to find an sample size that should suffice to attain a desired level of power. The relationship in thm. 1 shows that only σ_w^2, κ_w^2 and γ are required to calculate ν_* (after specifying μ_w in order to set the target effect τ, which fixes r'_w as well). Our task is therefore to hypothesize values or bounds for these parameters or to estimate them using pre-existing data.

The marginal variances σ_w^2 should be familiar to most analysts. Usually there is existing knowledge about the marginal variance under standard-of-care from registry data, electronic health records, or prior studies on similar populations. This variance is taken to be σ_0^2 and it is most often assumed that $\sigma_1 = \sigma_0$ because there is rarely reliable data on treatment-arm outcomes (since the treatment is often new and experimental and phase I/II trials are typically small). If treatment-arm data is available, then σ_1 can be estimated independently.

The average conditional variances κ_w^2 are less-familiar quantities, but they also admit natural interpretations and estimators. One way to estimate an upper bound on the average conditional variance in either treatment arm is to average known (marginal) variances across sub-populations defined by the planned adjustment covariates. For instance, presume the anticipated trial population consists of about equal numbers of men and women and that biological sex is a planned adjustment covariate. Then the mean of the marginal outcome variance among men in treatment arm w and the marginal outcome variance among women in treatment arm w is a consistent estimator for an upper bound on the average conditional variance in treatment arm w. The population can be arbitrarily divided in this way as many times as existing data permits as long as the splits are pre-specified. By the bounding argument, this estimate of average conditional variance is likely to be larger than the true value, yielding a conservative estimate for power that still offers gains over the traditional difference-in-means power calculation. If treatment-arm data are not available or are too scant to allow for subdivision, one might assume that $\kappa_1^2 = \kappa_0^2$, or, very conservatively, that $\kappa_1^2 = \sigma_1^2$ (recall $\kappa_w^2 \leq \sigma_w^2$).

An alternative and more data-intensive method for estimating κ_w^2 is to use a machine learning model in combination with historical data. Note that κ_w^2 is in fact the Bayes mean-squared error (MSE) for the prediction problem of estimating $\mathbb{E}[Y_w|X]$.[4] So if there are existing data for (X, Y_w), the test-set MSE for a machine learning model trained using those data is a consistent estimator for an upper bound on the Bayes MSE because the Bayes MSE is by definition the MSE of the best possible model.[5] This process would also produce a usable upper bound even if only a subset of the planned covariates were available in the historical data because $\mathbb{V}[Y|X,Z] \leq \mathbb{V}[Y|X]$.

Like σ_1^2 and κ_1^2, the value γ depends on the behavior of the treatment arm and so it is not estimable using

[4] By the law of total expectation, $\mathbb{E}\left[(Y_w - \mu_w(X))^2\right] = \mathbb{E}\left[\mathbb{E}\left[(Y_w - \mu_w(X))^2|X\right]\right]$. The internal expectation is the definition of the conditional variance, i.e. $\mathbb{E}\left[(Y_w - \mu_w(X))^2|X\right] \equiv \mathbb{V}[Y_w|X]$.

[5] Note that it is always the *MSE* that is the quantity of interest, even if the outcome is binary, etc. In the binary outcome case the model's prediction (an estimate of $\mathbb{E}[Y_w|X]$) represents the probability of the outcome. MSE in this case is equivalent to the "Brier score".

historical control-arm or standard-of-care data alone. However, there are interpretable domain assumptions that bound γ. For example, in many cases it is assumed that the effect of treatment is additively constant across the population (i.e. $\mu_1(X) = \mu_0(X) + \tau$). When that is the case, $\gamma = 1$. It is unreasonable to assume this relationship would obtain exactly, but we recommend $\gamma > 0$ as a conservative lower bound that provides substantial wiggle room for treatment effect heterogeneity. Indeed, Cauchy-Schwarz and a little algebra[6] show that $\mathbb{V}[\mu_0(X)] \geq \mathbb{V}[\mu_1(X) - \mu_0(X)]$ or $\mathbb{V}[\mu_1(X)] \geq \mathbb{V}[\mu_1(X) - \mu_0(X)]$ are sufficient conditions to ensure $\gamma > 0$. So as long as the heterogeneity of (additive) effect is less than the heterogeneity in expected outcomes in either treatment arm, $\gamma \geq 0$ will certainly hold. In the case where a substantial amount of treatment arm data is available, it may be possible to estimate γ by calculating the empirical correlation coefficient between the out-of-sample predictions of estimated treatment-arm and control-arm models. At a minimum, an estimate of this kind could be used to confirm $\gamma \geq 0$, which would justify the conservative assumption that $\gamma = 0$.

4 Demonstration

Here we demonstrate via simulations and a case study that our approach for prospective sample size estimation results in trials that attain their design power with fewer subjects that would be required with existing methods. We focus on the use-case where there are no available treatment arm data before the trial is run because these data are usually scant at best (e.g. come from a relatively small phase II trial) and likely insufficient to reliably estimate κ_1 and γ. On the other hand, standard-of-care data are often plentiful and can be used as a reasonable proxy for control-arm data. In practice we recommend using any existing treatment-arm data to verify that $\sigma_0 \gtrapprox \sigma_1$, $\kappa_0 \gtrapprox \kappa_1$, and $\gamma \geq 0$ so that power estimates are ensured to be conservative.

4.1 Simulation

Methods Our treatment effect of interest was the difference-in-means of a continuous outcome (thus $r'_0 = -1, r'_1 = 1$). In each simulation scenario we fixed a counterfactual data-generating process $P(Y_1, Y_0, X)$. We sampled a dataset of 10,000 observations of X, Y_0, which represented a historical sample (e.g. prior trial control arms, registry data) that we used to estimate the population parameters we needed for the sample size calculation. Our estimates of these parameters were:

$$\begin{aligned} \hat{\sigma}_0 &= \hat{\sigma}_1 = \widehat{\mathbb{V}}[Y_0] \\ \hat{\kappa}_0 &= \hat{\kappa}_1 = \text{cross validation RMSE of an ensemble model trained on } \boldsymbol{X}, \boldsymbol{Y}_0 \\ \hat{\gamma} &= 0 \end{aligned} \tag{7}$$

The ensemble model we used was a cross-validated selection between linear regression (LR), 5-nearest-neighbors (5-kNN), and gradient boosted trees (GBM; 50 depth-5 trees, all other parameters left at their scikit-learn defaults [33]). Note that kNN explicitly satisfies the mean-square consistency (eq. 3) required for AIPW to be semiparametric efficient [28]. Nested cross-validation to select between these three models was used within the outer cross validation loop used to estimate RMSE.

We used these parameter estimates to estimate the asymptotic variances of the AIPW and unadjusted estimators in a 1:1 randomized trial ($\pi_w = 1/2$) according to the formulae

$$\begin{aligned} \hat{\nu}^2_{\text{AIPW}} &= r'^2_0 \left(\frac{\pi_1}{\pi_0} \hat{\kappa}^2_0 + \hat{\sigma}^2_0 \right) + r'^2_1 \left(\frac{\pi_0}{\pi_1} \hat{\kappa}^2_1 + \hat{\sigma}^2_1 \right) - 2|r'_0 r'_1| \gamma \sqrt{(\hat{\sigma}^2_0 - \hat{\kappa}^2_0)(\hat{\sigma}^2_1 - \hat{\kappa}^2_1)} \\ \hat{\nu}^2_{\text{unadj}} &= r'^2_0 \frac{\hat{\sigma}^2_0}{\pi_0} + r'^2_1 \frac{\hat{\sigma}^2_1}{\pi_1} \end{aligned} \tag{8}$$

[6] Abbreviate $\mathbb{V}[\mu_0] = \mathbb{V}[\mu_0(X)]$ and $\mathbb{V}[\tau] = \mathbb{V}[\mu_1(X) - \mu_0(X)]$. $\mathbb{V}[\mu_0] \geq \mathbb{V}[\tau] \iff \mathbb{V}[\mu_0] \geq \sqrt{\mathbb{V}[\tau]\,\mathbb{V}[\mu_0]} \implies \mathbb{V}[\mu_0] \geq \left| \lambda \sqrt{\mathbb{V}[\tau]\,\mathbb{V}[\mu_0]} \right| \ \forall \lambda \in [-1, 1] \implies \mathbb{V}[\mu_0] \geq |\mathbb{C}[\tau, \mu_0]| \implies \mathbb{V}[\mu_0] + \mathbb{C}[\tau, \mu_0] \geq 0 \iff \mathbb{C}[\mu_0 + \tau, \mu_0] \geq 0 \iff \mathbb{C}[\mu_1, \mu_0] \geq 0 \iff \gamma \geq 0$. Replace $\mu_0(X)$ with $\mu_1(X)$ and proceed in the same way to show the equivalent for $\mu_1(X)$.

We also calculated the *true* asymptotic variance bound ν_*^2 (thm 1) by extracting the true values of σ_w, κ_w, and γ from the counterfactual distribution. Note that there is no "$\hat{\nu}^2_{\text{ANCOVA}}$" because there does not exist a parsimonious asymptotic variance formula for ANCOVA that involves only a small number of population parameters [6]. Trials analyzed with ANCOVA typically assume unadjusted estimation for sample size calculation, which is conservative.

We used the prospective variance estimates $\hat{\nu}^2_{\text{AIPW}}$ and $\hat{\nu}^2_{\text{unadj}}$ to choose sample sizes $n^\dagger_{\text{AIPW}}$, $n^\dagger_{\text{unadj}}$ for a hypothesized trial according to the formula

$$n^\dagger = \left\{ \begin{array}{l} \text{argmin}_n \; n \\ \text{s.t. } 1-\beta < \Phi\left(\Phi^{-1}(\alpha/2) + \sqrt{n}\frac{\tau}{\nu}\right) + \Phi\left(\Phi^{-1}(\alpha/2) - \sqrt{n}\frac{\tau}{\nu}\right) \end{array} \right. \tag{9}$$

plugging in $\hat{\nu}_{\text{AIPW}}$ and $\hat{\nu}_{\text{unadj}}$ for ν, respectively. $n^\dagger_{\text{unadj}}$ represents the enrollment of the trial that we would have calculated if we used the standard, unadjusted power formula, which we know to be overly conservative. $n^\dagger_{\text{AIPW}}$ represents the enrollment of the trial that we would have calculated if we used our proposed power formula implied from theorem 1 and specified that the analysis of the trial would be conducted with an AIPW estimator. We also calculated the enrollment using ν_*^2 to obtain a value $n^\dagger_*$, which represents the trial size we would need to attain design power if we were to use an "oracle" AIPW estimator with direct access to the true functions $\mu_w(X)$. It would not be possible to do this in practice, but it is useful to study the behavior of the oracle estimator because it represents an effective upper bound on performance.

The desired power level $1-\beta$ was set to 0.8 and the significance level α was set to 0.05 in all experiments. The target treatment effect τ was the true effect from the counterfactual distribution.

To calculate empirical power we then simulated 1:1 randomized trials of a variety of sizes (n). In each simulated trial, we drew data from the specified counterfactual distribution, randomly assigned treatment, and hid the unobserved potential outcomes. We then used an "oracle" AIPW estimator, a cross-fit AIPW estimator (using cross-validated selection between GBM, 5-kNN, and LR to learn $\hat{\mu}^{(-k)}(\cdot)$), a standard main-terms ANCOVA estimator with robust standard errors (HC0, [34]), and an unadjusted estimator to estimate the treatment effect and a standard error from the data. We recorded if the result was statistically significant in each case. We repeated each experiment 1000 times and calculated the power of each estimator by computing the fraction of significant results. This was repeated for each trial size n.

The objective of these simulations was to see if the power of the AIPW estimator for $n > n^\dagger_{\text{AIPW}}$ indeed exceeded the design power $1-\beta = 0.8$. Secondarily, we were also interested in the amount by which $n^\dagger_* < n^\dagger_{\text{AIPW}} < n^\dagger_{\text{unadj}}$, the overall amount by which AIPW increased power over the ANCOVA or unadjusted anayses, the amount of conservatism in $n^\dagger_{\text{AIPW}}$, and the difference in performance between the cross-fit and oracle AIPW estimators.

We also conducted simulations in which the average treatment effect was 0, in which case we interpreted the rate of significant results as the empirical type I error of each estimator. The purpose of those simulations was to confirm that all of our estimators control type I error in large samples.

Lastly, we repeated each of our experiments with the AIPW estimator using one of GBM, kNN, or LR by itself as the estimator of the conditional means. For these experiments we used the same hyperparameter settings for each learner as were used in the ensemble learner. We used the same methods as above to calculate prospective sample sizes for each of these designs. The purpose of these experiments was to assess how sensitive the AIPW estimator is to the choice of learner and, more importantly, to see whether the corresponding prospectively-calculated sample sizes (i.e. $n^\dagger_{\text{AIPW(kNN)}}$, $n^\dagger_{\text{AIPW(GBM)}}$, and $n^\dagger_{\text{AIPW(LR)}}$) resulted in well-powered trials in all cases.

Scenarios Each simulation scenario was fully specified by the choice of counterfactual distribution $P(Y_1, Y_0, X) = P(Y_1|X)P(Y_0|X)P(X)$. We chose the scenarios to reflect our real-world presumptions that $\sigma_0 \gtrapprox \sigma_1$ as well as $\kappa_0 \approx \kappa_1$ and $\gamma \geq 0$. In other words, treatment did not substantially affect the heterogeneity in outcomes and any heterogeneity in treatment effect was on the order of or smaller than the heterogeneity in expected

outcomes.[7] For convenience of plotting, we specified the distributions so that $n^{\dagger}_{\text{AIPW}}$ and $n^{\dagger}_{\text{unadj}}$ would come out as numbers on the order of 10^2.

Within these constraints we explored four different scenarios including linear and nonlinear functions for the conditional means and the presence or absence of treatment effect heterogeneity. While no set of simulations can possibly come close to spanning the full space of possible distributions, we chose these to be illustrative of important archetypal scenarios one might find in a randomized trial.

Our simulations follow the format of Schuler et al. [6]. In all cases, the distribution of covariates $P(X)$ was a 10-dimensional uniform random variable in the prism $[-1, 1]^{10}$. $P(Y_w|X)$ were of a Gaussian quadratic-mean form $\mathcal{N}(aX^\top \mathbb{1}X + bX^\top \mathbb{1} + c, 1)$ in all scenarios ($\mathbb{1}$ is a matrix or vector of 1s with appropriate shape implied). The parameter a controls the degree of non-linearity in this specification, with $a = 0$ representing the linear case. Treatment effect heterogeneity refers to the situation in which a or b is different for $P(Y_0|X)$, and $P(Y_1|X)$. The specific values of a, b, and c for each scenario are shown in Table 1.

Scenario	$P(Y_0\|X)$			$P(Y_1\|X)$		
	a_0	b_0	c_0	a_1	b_1	c_1
linear, constant	0	1	0	0	1	1/2
linear, heterogeneous	0	1	0	0	0	1/2
nonlinear, constant	1	1	0	1	1	1
nonlinear, heterogeneous	1	1	0	1	0	1

Table 1: Parameters for all simulation scenarios.

The scenarios and parameters for the type I error simulations were identical except that c_0 was modified in each scenario such that the average treatment effect became 0.

Results Our results (figure 1) show that trials analyzed with the AIPW estimator attain power greater than 80% at the corresponding enrollment targets. Since these targets were prospectively calculated, this provides empirical evidence that our methods can be applied in the design phase to power trials as long as the analysis plan specifies an AIPW (or other semiparametric efficient) estimator. Moreover the enrollment target is still likely to be conservative by some amount and robust to heterogeneity of the treatment effect and nonlinear conditional means.

The enrollment target calculated by assuming the unadjusted estimator is also robust to different conditions but is far more conservative than is necessary when any kind of covariate adjustment is employed in the analysis. Indeed, we observe potential sample size savings of $\sim 35\%$ in these simulations when using the AIPW-based enrollment target. However, one advantage of the unadjusted enrollment target is the trial will attain or exceed its design power regardless of what estimator is specified in the analysis plan, as long as it is as efficient or more efficient than the unadjusted estimator.

We also observe near-perfect agreement between the power predicted at the oracle enrollment target and the empirical power attained by the oracle estimator at that sample size. This is strong empirical evidence for the validity of theorem 1. More importantly, however, the discrepancy between the oracle enrollment target and the estimated AIPW enrollment target suggests that it is not necessarily beneficial to have access to the true population parameters (σ_w, κ_w, γ) at design time because the AIPW estimator is typically far from its optimal efficiency in practice. Trials analyzed with a real-world (not oracle) AIPW estimator will usually *not* attain their design power at the *oracle* enrollment target. In other words, there is benefit to having to *estimate* the population parameters from data because using those estimates in the sample size calculation better reflects the finite-sample performance of the AIPW estimator.

Our results also confirm what is known about the relative efficiency of AIPW, ANCOVA, and unadjusted estimation. Main-terms ANCOVA performs as well as the oracle in the scenario with linear conditional means and a constant effect because it is perfectly specified in that setting. Our AIPW estimator is also near-optimal in that case because it includes a linear regression in its ensemble model of the conditional

[7] Without these conditions, it is not possible to conservatively use historical control data to estimate the necessary parameters for either our power formula or the unadjusted power formula.

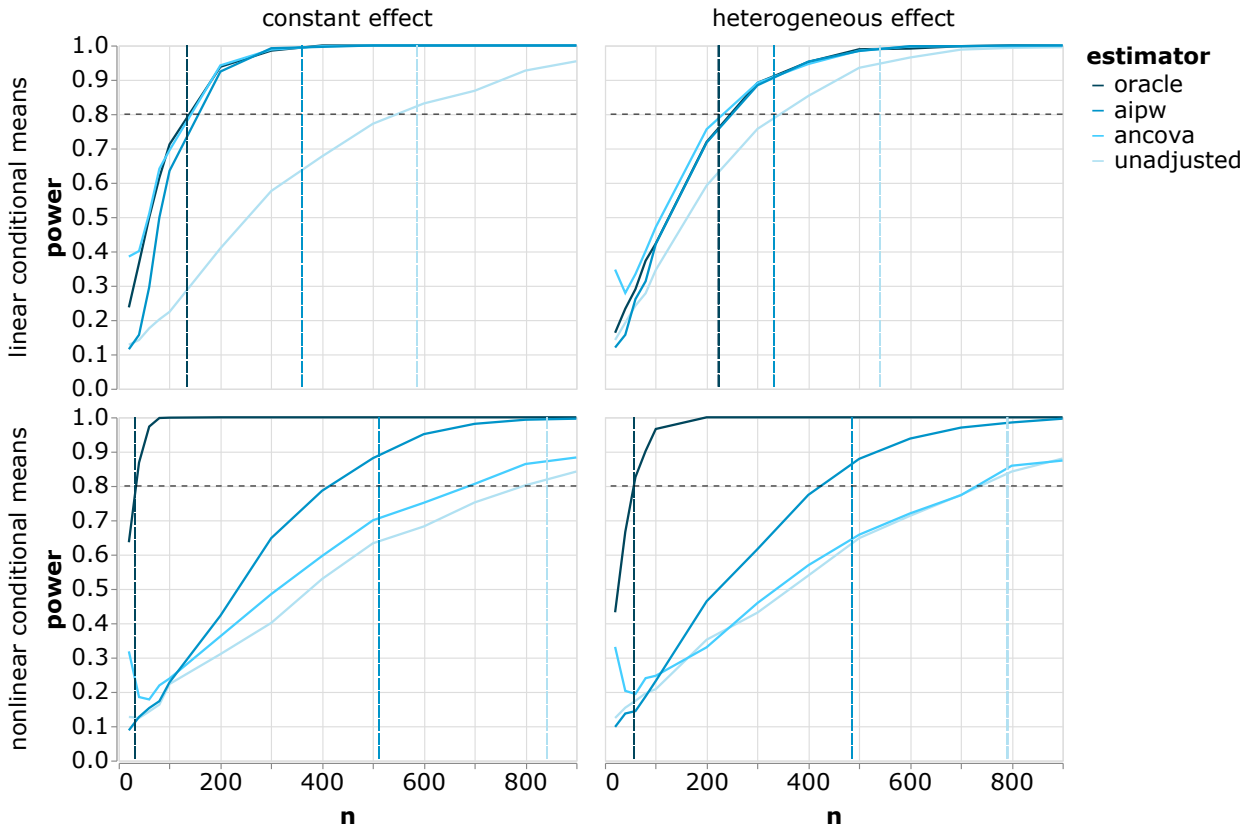

Figure 1: Empirical power and prospectively-calculated enrollment targets. The dotted vertical lines correspond to enrollment targets, which were prospectively calculated using parameters estimated from a "historical" dataset. The lightest blue corresponds to the enrollment required for 80% power as dictated by the unadjusted estimator, darker blue corresponds to the AIPW estimator, and darkest to the oracle AIPW estimator. The curves indicate the true empirical power of each estimator (including ANCOVA for comparison) as the enrollment is varied. The horizontal dotted line represents the 80% target design power. A cross-validated ensemble of GBM, 5-kNN, and linear regression was used for the AIPW estimator and to prospectively calculate the corresponding sample size.

means. On the other hand, AIPW greatly exceeds ANCOVA in power when there is any nonlinearity. The efficiency gain over unadjusted estimation is relatively lessened here, likely because our kNN estimation of the conditional mean functions converges much more slowly than the linear regression does in the linear case. In these settings, there is also a large difference between the performance of the oracle and real-world AIPW estimators. This suggests that there is a substantial opportunity to improve the quality of the conditional mean modeling in the AIPW estimator (e.g. with stronger machine learning methods or via models pre-trained on large external datasets and fine-tuned on the trial data).

The results of our simulations in scenarios with no treatment effect (figure 2) confirm that the AIPW estimator controls type I error in large samples [19]. As expected due to the asymptotic nature of the inference, we observed some inflation of type I error for all estimators in very small samples ($n \lesssim 100$).[8]

Our experiment varying the learner used to estimate the conditional means for the AIPW estimator shows that our procedure is largely robust to what machine learning method is chosen (figure 3). However, the experiments with GBM produced trials that in some cases were slightly under-powered, which could be of concern. Because of this, we recommend using a diverse cross-validated (or stacked) ensemble of learning algorithms with varied tuning parameters.

[8]The small-sample behavior of ANCOVA with robust standard errors has been the subject of extensive prior work [34].

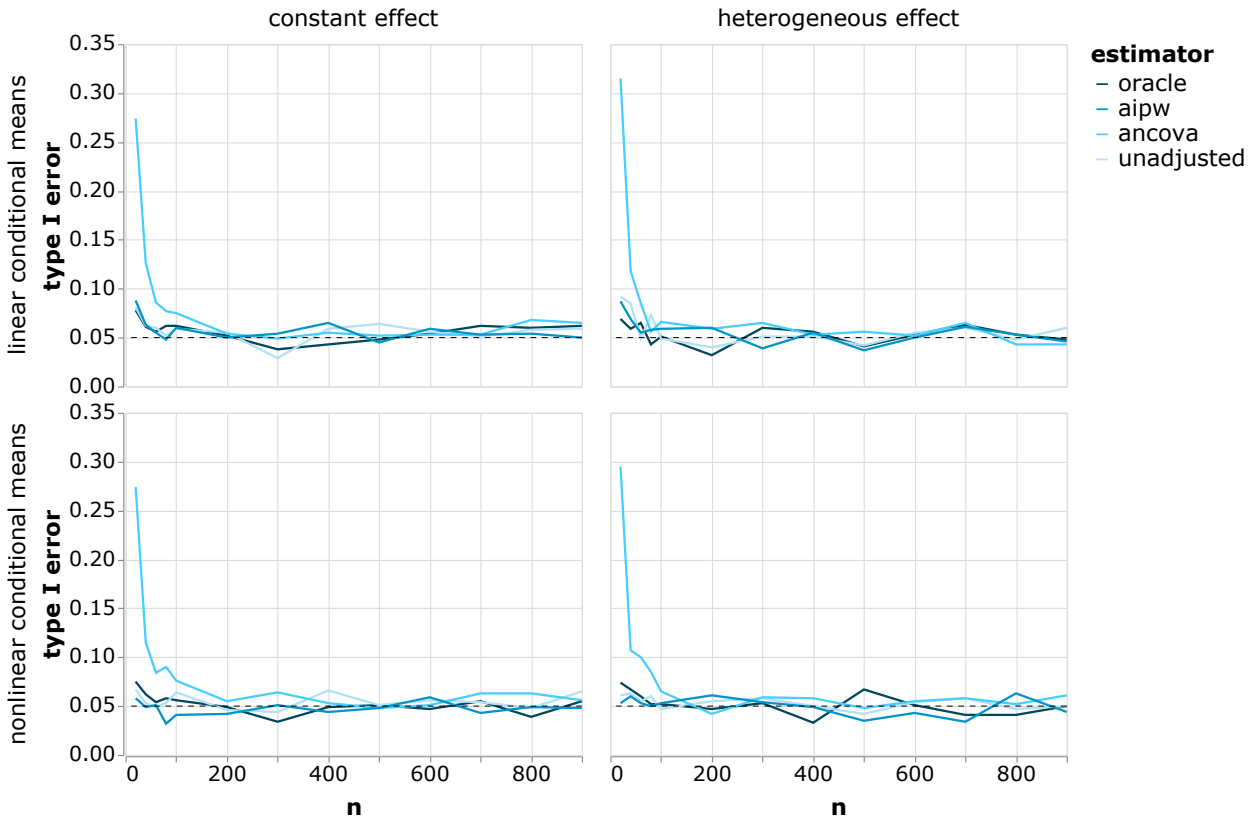

Figure 2: Empirical type I error of the orracle, AIPW, ANCOVA, and unadjusted estimators in our simulations. The horizontal dotted line represents the 5% target type I error rate.

Surprisingly, prospective power estimation worked well when linear regression was used to estimate the conditional means for AIPW. In this case the mean-square consistency condition (eq. 3) is clearly not satisfied so theorem 1 may not hold; thus it may not be prudent to rely on this empirical result. Using linear regression in the AIPW estimator gave less power than other learners when there was any nonlinearity in the conditional means. Note that including linear regression in an ensemble model ensemble effectively guarantees that AIPW will exceed the power of ANCOVA.

4.2 Case Study

In addition to the simulations presented above, we re-analyzed data from an existing trial to demonstrate how our semiparametric efficient estimation can be taken into account when prospectively powering trials. Much of the subsequent description is shared with Schuler et al. [6].

Trial Our demonstration trial, reported by Quinn et al. [35], was conducted to determine if docosahexaenoic acid (DHA) supplementation slows cognitive and functional decline for individuals with mild to moderate Alzheimer's disease. The trial was performed through the Alzheimer's Disease Cooperative Study (ADCS), a consortium of academic medical centers and private Alzheimer disease clinics funded by the National Institute on Aging to conduct clinical trials on Alzheimer disease.

Quinn et al. [35] randomized 238 subjects to a treatment arm given DHA and 164 subjects to a control arm given placebo. This trial measured a number of covariates at baseline including demographics and patient characteristics (e.g. sex, age, region, weight), lab tests (e.g. blood pressure, ApoE4 status [36]), and component scores of cognitive tests. A full list of the 37 covariates we used is reported in Schuler et al. [6].

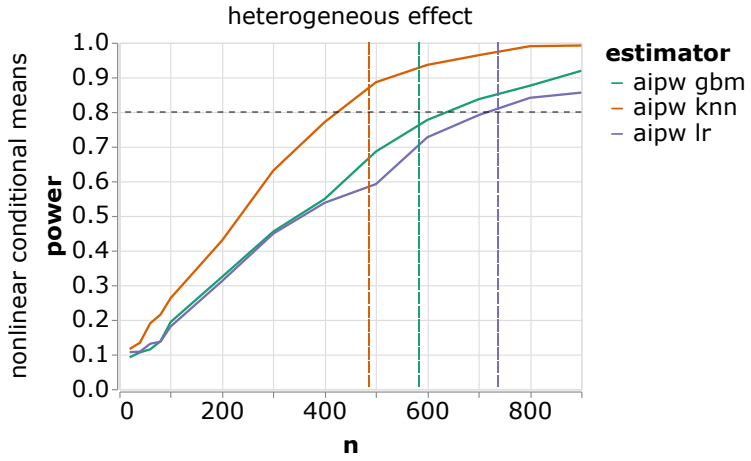

Figure 3: Empirical power and prospectively-calculated enrollment targets for AIPW estimators with different learners used to estimate the conditional means. Colors differentiate the learners (gbm: gradient boosted trees, knn: k nearest neighbors, lr: linear regression) with hyperparameters as described above. Other visual elements are as in figure 1. Additional simulation scenarios are shown in the appendix.

The primary outcome of interest for our reanalysis was the increase in the Alzheimer's Disease Assessment Scale - Cognitive Subscale (ADAS-Cog 11, a quantitative measure of cognitive ability) [37] over the duration of the trial (18 months).

Methods Before examining the trial data, we estimated the population parameters required for a sample size calculation from a historical dataset comprised of 6,919 early-stage Alzheimer's patients. These data came from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Critical Path for Alzheimer's Disease (CPAD) database [38, 39], and included measurements of ADAS-Cog 11 at 6-month, or more frequent, intervals post-baseline. The ADNI dataset is made up of longitudinal data from 4 sequential large observational studies in Alzheimer's disease, while the CPAD dataset is made up of control arm data from 29 Alzheimer's disease randomized trials. These data also included the same baseline covariates as were measured in the DHA trial (imputed to a column mean where missing).

Once these data were prepared, we used the procedure detailed in eq. 7 to estimate values for σ_w, κ_w, and γ. We set π_w to the observed treatment ratios in the trial. We then used these values to calculate the sample sizes that would be required to detect a clinically meaningful target treatment effect of 2.7 points with power $> 80\%$ given a significance level of 0.05, assuming either an unadjusted analysis ($n^\dagger_{\text{unadj}}$) or an efficient analysis ($n^\dagger_{\text{AIPW}}$) according to the procedure described in eqs. 8 and 9. This was done "prospectively", i.e. before we performed any analysis of the trial data.

We then sought to emulate what *would* have happened had we executed the trial with either a) the smaller enrollment ($n^\dagger_{\text{AIPW}}$) and AIPW in the analysis plan or b) the larger enrollment ($n^\dagger_{\text{unadj}}$) and unadjusted estimation in the analysis plan. We consider these to be two different "trial designs". As in our simulations, we pre-specified the AIPW estimator to use a cross-validated selection between 5-kNN, GBM, and linear regression in the estimation of conditional means.

The idea of this case study was to see whether the final inference about the treatment effect in both designs turned out to be similar, despite the different number of subjects. To emulate running these two trials side-by-side, in each case we took a bootstrap sample of $n^\dagger$ subjects with observed outcomes from the full trial dataset (in the original randomization ratio) and then analyzed that sample of data with the specified estimator. We repeated this process 500 times for each design to average out the resampling variability. We report the mean estimated treatment effect and the mean estimated variance from executing each trial design.

Results Our sample size calculation assuming unadjusted estimation yielded a required enrollment of $n^{\dagger}_{\text{unadj}} = 272$ (115 controls, 157 treated). Assuming an efficient estimator we obtained $n^{\dagger}_{\text{AIPW}} = 243$ (103 controls, 140 treated). Sample size calculation leveraging efficient estimation therefore yielded a sample size savings of $\sim 10\%$ in this example.

The mean effect estimate in the unadjusted design was -0.140 with a mean variance of 1.047. The mean effect estimate in the AIPW design was 0.002 with a mean variance of 1.059. The two estimates are qualitatively identical (null effect). Moreover, the variance of both estimates are nearly identical despite the fact that the AIPW design used 10% fewer subjects. Although we cannot estimate empirical power in this case because we do not know the true treatment effect and cannot repeat the trial, the similarity of the estimated variances implies that both designs are likely to have about equal power.

5 Discussion

Our theory, simulations, and case study all demonstrate how efficient estimation can be exploited to design smaller trials that attain their design power without sacrificing type I error control. The required sample size calculation is parsimonious and requires the analyst to provide only a small number of population parameters. Under mild assumptions, these parameters are all estimable from historical data with off-the-shelf techniques. The resulting sample size is typically much smaller than would otherwise be required and often still provides a margin of safety. This margin can always be increased by conservative estimation of the required population parameters.

Although our simulations and case study focused on the case of a continuous outcome, our theory shows that theorem 1 applies equally well in, e.g., a trial with a binary outcome and an odds ratio estimand. Moreover, although our simulations and case study focused only on AIPW, our theory holds for any semi-parametric efficient estimator in the context of randomized trial data (e.g. TMLE).

Also, for simplicity of presentation, we avoided discussion of missing covariate data, dropout, intercurrent events, and other complicating factors in trial design. For example, consider a trial where the outcome is accurately measured, but less attention is paid to the covariates. Any missing covariate values would need to be imputed (e.g. using column means) before adjusted analysis. Imputation of covariates does not affect type I error control, but the average conditional variances κ_w estimated from historical data with less missingness may be too small in this case and the power may be overestimated. These phenomena were not modeled in our simulations or case study. Indeed, there are no general-purpose solutions to problems like these, many of which also affect standard trial designs. In practice, statisticians should use heuristic methods (e.g. artificially censoring historical covariate data) and/or conservative assumptions (e.g. multipliers on estimated parameters) as necessary to counteract the effects of these imperfections on sample size estimation.

In addition to our power formula, we contribute further evidence that cross-fit AIPW in particular is a safe and effective estimator in the context of randomized trials. Our results show that it can easily meet or exceed the power of ANCOVA and even appears to have better control of type I error in small samples. Although AIPW makes use of innovations in "black-box" machine learning, it is no harder to interpret the estimated treatment effect. If anything, inference becomes clearer because it provides more conclusive evidence. This benefit comes without any meaningful cost because AIPW estimation requires no additional assumptions to ensure type I error control in randomized trials.

Moreover, AIPW with black-box machine learning is still perfectly compatible with pre-specification. In order to pre-specify an analysis with AIPW, one must clearly lay out the set of learners (e.g. kNN, linear regression) and hyperparameters (e.g. number of neighbors, regularization strength) that will be considered and how their predictions will be selected among, combined, etc. (e.g. selection via cross-validation, super-learning [40, 41]). All of these choices can and should be made at design time, as we demonstrated in our case study. The only requirement for theorem 1 to hold for the AIPW estimator is that the learner (or ensemble) be mean-square consistent in estimating the conditional mean functions. As such, it is generally not appropriate to use (only) linear models for this purpose. However, the analyst can rest assured that the AIPW estimator will control type I error in large samples regardless of the learner as long as it is pre-specified. Moreover, our empirical evidence suggests that trials may be well-powered even if the assumptions

of theorem 1 are violated.

Our results also suggest that further sample size reductions are possible. In cases where the conditional mean functions were nonlinear, the oracle AIPW estimator outperformed the real-world AIPW estimator by a large margin. The only difference between these two estimators is that the true conditional means are known to the oracle instead of estimated. This implies that better conditional mean models could yield large improvements in power or be translated to larger sample size reductions. Our experiments used an ensemble of 5-kNN, GBM, and linear regression for simplicity and speed, but in practice a larger number of learners with greater diversity could be leveraged so that the learned models approach the true conditional means with as little data as possible. Conversely, historical data might be combined with the trial data to facilitate learning of these models. Since the AIPW estimator is unbiased regardless of the model used, including historical data in this fashion would not sacrifice type I error control. It would also be possible to leverage pre-training and transfer learning methods if deep learning were used to model the conditional means [42, 43].

The divergence between the oracle and real-world AIPW estimator performance also highlights a crucial consideration for estimating population parameters for the sample size calculation. If the true population parameters were actually known, it would not behoove the analyst to use them in the power calculation because the real-world AIPW estimator would be unlikely to attain design power at the enrollment target specified with the oracle parameters. In other words, having to estimate the required population parameters from external data is actually helpful because the estimate of the average conditional variance ($\hat{\kappa}_0^2$) is usually closer to the amount of variance that is still "unexplainable" with a real-world model than the true average conditional variance (κ_w^2) is. Indeed, our experiments with isolated GBM models were in some cases slightly under-powered, which suggests that our powering method can be sensitive to the learning curve (RMSE vs. n) of the chosen algorithm. Using a large, diverse ensemble is therefore recommended. This may also have implications for the amount of historical data used to estimate κ_w relative to the amount of data used for learning in the trial, although in our experiments we used an order of magnitude more historical than trial data and still obtained conservative sample sizes with AIPW. Pre-training or combining historical and trial data for estimation of the conditional mean models is also an attractive avenue to address this concern in future work.

Data Availability

Certain data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

Certain data used in the preparation of this article were obtained from the Critical Path for Alzheimer's Disease (CPAD) database. In 2008, Critical Path Institute, in collaboration with the Engelberg Center for Health Care Reform at the Brookings Institution, formed the Coalition Against Major Diseases (CAMD), which was then renamed to CPAD in 2018. The Coalition brings together patient groups, biopharmaceutical companies, and scientists from academia, the U.S. Food and Drug Administration (FDA), the European Medicines Agency (EMA), the National Institute of Neurological Disorders and Stroke (NINDS), and the National Institute on Aging (NIA). CPAD currently includes over 200 scientists, drug development and regulatory agency professionals, from member and non-member organizations. The data available in the CPAD database has been volunteered by CPAD member companies and non-member organizations.

Certain data used in the preparation of this article were obtained from the University of California, San Diego Alzheimer's Disease Cooperative Study Legacy database.

Acknowledgments

The author thanks Xinkun Nie, Charles Fisher, and David Miller for helpful conversations.

Data collection and sharing for this project was funded in part by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Data collection and sharing for this project was funded in part by the University of California, San Diego Alzheimer's Disease Cooperative Study (ADCS) (National Institute on Aging Grant Number U19AG010483).

References

[1] George Maldonado and Sander Greenland. Estimating causal effects, 2002. URL https://scholar.harvard.edu/files/cwinship/files/estimating_causal.pdf.

[2] Harold C. Sox and Steven N. Goodman. The Methods of Comparative Effectiveness Research. *Public Health*, 33(1):425–445, 2012. ISSN 0163-7525. doi: 10.1146/annurev-publhealth-031811-124610. URL http://www.annualreviews.org/doi/abs/10.1146/annurev-publhealth-031811-124610.

[3] J. Marc Overhage, Patrick B. Ryan, Martijn J. Schuemie, and Paul E. Stang. Desideratum for Evidence Based Epidemiology. *Drug Safety*, 36(Suppl 1):5–14, 10 2013. ISSN 0114-5916. doi: 10.1007/s40264-013-0102-2. URL https://content.ebscohost.com/ContentServer.asp?T=P&P=AN&K=91672951&S=R&D=aph&EbscoContent=dGJyMNLr40Sep7E40dvuOLCmr06ep7VSsq24TbCWxWXS&ContentCustomer=dGJyMPGqsVGyrbdJuePfgeyx44Dt6fIA.

[4] Edward L. Hannan. Randomized Clinical Trials and Observational Studies Guidelines for Assessing Respective Strengths and Limitations. *JACC: Cardiovascular Interventions*, 1(3):211–217, 06 2008. ISSN 1936-8798. doi: 10.1016/j.jcin.2008.01.008.

[5] S R Jones, S Carley, and M Harrison. An introduction to power and sample size estimation. *Emergency Medicine Journal*, 20(5):453, 2003. ISSN 1472-0205. doi: 10.1136/emj.20.5.453.

[6] Alejandro Schuler, David Walsh, Diana Hall, Jon Walsh, and Charles Fisher. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. *arXiv*, 2020.

[7] William D. Dupont and Walton D. Plummer. Power and sample size calculations A review and computer program. *Controlled Clinical Trials*, 11(2):116–128, 1990. ISSN 0197-2456. doi: 10.1016/0197-2456(90)90005-m.

[8] Selene Leon, Anastasios A Tsiatis, and Marie Davidian. Semiparametric Estimation of Treatment Effect in a Pretest-Posttest Study. *Biometrics*, 59(4):1046–1055, 2003. ISSN 0006-341X. doi: 10.1111/j.0006-341x.2003.00120.x.

[9] Elizabeth L Turner, Pablo Perel, Tim Clayton, Phil Edwards, Adrian V Hernández, Ian Roberts, Haleema Shakur, Ewout W Steyerberg, and CRASH trial collaborators. Covariate adjustment increased power in randomized controlled trials: an example in traumatic brain injury. *Journal of clinical epidemiology*, 65(5):474–81, 2011. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2011.08.012.

[10] Alice S Whittemore. Sample Size for Logistic Regression with Small Response Probability. *Journal of the American Statistical Association*, 76(373):27–32, 1981. ISSN 0162-1459. doi: 10.1080/01621459.1981.10477597.

[11] Eugene Demidenko. Sample size determination for logistic regression revisited. *Statistics in Medicine*, 26(18):3385–3397, 2007. ISSN 1097-0258. doi: 10.1002/sim.2771.

[12] F. Y. Hsieh, Daniel A. Bloch, and Michael D. Larsen. A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine*, 17(14):1623–1634, 1998. ISSN 1097-0258. doi: 10.1002/(sici)1097-0258(19980730)17:14⟨1623::aid-sim871⟩3.0.co;2-s.

[13] DAVID F SIGNORINI. Sample size for Poisson regression. *Biometrika*, 78(2):446–450, 1991. ISSN 0006-3444. doi: 10.1093/biomet/78.2.446.

[14] A Tsiatis. *Semiparametric theory and missing data*. 2007.

[15] Michael C Knaus. Double Machine Learning based Program Evaluation under Unconfoundedness. *arXiv*, 2020.

[16] Jui-Chung Yang, Hui-Ching Chuang, and Chung-Ming Kuan. Double machine learning with gradient boosting and its application to the Big N audit quality effect. *Journal of Econometrics*, 216(1):268–283, 2020. ISSN 0304-4076. doi: 10.1016/j.jeconom.2020.01.018.

[17] K. L. Moore and M. J. van der Laan. Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in Medicine*, 28(1):39–64, 2009. ISSN 1097-0258. doi: 10.1002/sim.3445.

[18] Zhiwei Zhang and Shujie Ma. Machine learning methods for leveraging baseline covariate information to improve the efficiency of clinical trials. *Statistics in Medicine*, 38(10):1703–1714, 2019. ISSN 0277-6715. doi: 10.1002/sim.8054.

[19] Paul N Zivich and Alexander Breskin. Machine Learning for Causal Inference: On the Use of Cross-fit Estimators. *Epidemiology*, 32(3):393–401, 2021. ISSN 1044-3983. doi: 10.1097/ede.0000000000001332.

[20] Christoph Rothe. Flexible Covariate Adjustments in Randomized Experiments.

[21] Stefan Wager, Wenfei Du, Jonathan Taylor, and Robert J Tibshirani. High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(45):12673 12678, 10 2016. ISSN 0027-8424. doi: 10.1073/pnas.1614732113.

[22] Donald B Rubin. Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005. ISSN 0162-1459. doi: 10.1198/016214504000001880. URL http://amstat.tandfonline.com/doi/abs/10.1198/016214504000001880.

[23] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1 C68, 2018. doi: 10.1111/ectj.12097.

[24] Mark J van der Laan and Sherri Rose. *Targeted Learning*. Springer Series in Statistics. Springer New York, 2011. ISBN 978-1-4419-9781-4. doi: 10.1007/978-1-4419-9782-1.

[25] Iván Díaz. Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*, 21(2):353–358, 2019. ISSN 1465-4644. doi: 10.1093/biostatistics/kxz042.

[26] Whitney K Newey and James R Robins. Cross-Fitting and Fast Remainder Rates for Semiparametric Estimation. *arXiv.org*, math.ST, 01 2018. URL arXiv.org.

[27] Stefan Wager. Course notes, Stanford Stats 361. URL https://web.stanford.edu/\~swager/stats361.pdf.

[28] Philip E Cheng. Strong consistency of nearest neighbor regression function estimators. *Journal of Multivariate Analysis*, 15(1):63–72, 1984.

[29] Ye Luo and Martin Spindler. High-Dimensional L_2Boosting: Rate of Convergence. *arXiv*, 2016.

[30] Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep Neural Networks for Estimation and Inference. *arXiv*, 2018.

[31] Vasilis Syrgkanis and Manolis Zampetakis. Estimation and Inference with Trees and Forests in High Dimensions. *arXiv*, 2020.

[32] Michael Rosenblum and Mark J van der Laan. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *The international journal of biostatistics*, 6(1):Article 13, 2010. doi: 10.2202/1557-4679.1138.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

[34] J. Scott Long and Laurie H. Ervin. Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. *The American Statistician*, 54(3):217–224, 2012. ISSN 0003-1305. doi: 10.1080/00031305.2000.10474549.

[35] Joseph F. Quinn, Rema Raman, Ronald G. Thomas, Karin Yurko-Mauro, Edward B. Nelson, Christopher Van Dyck, James E. Galvin, Jennifer Emond, Clifford R. Jack, Michael Weiner, Lynne Shinto, and Paul S. Aisen. Docosahexaenoic Acid Supplementation and Cognitive Decline in Alzheimer Disease: A Randomized Trial. *JAMA*, 304(17):1903–1911, 2010. ISSN 0098-7484. doi: 10.1001/jama.2010.1510.

[36] Keith D Coon, Amanda J Myers, David W Craig, Jennifer A Webster, John V Pearson, Diane Hu Lince, Victoria L Zismann, Thomas G Beach, Doris Leung, Leslie Bryden, Rebecca F Halperin, Lauren Marlowe, Mona Kaleem, Douglas G Walker, Rivka Ravid, Christopher B Heward, Joseph Rogers, Andreas Papassotiropoulos, Eric M Reiman, John Hardy, and Dietrich A Stephan. A High-Density Whole-Genome Association Study Reveals That APOE Is the Major Susceptibility Gene for Sporadic Late-Onset Alzheimer's Disease. *The Journal of Clinical Psychiatry*, 68(04):613–618, 2007. ISSN 0160-6689. doi: 10.4088/jcp.v68n0419.

[37] Wilma G. Rosen, Richard C. Mohs, and Kenneth L. Davis. A New Rating Scale for Alzheimer's Disease. *American Journal of Psychiatry*, 1984.

[38] Jon Neville, Steve Kopko, Steve Broadbent, Enrique Avilés, Robert Stafford, Christine M. Solinsky, Lisa J. Bain, Martin Cisneroz, Klaus Romero, Diane Stephenson, and Coalition Against Major Diseases. Development of a unified clinical trial database for Alzheimer's disease. *Alzheimer's & Dementia*, 11 (10):1212–1221, 2015. ISSN 1552-5260. doi: 10.1016/j.jalz.2014.11.005.

[39] K Romero, M Mars, D Frank, M Anthony, J Neville, L Kirby, K Smith, and R L Woosley. The Coalition Against Major Diseases: Developing Tools for an Integrated Drug Development Process for Alzheimer's and Parkinson's Diseases. *Clinical Pharmacology & Therapeutics*, 86(4):365–367, 2009. ISSN 1532-6535. doi: 10.1038/clpt.2009.165.

[40] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 7. Model Assessment and Selection. The Elements of Statistical Learning, page 1 42. Springer New York, 2009. ISBN 978-0-387-84857-0.

[41] Mark J. van der Laan, Eric C Polley, and Alan E. Hubbard. Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):Article25, 2007. ISSN 2194-6302. doi: 10.2202/1544-6115.1309.

[42] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015. ISSN 1476-4687. doi: 10.1038/nature14539.

[43] Sebastien Dubois, Nathanael Romano, Kenneth Jung, Nigam Shah, and David Kale. The Effectiveness of Transfer Learning in Electronic Health Records Data. *Workshop Track - ICLR*, 2017.

A Mathematical Results

A.1 Details of the cross-fit AIPW estimator in randomized trials

Here we discuss the AIPW estimator and its semiparametric efficiency in the context of randomized trials. While the conclusions of this paper apply to any semiparametric efficient estimator, it may help the reader to understand semiparametric efficiency in the context of a single estimator. The details provided here are all available elsewhere in the literature or follow immediately from known results [14, 27, 20, 32]. We reframe them here to provide a quick reference and starting point for further reading.

Preliminaries A scalar *parameter* of a distribution is a functional that ingests the distribution and returns a number. For example, consider the distribution of a scalar random variable Y defined by the PDF f_y. The mean of Y (which is a parameter) is the functional $\int y f_y(y) dy$. An *estimator* of a (scalar) parameter is a function that takes data sampled from the distribution in question and returns a number which is meant to approximate the parameter. For example, an estimate of the mean of Y is $\frac{1}{n}\sum_i y_i$ when y_i are assumed to be draws from the distribution of Y. An estimator is *consistent* if it recovers the true value of the parameter as the sample size grows. It is usually possible to construct many different consistent estimators of the same parameter, so we are interested in finding the one that has the smallest possible sampling variance. We call this the *efficient* estimator.

It becomes much easier to find the efficient estimator if we restrict our attention to the class of *regular* and *asymptotically linear* (RAL) estimators, and we actually don't lose anything by doing so. It is not important to understand the precise mathematical definition of regularity, but heuristically, a regular estimator does not have anomalously bad performance for certain special values of the parameter. Thus restricting ourselves to regular estimators is a good and sensible thing to do. Effectively all estimators which can be used in practice are regular. Moreover, among regular estimators of some parameter, a result called the Hájek-Le Cam convolution theorem guarantees that the estimator with the smallest asymptotic variance is asymptotically linear, so we lose nothing by further restricting ourselves to asymptotically linear estimators once we've excluded irregular estimators [14].

The definition of asymptotic linearity for an estimator $\hat{\psi}$ of a parameter ψ is that there exists an *influence function* ϕ such that $\sqrt{n}(\hat{\psi}-\psi) = \widehat{\mathbb{E}}[\phi]+o_p(1)$. In other words, the estimator $\hat{\psi}$ behaves like an IID average of some random variable ϕ in large samples. Asymptotic linearity immediately implies $\sqrt{n}(\hat{\psi}-\psi) \rightsquigarrow N(0, \mathbb{V}[\phi])$ by the central limit theorem. Therefore the asymptotic variance of any asymptotically linear estimator is given by the variance of its influence function.

Finding the most efficient RAL estimator thus boils down to finding the influence function with the smallest variance, which we call the *efficient influence function* (EIF). The EIF defines a lower bound on the achievable sampling variance when estimating a parameter. It so happens that it is often possible to characterize the space of all influence functions of RAL estimators of a given parameter in a generic generative model, which makes it possible to derive the EIF. Obtaining an efficient estimator is therefore a matter of deriving the EIF for the class of RAL estimators of a parameter and then constructing an estimator that has that influence function.

Application In the semiparametric statistical model $P(Y,W,X) = P(Y|X)P(W|X)P(X)$ (with $P(Y|X)$, $P(W|X)$, and $P(X)$ free to be any distributions that satisfy mild regularity conditions), it turns out that the efficient influence function of the class of RAL estimators for the parameter $\mu_w = \mathbb{E}[Y|W=w]$ is $\phi_w = \frac{W_w}{\pi_w(X)}(Y-\mu_w(X)) + (\mu_w(X)-\mu_w)$. Proving this is not simple (see Tsiatis [14]), but after the fact has been established all our subsequent theory requires only algebra and elementary tools from large-sample theory.

As might be expected, the EIF for the two-dimensional parameter $\mu = [\mu_0, \mu_1]^\top$ is $\phi_\mu = [\phi_0, \phi_1]^\top$. Influence functions obey a "chain rule" such that if the EIF of ψ is ϕ, the EIF of $g(\psi)$ is $\nabla g^\top \psi$. For us, that means that the EIF of $\tau = r(\mu_0, \mu_1)$ is $\phi = r'_0(\mu_0,\mu_1)\phi_0 + r'_1(\mu_0,\mu_1)\phi_1$.

In a two arm randomized trial with treatment fractions π_w, one estimator of μ_w that has the efficient

influence function is $\hat{\mu}_w^* = \widehat{\mathbb{E}}\left[\frac{W_w}{\pi_w}(Y - \mu_w(X)) + \mu_w(X)\right]$. The fact is easily verified by showing $\sqrt{n}(\hat{\mu}_w - \mu_w) = \widehat{\mathbb{E}}\left[\phi_w\right] + o_p(1)$ with an application of the central limit theorem. An application of the delta method shows that $\hat{\tau}^* = r(\hat{\mu}_0^*, \hat{\mu}_1^*)$ attains the asymptotic variance $\nu^2 = \mathbb{V}\left[\phi\right]$ and is therefore efficient. We'll call this the "oracle estimator".

Unfortunately, the oracle estimator $\hat{\tau}^*$ is infeasible in practice because the conditional means $\mu_w(X)$ are not known. However, it turns out that estimates can be substituted without sacrificing the optimality properties. Let $\hat{\mu}_w^{(k)} = \widehat{\mathbb{E}}\left[\frac{W_w}{\pi_w}(Y - \hat{\mu}_w^{(-k)}(X)) + \hat{\mu}_w^{(-k)}(X) \mid k(i) = k\right]$. This is the marginal mean estimated from the kth fold of data using the conditional mean function *estimated* from the rest of the data. Let $\hat{\mu}_w^{*(k)}$ be the oracle equivalent that uses the *true* conditional mean $\mu_w(X)$. Clearly, $\hat{\mu}_w = \sum \frac{n^{(k)}}{n}\hat{\mu}_w^{(k)}$ where $n^{(k)}$ is the number of observations in fold k (and similarly for $\hat{\mu}_w$). If we can show that $\sqrt{n}(\hat{\mu}_w^{(k)} - \hat{\mu}_w^{*(k)}) \xrightarrow{p} 0$, then Slutsky's theorem and the delta method imply that $\hat{\tau}$ has the same asymptotic properties as $\hat{\tau}^*$, i.e. $\sqrt{n}(\hat{\tau} - \tau) \rightsquigarrow N(0, \nu^2)$. In other words, since the oracle estimator is efficient with a known asymptotic variance, the feasible estimator is also efficient and has the same asymptotic variance. It turns out that if we assume that $\mathbb{E}\left[\left(\hat{\mu}_w^{(-k)}(X) - \mu_w(X)\right)^2\right] \to 0$ (a very weak condition), then it is possible to show $\sqrt{n}(\hat{\mu}_w^{(k)} - \hat{\mu}_w^{*(k)}) \xrightarrow{p} 0$ as desired.

We begin by deriving an expression for the difference between the oracle and feasible estimators:

$$\begin{aligned}
\hat{\mu}_w^{(k)} - \hat{\mu}_w^{*(k)} &= \widehat{\mathbb{E}}\left[\left(\frac{W_w}{\pi_w}(Y - \hat{\mu}_w^{(-k)}(X)) + \hat{\mu}_w^{(-k)}(X)\right) - \left(\frac{W_w}{\pi_w}(Y - \mu_w(X)) + \mu_w(X)\right) \;\middle|\; k(i) = k\right] \\
&= \widehat{\mathbb{E}}\left[\left(1 - \frac{W_w}{\pi_w}\right)\left(\hat{\mu}_w^{(-k)}(X) - \mu_w(X)\right) \;\middle|\; k(i) = k\right] \\
&= \frac{1}{n^{(k)}}\sum_{i \in \mathcal{I}_k}^{n^{(k)}}\left[\left(1 - \frac{W_{w,i}}{\pi_w}\right)\left(\hat{\mu}_w^{(-k)}(X_i) - \mu_w(X_i)\right)\right]
\end{aligned} \tag{10}$$

where in the last line all we've done is expand the $\widehat{\mathbb{E}}\left[\cdot\right]$ notation and introduce $\mathcal{I}_k = \{i : k(i) = k\}$. Notice that if we condition on the dataset $\mathcal{I}_{(-k)}$, the estimated function $\hat{\mu}_w^{(-k)}(\cdot)$ becomes fixed because it does not depend on the data in fold k. Therefore $\mathbb{E}\left[\hat{\mu}_w^{(k)} - \hat{\mu}_w^{*(k)} \mid \mathcal{I}_{(-k)}\right] = 0$ because of the conditional independence between $\left(1 - \frac{W_{w,i}}{\pi_w}\right)$ and $\left(\hat{\mu}_w^{(-k)}(X_i) - \mu_w(X_i)\right)$ in fold k and the fact that $\mathbb{E}\left[\left(1 - \frac{W_{w,i}}{\pi_w}\right)\right] = 0$. This, in turn, implies that $\mathbb{E}\left[\left(\hat{\mu}_w^{(k)} - \hat{\mu}_w^{*(k)}\right)^2 \mid \mathcal{I}_{(-k)}\right] = \mathbb{V}\left[\hat{\mu}_w^{(k)} - \hat{\mu}_w^{*(k)} \mid \mathcal{I}_{(-k)}\right]$, which we will use shortly. Moreover, the terms within the sum of eq. 10 are all IID conditional on $\mathcal{I}_{(-k)}$ so we can pass the variance through the sum (and gain a $1/n^{(k)}$ in the process):

$$\begin{aligned}
\mathbb{V}\left[\hat{\mu}_w^{(k)} - \hat{\mu}_w^{*(k)} \mid \mathcal{I}_{(-k)}\right] &= \frac{1}{(n^{(k)})^2}\sum_{i \in \mathcal{I}_k}^{n^{(k)}}\mathbb{V}\left[\left(1 - \frac{W_{w,i}}{\pi_w}\right)\left(\hat{\mu}_w^{(-k)}(X_i) - \mu_w(X_i)\right) \mid \mathcal{I}_{(-k)}\right] \\
&= \frac{1}{n^{(k)}}\left(\frac{1 - \pi_w}{\pi_w}\right)\mathbb{E}\left[\left(\hat{\mu}_w^{(-k)}(X_i) - \mu_w(X_i)\right)^2 \mid \mathcal{I}_{(-k)}\right]
\end{aligned} \tag{11}$$

Our plan is to show $\sqrt{n}(\hat{\mu}_w^{(k)} - \hat{\mu}_w^{*(k)}) \xrightarrow{L^2} 0$, which implies $\sqrt{n}(\hat{\mu}_w^{(k)} - \hat{\mu}_w^{*(k)}) \xrightarrow{p} 0$ because L^2 convergence implies convergence in probability. By the definition of L^2 convergence, that means we must show $\mathbb{E}\left[n\left(\hat{\mu}_w^{(k)} - \hat{\mu}_w^{*(k)}\right)^2\right] \to 0$, which we can do by using what we've derived above:

$$
\begin{aligned}
\mathbb{E}\left[n\left(\hat{\mu}_w^{(k)} - \hat{\mu}_w^{*(k)}\right)^2\right] &= n\mathbb{E}\left[\mathbb{E}\left[\left(\hat{\mu}_w^{(k)} - \hat{\mu}_w^{*(k)}\right)^2 \mid \mathcal{I}_{(-k)}\right]\right] \\
&= n\mathbb{E}\left[\mathbb{V}\left[\hat{\mu}_w^{(k)} - \hat{\mu}_w^{*(k)} \mid \mathcal{I}_{(-k)}\right]\right] \\
&= \frac{n}{n^{(k)}}\left(\frac{1-\pi_w}{\pi_w}\right)\mathbb{E}\left[\mathbb{E}\left[\left(\hat{\mu}_w^{(-k)}(X_i) - \mu_w(X_i)\right)^2 \mid \mathcal{I}_{(-k)}\right]\right] \\
&= \frac{n}{n^{(k)}}\left(\frac{1-\pi_w}{\pi_w}\right)\mathbb{E}\left[\left(\hat{\mu}_w^{(-k)}(X_i) - \mu_w(X_i)\right)^2\right]
\end{aligned}
\tag{12}
$$

This must converge to 0 because n and $n^{(k)}$ grow in proportion to each other and the expectation $\mathbb{E}\left[\left(\hat{\mu}_w^{(-k)}(X_i) - \mu_w(X_i)\right)^2\right]$ converges to 0 by our mean-square consistency assumption (eq. 3).

By the arguments above, this is enough to establish the asymptotic normality of our estimator with the efficient asymptotic variance: $\sqrt{n}(\hat{\tau} - \tau) \rightsquigarrow N(0, \nu^2)$.

Similar arguments are required to show that the plug-in variance estimator is consistent (when cross-fitting is used). Some care is required to address the terms $r'_w(\hat{\mu}_0, \hat{\mu}_1)$, etc., but the elementary tools of asymptotic statistics suffice.

A.2 Proof of theorem 1

Beginning with eq. 4, we have that

$$
\begin{aligned}
\nu_*^2 &= \mathbb{V}\left[r'_0\phi_0 + r'_1\phi_1\right] \\
&= r_0'^2\mathbb{V}\left[\phi_0\right] + r_1'^2\mathbb{V}\left[\phi_1\right] + 2r'_0r'_1\mathbb{C}\left[\phi_0, \phi_1\right]
\end{aligned}
\tag{13}
$$

from expansion of the variance. We will use several tricks to analyze these terms. The first is that $W_wY = W_wY_w$, by the fact that $Y = Y_0W_0 + Y_1W_1$ (note $W_w^2 = W_w$ and $W_0W_1 = 0$, which we will also use). Secondly, $W_w \perp Y_w, \hat{\mu}_w(X)$ by unconfoundedness, allowing for factorization of expectations.

These tricks and a few lines of algebra show that the variances in the first and second terms above (eq. 13) are

$$
\begin{aligned}
\mathbb{V}\left[\phi_w\right] &= \frac{1-\pi_w}{\pi_w}\mathbb{E}\left[\sigma_w^2(X)\right] + \mathbb{V}\left[Y_w\right] \\
&= \frac{1-\pi_w}{\pi_w}\kappa_w^2 + \sigma_w^2
\end{aligned}
\tag{14}
$$

where $\sigma_w^2(X) \equiv \mathbb{V}\left[Y_w|X\right]$ is the conditional variance function in treatment arm w. In the last line we define $\kappa_{w\cdot}^2 \equiv \mathbb{E}\left[\sigma_w^2(X)\right]$ (the average conditional variance) and $\sigma_w^2 \equiv \mathbb{V}\left[Y_w\right]$ (the marginal variance) to simplify notation.

Similar manipulation reduces the covariance term in eq. 13 to

$$
\begin{aligned}
\mathbb{C}\left[\phi_0, \phi_1\right] &= \mathbb{C}\left[\mu_0(X), \mu_1(X)\right] \\
&= \mathrm{Corr}[\mu_0(X), \mu_1(X)]\sqrt{\mathbb{V}\left[\mu_0(X)\right]\mathbb{V}\left[\mu_1(X)\right]} \\
&= \gamma\sqrt{(\sigma_0^2 - \kappa_0^2)(\sigma_1^2 - \kappa_1^2)}
\end{aligned}
\tag{15}
$$

by exploitation of the law of total variance $\underbrace{\mathbb{V}[Y_w]}_{\sigma_w^2} = \underbrace{\mathbb{E}\left[\sigma_w^2(X)\right]}_{\kappa_w^2} + \mathbb{V}[\mu_w(X)]$ and definition of $\gamma \equiv \text{Corr}[\mu_0(X), \mu_1(X)]$. Assembling the above and making explicit the known signs of r'_w gives

$$\nu_*^2 = r_0'^2 \left(\frac{\pi_1}{\pi_0}\kappa_0^2 + \sigma_0^2\right) + r_1'^2 \left(\frac{\pi_0}{\pi_1}\kappa_1^2 + \sigma_1^2\right) - 2|r'_0 r'_1|\gamma\sqrt{(\sigma_0^2 - \kappa_0^2)(\sigma_1^2 - \kappa_1^2)} \tag{16}$$

B Additional Simulation Results

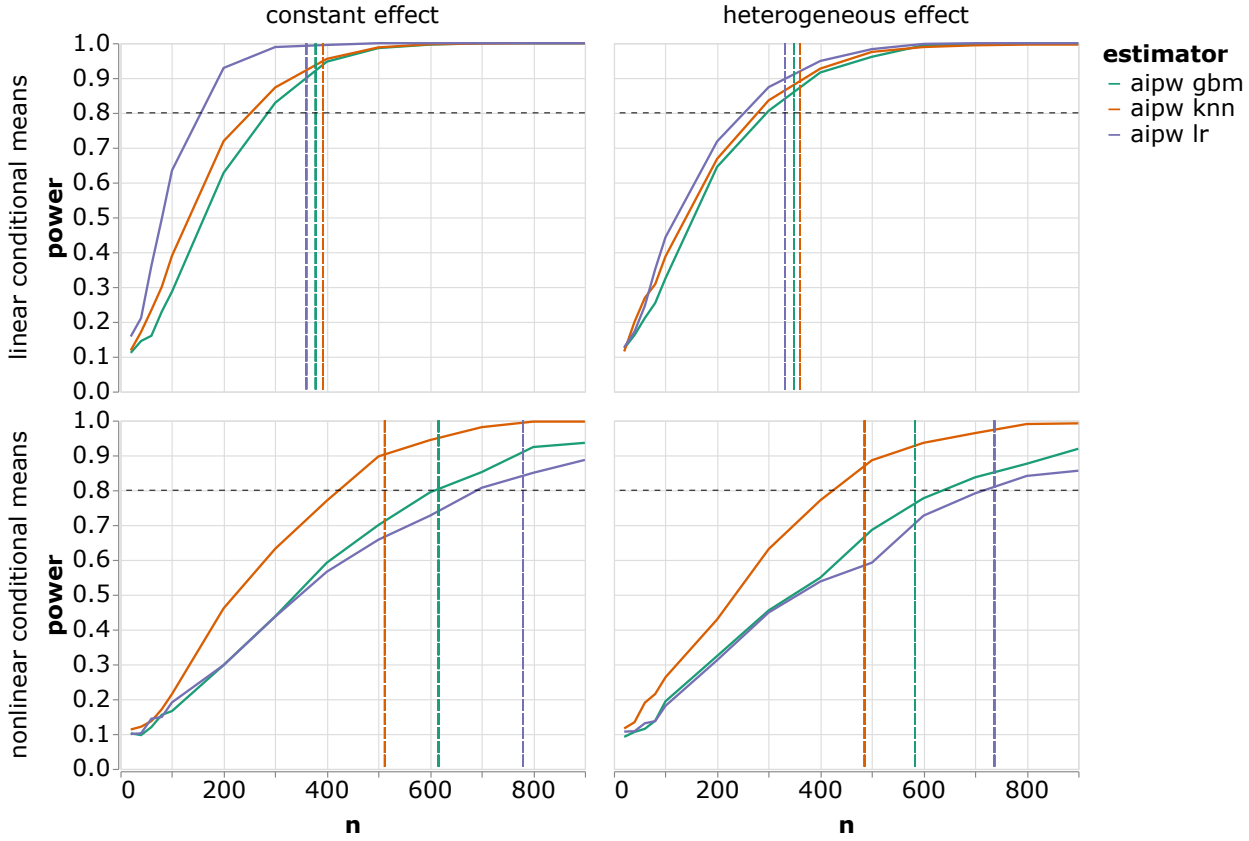

Figure 4: Empirical power and prospectively-calculated enrollment targets for AIPW estimators with different learners used to estimate the conditional means. Visual elements are as in figures 1 and 3.