Alejandro Schuler*, David Walsh, Diana Hall, Jon Walsh, Charles Fisher, for the Critical Path for Alzheimer's Disease, the Alzheimer's Disease Neuroimaging Initiative and the Alzheimer's Disease Cooperative Study

Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score

Abstract: Estimating causal effects from randomized experiments is central to clinical research. Reducing the statistical uncertainty in these analyses is an important objective for statisticians. Registries, prior trials, and health records constitute a growing compendium of historical data on patients under standard-of-care that may be exploitable to this end. However, most methods for historical borrowing achieve reductions in variance by sacrificing strict type-I error rate control. Here, we propose a use of historical data that exploits linear covariate adjustment to improve the efficiency of trial analyses without incurring bias. Specifically, we train a prognostic model on the historical data, then estimate the treatment effect using a linear regression while adjusting for the trial subjects' predicted outcomes (their *prognostic scores*). We prove that, under certain conditions, this prognostic covariate adjustment procedure attains the minimum variance possible among a large class of estimators. When those conditions are not met, prognostic covariate adjustment is still more efficient than raw covariate adjustment and the gain in efficiency is proportional to a measure of the predictive accuracy of the prognostic model above and beyond the linear relationship with the raw covariates. We demonstrate the approach using simulations and a reanalysis of an Alzheimer's disease clinical trial and observe meaningful reductions in mean-squared error and the estimated variance. Lastly, we provide a simplified formula for asymptotic variance that enables power calculations that account for these gains. Sample size reductions between 10% and 30% are attainable when using prognostic models that explain a clinically realistic percentage of the outcome variance.

Keywords: historical controls; prognostic score; randomized trial; sample size estimation.

*Corresponding author: Alejandro Schuler, UC Berkeley Center for Targeted Learning, Berkeley, CA, USA,
E-mail: alejandro.schuler@berkeley.edu. https://orcid.org/0000-0003-4853-6130
David Walsh, Diana Hall, Jon Walsh and Charles Fisher, Unlearn.AI, Inc., San Francisco, CA, USA

1 Introduction

The goal of much clinical research is to estimate the effect of a treatment on an outcome of interest (causal inference) [1]. The randomized trial is the gold standard for causal inference because randomization cancels out the effects of any unobserved confounders in expectation [2–4]. Although unobserved confounding is not a concern in randomized studies, we must still contend with the statistical uncertainty inherent to finite samples if we conduct our work in a population inference framework. Because of this, methods for the analysis of trial data should be chosen to safely minimize the resulting statistical uncertainty about the causal effect.

For a given treatment effect estimator and data-generating process, sample size is the primary determinant of sampling variance and power. Therefore the most straightforward method to reduce sampling variance is to run a larger trial that includes more subjects. However, trial costs and timelines typically scale with the number of subjects, making large trials economically and logistically challenging. Moreover, ethical considerations suggest that human subjects research should use the smallest sample sizes possible that allow for reliable decision making.

As most clinical trials compare an active treatment to a standard-of-care, often in combination with a placebo, there is a possibility to use existing control arm data[1] to augment clinical trials and reduce variance. Such "historical borrowing" methods are becoming increasing attractive as the creation of large, electronic patient datasets in the past decade facilitates this process by making it easier to find a suitably matched historical population. Various approaches ranging from directly inserting subjects from previous studies into the current sample to using previous studies to derive prior distributions for Bayesian analyses have been proposed [5–8]. Although such methods do generally increase power, they cannot strictly control the rate of type I error [5, 8, 9].

Here, we describe an alternative approach that exploits machine learning models and historical control arm data to decrease the uncertainty in effects estimated from randomized trials without compromising strict type-I error rate control in the large-sample setting. The gist of our proposal is to use the historical data to train a prognostic model that predicts a patient's outcome given their baseline covariates. This prognostic model is applied to all patients in the current trial in order to generate a prediction of their outcome (their "prognostic score" [10–12]). The score is then adjusted for as a covariate in a (linear) regression model of the trial outcome in order to estimate the treatment effect. This amounts to adding a single (constructed) adjustment covariate into an adjusted analysis. As such, it poses no additional statistical risk[2] over any other pre-specified adjusted analyses, which are preferable to unadjusted analyses in almost every case [13–16]. Our approach is entirely pre-specifiable, is generic enough to be integrated into many analysis plans, and is supported by regulatory guidance [17]. Moreover, there are no practical restrictions on the type of prognostic model used, enabling one to leverage machine learning-based methods that can learn nonlinear predictive models from large databases.

Below, we show that this prognostic covariate adjustment procedure attains the minimum possible asymptotic variance among "reasonable" estimators as long as the prognostic model improves with more data and the effect of treatment is constant. The uncertainty in the estimate for the treatment effect is minimized when the prognostic model predicts the control potential outcome of each subject. However, one can realize gains in efficiency even with imperfect prognostic models or in the presence of heterogeneous

---

1 Treatment-arm historical data is usually more difficult to come by, especially if the active treatment under consideration is experimental and has not been previously tested. However the approach we propose is generic enough to handle relatively arbitrary historical data. The improvements in efficiency will depend on how similar the historical data-generating process is to the current one.

2 There is some confusion and subtlety around this point. Freedman [52] is sometimes referenced to claim that covariate adjustment can incur bias, but in fact, the presence of treatment interaction terms resolves the issue [13]. Moreover, both of these results apply only to the "randomization inference" setting where the covariates and potential outcomes are considered fixed (i.e. randomness arises only from treatment assignment, not sampling). We use the more-common population inference framework in this work, in which these criticisms do not apply [51].

effects. In general, our procedure decreases the variance of the estimated treatment effect in proportion to the squared correlation of the prognostic model with the outcome while guaranteeing unbiasedness, control of type-I error rate, and confidence interval coverage. We demonstrate the efficiency gain in simulations and through a reanalysis of a previously reported clinical trial studying the effect of docosahexaenoic acid (DHA) on cognition in patients with Alzheimer's disease.

Although we are (to our knowledge) the first to formally characterize it, prognostic covariate adjustment has been used in trials for a long time. The baseline covariates in a trial are often "atomic" measurements such as sex, age, or lab values, but composite or computed covariates such as body mass index, Charlson comorbidity index, or Framingham risk score, are also frequently used [14, 15, 18–20]. These "indices" or "scores" are usually the output of a simplified prognostic model that has been learned (at least implicitly) from historical data. For instance, the Framingham cardiovascular risk score was developed by training Cox and logistic regression models using a large community-based cohort to obtain a single covariate that is highly predictive of cardiovascular outcomes. From that perspective, our proposed approach is a formalization of what has previously been an ad-hoc procedure.

The novel contributions of this paper are therefore threefold. Firstly, we provide a formal characterization of prognostic covariate adjustment. Although this process has already been used in trials (if one considers a baseline outcome measurement or risk score as a rudimentary prognostic score), it has not been formally described as a method for leveraging historical data to improve the efficiency of a proposed trial. Our second novelty is an asymptotic proof that shows prognostic covariate adjustment is semiparametric efficient if the effect of treatment is constant, the historical data follow the same distribution as the trial control arm, and the prognostic model improves with the amount of historical data. Roughly translated, this means that the power of a trial using prognostic covariate adjustment will be higher or equal to the power of any other trial design that controls type I error. Thirdly, we provide a method of sample size estimation that allows for the design of smaller trials with prognostic covariate adjustment that maintain their power. The formula we derive coincides in certain special cases with previously known results, but to our knowledge our derivation is more general and rests on fewer assumptions than what is available elsewhere in the literature [21].

2 Setting and notation

Our setting is a two-arm randomized clinical trial. Denote the outcome for subject i in the clinical 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). Throughout the paper we assume a continuous outcome. 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,

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

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

We denote the population average outcomes under each treatment condition w as $\mu_w = \mathbb{E}[Y_w]$ and the conditional means as $\mu_w(X) = \mathbb{E}[Y_w|X]$. In general, a "treatment effect" is any function of these marginal means, i.e. $\tau = r(\mu_0, \mu_1)$, but when we say treatment effect here we specifically mean the difference in means, $\tau = \mu_1 - \mu_0$.

Finally, we denote the treatment indicators as $W_{1,i} = W_i$ and $W_{0,i} = 1 - W_i$ to allow for symmetric notation. 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, 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 n is increasing while other properties of the data-generating process remain fixed. As usual, the "asymptotic varaince" v^2 of an asymptotically normal estimator $\hat{\tau}_n$ of a parameter τ refers to the variance of the limiting distribution of the sequence $\sqrt{n}(\hat{\tau}_n - \tau) \rightsquigarrow N(0, v^2)$. We omit the subscript n where the relevant sequence is clear from context.

3 Prognostic covariate adjustment

At a high level, our approach to historical borrowing (which we call prognostic covariate adjustment) is to use the historical data to learn a prognostic model, then apply it to the trial to generate an additional adjustment covariate. Specifically, let $\mathcal{M}$ be some learning algorithm (e.g. a linear regression, random forest, deep neural network, etc.) which, when trained on our historical dataset $(\boldsymbol{X'}, \boldsymbol{Y'})$, outputs a "fit" predictive model, $m: \mathcal{X} \to \mathcal{Y'}$. We assume throughout that the historical and trial data are statistically independent. This is natural since we already assume that the individual observations *within* each dataset are independent, but does preclude scenarios where a subject present in the historical data later enters the trial, for example.

The treatment effect estimate we will use is consistent for any choice of prognostic model, so we can discuss the method in great generality without assuming the prognostic model takes a specific form or attains a certain level of predictive performance. That said, it is particularly interesting to consider the "best case scenario" in which the historical data are representative of the trial control arm, i.e. $P(X', Y') = P(X, Y_0)$. It will be later be shown that the prognostic model that endows our estimator with minimum variance is the conditional mean, $m(X) = \mathbb{E}[Y_0|X]$, which is the same as $m(X') = \mathbb{E}[Y'|X']$ in the best case scenario. Therefore, the construction of the optimal prognostic score involves estimating a function equivalent to the conditional mean of the outcome given the covariates in the historical sample of controls. As such, constructing the optimal prognostic model is a standard machine learning problem. In the results that follow, we make no presumption about the particular choice of $\mathcal{M}$. We refer to $m(\cdot)$ as a *prognostic model* and to the quantity $M_i = m(X_i)$ as the *prognostic score*. Hansen [10] defines the prognostic score as any quantity $f(X)$ which induces $Y_0 \perp X|f(X)$. In the literature, the prognostic score is often treated more informally as the expected value of the control outcome given the baseline covariates, which motivates our terminology here.

While it is optimal to construct a prognostic model that accurately approximates $m(X) = \mathbb{E}[Y_0|X]$, this is not strictly necessary for using our proposed estimator. Indeed, there may be situations in which this is not possible in practice. For example, the historical sample may be too small to reliably learn the relationship, the distribution of the covariates in the historical population may not reflect the trial population (i.e. a "domain shift"), or the outcome in the trial may not even have been measured in the historical data (e.g. a biomarker measured with a new technology). In this latter case, we refer to Y' as a "surrogate" outcome for Y. Whatever the reason, the proposed estimator will be unbiased, retain control of the type-I error rate, and decrease variance (to some degree) even if the prognostic model does not accurately approximate the control potential outcomes.

In general, we can assume that the prognostic model, m, is given at the time of the trial. To analyze the trial data, we first use the prognostic model to generate the prognostic score, $M_i = m(X_i)$, for each subject given their baseline covariates, X_i. Then, we estimate the treatment effect using a linear regression adjusted

for the empirically centered covariates, prognostic score, and their interactions with the treatment.[3] Letting $Z^\top = [1, \tilde{W}, \tilde{X}^\top, \tilde{M}^\top, \tilde{W}\tilde{X}^\top, \tilde{W}\tilde{M}^\top]$ be the regressors,[4] we fit $\mathbb{E}\left[Y_i|Z_i\right] = Z_i^\top \beta$ using ordinary least squares to obtain the fit coefficients, $\hat{\beta}$. Our estimate of the treatment effect is $\hat{\tau} = \hat{\beta}_W$, i.e. the coefficient corresponding to the W term in the regression. This specification is directly based on the "ANCOVA II" estimator analyzed in Yang and Tsiatis [16]. It is well-known that this is a consistent and asymptotically normal estimator of the treatment effect when treatment is randomized, even if the regression is misspecified (i.e. the true relationship is nonlinear)[5] [16, 23, 24].

We estimate sampling variance with the usual "Sandwich" estimator $\widehat{\mathbb{V}\left[\hat{\tau}\right]} = (\boldsymbol{ZZ}^\top)^{-1}DD^T(\boldsymbol{ZZ}^\top)^{-1}$ in which $D = \boldsymbol{Z}\,\mathrm{diag}\,(\boldsymbol{Y} - \boldsymbol{Z}^\top\hat{\beta})$. This variance estimator is consistent whether or not the regressions are correctly specified and is consequently robust against deviations from linearity and homoscedasticity [25]. This guarantees strict (large-sample)[6] type-I error rate control based on p-value cutoffs and valid confidence intervals in practically all cases (i.e. as long as $P(Y_0, Y_1, X)$ satisfies mild regularity conditions). In particular, the statistical validity of the inference is not impacted by the nature of the prognostic score M because it enters into the analysis like any other covariate.

3.1 Statistical properties

3.1.1 Optimality under constant effects

Although prognostic covariate adjustment provides valid inference (i.e. proper coverage and asymptotic unbiasedness) under general conditions, we first motivate it by showing that it is in fact *optimal* under the assumption of a constant treatment effect and certain assumptions about the prognostic model.

Theorem 1. *Presume a constant treatment effect $\mu_1(X) = \mu_0(X) + \tau$. Then the linear prognostic covariate adjustment procedure that uses $m(X) = \mathbb{E}\left[Y_0|X\right]$ as the prognostic model has the lowest possible asymptotic variance among all regular and asymptotically linear estimators with access to the covariates X.*

This is restated and proved as Corollary A.6.1 in the Appendix A. Corollary A.6.2 shows the statement also holds when the interaction terms are omitted from the working regression model. All practical and reasonable estimators in the context of trial analyses with continuous outcomes are regular and asymptotically linear,[7] so this result means linear adjustment for the true control conditional mean is, in some sense, the absolute "best" possible estimation procedure [26]. Since the prognostic model is trained to approximate $\mathbb{E}\left[Y'|X'\right]$, it is best for the historical data to be drawn from the same distribution as the trial control arm so that $\mathbb{E}\left[Y'|X'\right] = \mathbb{E}\left[Y_0|X\right]$.

---

3 Theorem A.5 in the appendix shows that including interactions is necessary to ensure that the estimator is more efficient than difference-in-means estimation. The estimated coefficients of these interactions are of no interest to us here—they are merely a tool to reduce variance in estimation of the main treatment effect. There is a separate and large literature on the estimation of heterogeneous treatment effects, of which testing for linear treatment-covariate interactions is one small part.

4 If the prognostic score M is identically equal to any of the covariates or is numerically constant, we omit including it in the regression since in this case it cannot do anything to reduce variance.

5 See footnote .

6 In practice, it may be advisable to employ corrections for this estimator if one is working with a small sample. The estimators known as HC1, HC2, and HC3 are all reasonable options that make minor changes to the matrix D, e.g. $D_{\mathrm{HC1}} = \sqrt{\frac{n}{n-p}}D$. See Long and Ervin [53].

7 Regularity and asymptotic linearity are specific technical conditions. Definitions may be found in Tsiatis [26].

We can in fact weaken our assumption that the prognostic score is perfect ($m(X) = \mathbb{E}[Y_0|X]$) and still obtain the same conclusion as long as the prognostic model approaches the truth as the size of the external dataset increases. The proof of the following is also given in the appendix (Theorem A.8).

Theorem 2. *Presume X has compact support and there is a constant treatment effect: $\mu_1(X) = \mu_0(X) + \tau$ with $|\mu_0(x)| < b$ bounded. Let $m(x)$ be a (random) function learned from the external data $(\boldsymbol{Y'}, \boldsymbol{X'})_n$, such that $|m(x)| < b$ is also bounded and $|m(X) - \mu_0(X)| \xrightarrow{L_2} 0$ so that the learned model approaches the truth in mean-squared error (MSE) as $n' \to \infty$. If the number of trial samples n grows in tandem with the size of the historical data n' (i.e. $n = O(n')$), then the linear prognostic covariate adjustment estimator that uses the learned model $m(X)$ in the role of X has the lowest possible asymptotic variance among all regular and asymptotically linear estimators with access to the covariates X.*

This also holds without the use of interaction terms in the regression (Corollary A.8.1). Of course, $m(X) = \mathbb{E}[Y_0|X]$ will never hold for any given prognostic model, either because the model is not perfectly learning the relationship or because the training data are not entirely representative of the trial control arm. Despite this, the theorem means that our statement of optimality is not meaningless because with enough external data we will come closer and closer to attaining the minimum possible variance. Note that the condition that $|m(X) - \mu_0(X)| \xrightarrow{L_2} 0$ is relatively weak and justifies the use of a number of machine learning algorithms $\mathcal{M}$ to learn the prognostic model [27–30].

3.1.2 Superiority over covariate adjustment without prognostic score

It is not always reasonable to assume a constant effect. However, we can show that even if it is not optimal among all estimators, prognostic covariate adjustment still retains an advantage over standard covariate adjustment (Definition A.1). This holds for any fixed prognostic model $m(x)$.

A note on interaction terms. The following results require the aforementioned interaction terms to be present in the regression model. If interactions are omitted, the results below only hold if either the treatment effect is constant or the randomization is 1:1. Without these conditions and without interaction terms, there are cases in which adding the prognostic score (or any covariate) could actually increase the asymptotic variance (Theorem A.5). These cases may be rare in practice but either way the problem can be entirely avoided simply by including interactions. Theorem A.5 also shows that these interactions are always useful even without the prognostic score.

Theorem 3. *Assume only mild regularity conditions to ensure that the usual sandwich estimator is consistent. Consider linearly adjusting (with treatment-covariate interactions) for covariates X with variance $\mathbb{V}[X] = \Sigma_x$ and covariance with Y_w of $\mathbb{C}[Y_w, X] = \xi_{w,x}$ versus a set of covariates $[X, M]$ ($M \in \mathbb{R}$) (again using interactions) with $\mathbb{C}[X, M] = \zeta$, $\mathbb{V}[M] = \sigma_m^2$ and $\mathbb{C}[Y_w, M] = \xi_{w,m}$. Let $\xi_{m*} = \pi_0\xi_{1,m} + \pi_1\xi_{0,m}$ and $\xi_{x*} = \pi_0\xi_{1,x} + \pi_1\xi_{0,x}$. Assume M is not a linear combination of the variables in X. The reduction in asymptotic variance from including the prognostic score M in the regression is always nonnegative and given by*

$$\left(\frac{1}{\pi_0\pi_1}\right) \frac{\left(\xi_{m*} - \xi_{x*}^\top \Sigma_x^{-1} \zeta\right)^2}{\sigma_m^2 - \zeta^\top \Sigma_x^{-1} \zeta} \tag{2}$$

This is restated and proved as Corollary A.4.2 in the Appendix A. The conclusions are that (i) including a prognostic score as a covariate should never hurt the asymptotic variance and (ii) the efficiency gain depends on how well correlated the prognostic score is with the outcome (above and beyond any correlation with the raw covariates). Prognostic models that are better correlated with the outcome thus offer larger efficiency gains, but the presence of any correlation at all could still decrease the variance. This justifies the use of prognostic covariate adjustment for surrogate outcomes (i.e. when Y' and Y represent different but, perhaps

correlated, outcomes) and in cases with heterogeneous treatment effects. It is worth mentioning that the prognostic score must be a nonlinear function of the included covariates X for there to be any benefit. In fact, there would not be a unique solution for the coefficients of the regression if the prognostic score and covariates were exactly colinear.

3.1.3 Pragmatic sample size calculation

Our primary goal in this paper is to demonstrate how prognostic covariate adjustment decreases the variance for a fixed trial relative to raw covariate adjustment. However, it is also useful to consider how knowledge of this efficiency gain could be exploited to conduct *smaller* trials that attain a desired level of confidence. Since the asymptotic variance of the estimate determines the power of the trial, smaller trials using prognostic covariate adjustment may attain equal power to a larger trial using raw covariate adjustment. Power must be estimated *before* running the trial in order to gauge the number of subjects to be enrolled. This requires a formula for sampling variance. Although Theorem 3 gives a precise quantification of the efficiency gained by including the prognostic score, it may be impractical to calculate this quantity because one would need to estimate a potentially large number of population parameters (e.g. the entries of Σ_x). Here we provide an upper bound on the resulting sampling variance that uses fewer population parameters and is therefore easier to apply in a prospective setting.

Theorem 4. *Assume only mild regularity conditions to ensure that the usual sandwich estimator is consistent. Given an arbitrary, fixed prognostic score $M = m(X)$, the asymptotic variance of our proposed estimation procedure is no greater than*

$$\frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} - \pi_0 \pi_1 \left(\frac{\rho_1 \sigma_1}{\pi_1} + \frac{\rho_0 \sigma_0}{\pi_0} \right)^2 \tag{3}$$

where $\sigma_w = \mathbb{V}[Y_w]$ and ρ_w is the population correlation between M and Y_w, i.e. $\mathbb{C}[M, Y_w] / \sqrt{\mathbb{V}[M] \mathbb{V}[Y_w]}$. This is always less than the asymptotic variance of the unadjusted estimator (an upper bound on the variance of the standard covariate-adjusted estimator that uses the same population parameters), which is

$$\frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} \tag{4}$$

The bound follows directly from 8, 9, and 8.2 in the appendix. This bound does not account for any reduction in variance due to adjustment for the raw covariates, X. However, if the prognostic model is accurate, the raw covariates are unlikely to provide substantial efficiency gains because their effects are already "soaked up" by the prognostic score. In addition, the only population parameters in this bound are the marginal outcome variances and model-outcome correlations in each treatment arm.

In conjunction with estimates of σ_w^2 and ρ_w, this formula may be used to prospectively calculate a lower bound on the power of a clinical trial analyzed with prognostic covariate adjustment. Asymptotic normality means that, in the limit, the probability of a two-sided p-value being less than α (i.e. a "statistically significant" result) is

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

where Φ is the CDF of the standard normal, τ is the true (target) treatment effect and v^2 is the asymptotic variance of whatever asymptotically normal estimator is being used. Composing the variance bound for the prognostic covariate adjustment estimator given in Theorem 4 with this formula gives an upper bound for the power of trial analyzed with that estimator that depends only on the target effect τ, sample sizes n_w, potential outcome variances σ_w^2, and potential outcome-prognostic score correlations ρ_w. The target effect is usually fixed *a priori*. The latter two quantities may be estimated using historical data and/or expert opinion (see Appendix B). After fixing them, the sample sizes may be varied in a desired ratio until the desired power (e.g. 80%) is achieved.

Because of efficiency gains, the required sample size for a trial powered with this method and analyzed with prognostic covariate adjustment will be lower than for a trial powered without exploiting covariate information. We can equate the power for a target effect τ of an estimator with variance given by Eq. (3) (upper bound on prognostic covariate adjustment) to one with variance given by 4 (upper bound on raw covariate adjustment) to algebraically discover the relationship between n, the sample size required for a well-powered analysis with raw covariate adjustment, and $n^\dagger$, the sample size required for a well-powered analysis with prognostic covariate adjustment. In the case of a 1:1 randomization ratio ($\pi_0 = \pi_1 = 0.5$), this relationship is

$$\frac{n^\dagger}{n} = 1 - \frac{(\sigma_0 \rho_0 + \sigma_1 \rho_1)^2}{2\left(\sigma_0^2 + \sigma_1^2\right)} \tag{6}$$

With a common variance $\sigma_0 = \sigma_1 = \sigma$ and correlation $\rho_0 = \rho_1 = \rho$ this simplifies to $\frac{n^\dagger}{n} = 1 - \rho^2$, or, in terms of a percent reduction in sample size from n, $\frac{n-n^\dagger}{n} = \rho^2$. This is precisely the out-of-sample R^2 of the prognostic model m. This relationship holds regardless of the value of the target effect and the desired power. The result coincides with the formula from Borm et al. [21], although our derivation is more robust because we do not rely on parametric assumptions. A generic relationship that does not assume 1:1 randomization is easily derived in the same fashion.

R^2 values can vary wildly depending on the model, outcome, and population, but folk wisdom among biostatisticians has it that values between 0.1 and 0.3 might be expected from a good prognostic model. Those values translate to meaningful sample size reductions between 10% and 30%.

Note that the bound in Theorem 4 is actually generic to any covariate because M enters the estimator the same way any other covariate does. In other words, the result still holds if one substitutes any X_j for M in the theorem. For instance, presume baseline age X_0 is known to be correlated with the standard-of-care (control) outcome at a strength of $\rho_{X_0} = 0.2$. Then in a 1:1 trial (presuming common σ^2 and ρ) we could reduce the sample size by 4% relative to the unadjusted power calculation and still maintain the same design power for the same target effect, all without the need for a dedicated prognostic model.[8]

4 Simulations

Unlike analyses of real data, analyses of simulated data can be compared to ground truth to gauge error. We used simulation to explore how mean-squared estimation error of the treatment effect varies with and without prognostic covariate adjustment. In particular, we were interested in cases with or without strong effects, non-linearity in the outcome-covariate relationship, heterogeneity in the treatment effect, surrogate outcomes, or distributional shifts between the historical and trial data-generating processes. We show that prognostic covariate adjustment performs better than raw covariate adjustment in all cases where theory expects it to and performs worse in no cases. The amount of improvement from scenario to scenario is in line with intuition gained from theory.

We chose our simulation scenarios to demonstrate the different performance benefits of effect estimation with prognostic covariate adjustment relative to raw covariate adjustment in cases that might come up in real trials. Effect sizes vary from trial to trial but the overall size of the true treatment effect should not change the relative efficiency of prognostic and raw covariate adjustment (i.e. because the asymptotic variances do not depend on the treatment effect) so we chose this scenario to demonstrate the point. Possible nonlinearities must always be contended with in trial analyses. For our purposes, outcomes that are linearly related to the covariates should not benefit from prognostic covariate adjustment above and beyond raw covariate adjustment because all of the information from the raw covariates is exploitable by the linear model. Heterogeneity of effect is always possible and should decrease the advantage of prognostic covariate

---

8 A prognostic model will, of course, be expected to approach the maximum possible correlation with the outcome that could be attained as a function of the baseline covariates. Thus there is usually a benefit to using such a model if one is available.

adjustment to some extent because the prognostic score is less predictive in the treatment arm. Using a surrogate outcome should also decrease the advantage of the prognostic score because its predictive capacity is lessened in both treatment arms of the trial – this case may arise if the trial outcome is not well-measured in historical data. Lastly, it is always possible that the historical data are not representative of the current trial population. When this is the case, we expect the benefits of prognostic covariate adjustment to be attenuated because the model must extrapolate outside of its trained range to perform on the trial population.

Each of our simulation scenarios is defined by particular choices for the pair of distributions $P(X', Y')$ and $P(X, Y_0, Y_1)$. In all cases, the distribution of covariates in the simulated historical and trial data were 10-dimensional uniform random variables in the prism $[l, h]^{10}$. Distributional shift was modeled by choosing different values of l and h for $P(X')$ and $P(X)$. The distributions $P(Y'|X')$, $P(Y_0|X)$, and $P(Y_1|X)$ 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, with $a = 0$ representing the linear case. In this context, treatment effect heterogeneity refers to the situation in which a or b is different for $P(Y_0|X)$, and $P(Y_1|X)$, whereas surrogate outcome refers to the situation in which a or b is different for $P(Y'|X')$ and $P(Y_0|X)$. Large constant effects are encoded with different values for c in $P(Y_0|X)$, and $P(Y_1|X)$ while keeping a and b the same. The specific values of l, h for each covariate distribution and of a, b, and c are shown in Table 1.

The "baseline" simulation scenario included some moderate outcome non-linearity, constant treatment effect, and no distributional shift between the historical data and the trial control arm. We tested four variations of this scenario. In the first ("linear") we examined what happened when the outcome-covariate relationship was precisely linear in both treatment arms. In the second ("heterogeneous effect") we tested a variation of the baseline scenario where the conditional average effect $\mathbb{E}[Y_1 - Y_0|X]$ was no longer a constant. In the third ("surrogate outcome") we tested a variation where the relationship between outcome and covariates in the historical data was not representative of the corresponding relationship in the control arm of the trial, i.e. $P(Y'|X' = x) \neq P(Y_0|X_0 = x)$. In the fourth ("covariate shift") we tested a variation where the historical population was not representative of the trial population in terms of the baseline covariates, i.e. $P(X' = x) \neq P(X = x)$.

In each simulation scenario, we generated a historical control dataset $(\boldsymbol{X}', \boldsymbol{Y}')$ by drawing 10,000 IID samples from a specified distribution, $P(X', Y') = P(Y'|X')P(X')$. These simulated historical data were used to train to a random forest (1000 trees, with other parameters set to defaults in the python package sklearn [31]) as a prognostic model, $m: \mathcal{X} \to \mathcal{Y}$. Then, we simulated a randomized trial dataset $(\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{Y})$ with 500 subjects, evenly split between treatment and control. The data-generating process for these data involved drawing 500 IID samples from a counterfactual distribution $P(Y_1, Y_0, X) = P(Y_1|X)P(Y_0|X)P(X)$, evenly splitting the sample into treatment and control arms, and then setting $Y = Y_1$ for the treated and $Y = Y_0$ for the controls. Finally, we used the prognostic model to generate the prognostic score, $\boldsymbol{M} = m(\boldsymbol{X})$, and analyzed the data using four estimation procedures: unadjusted, covariate-adjusted, covariate-adjusted with prognostic score,

Table 1: Parameters for all simulation scenarios.

Scenario	$P(X')$		$P(X)$		$P(Y'|X')$			$P(Y_0|X)$			$P(Y_1|X)$		
	l'	h'	L	H	a'	b'	c'	a_0	b_0	c_0	a_1	b_1	c_1
Baseline	**−1**	**1**	**−1**	**1**	**0.5**	**1**	**0**	**0.5**	**1**	**0**	**0.5**	**1**	**0**
Strong effect	−1	1	−1	1	0.5	1	0	0.5	1	0	0.5	1	5
Linear	−1	1	−1	1	0	1	0	0	1	0	0	1	0
Heterogeneous effect	−1	1	−1	1	0.5	1	0	0.5	1	0	0	1	0
Surrogate outcome	−1	1	−1	1	0.5	−1	0	0.5	1	0	0.5	1	0
Covariate shift	−2	0	−1	1	0.5	1	0	0.5	1	0	0.5	1	0

Parameters for the baseline scenario are shown in bold.

Table 2: Mean-squared errors (MSEs) of each estimator in each simulation scenario.

Scenario	Unadjusted	+Covariates + interactions	+Covariates + prognostic score + interactions	+Covariates + prognostic score	Oracle
Baseline	0.076	0.051	**0.017**	**0.017**	0.008
Strong effect	0.077	0.051	**0.018**	**0.018**	0.008
Linear	0.035	**0.008**	**0.008**	**0.008**	0.008
Heterogeneous effect	0.055	0.030	**0.021**	0.022	0.020
Surrogate outcome	0.075	0.050	0.038	**0.037**	0.008
Covariate shift	0.077	0.050	**0.049**	**0.049**	0.008

The result with the smallest MSE (excluding the oracle estimator) is shown in bold. The headings correspond to: $Z^\top = [1, \tilde{W}]$ (unadjusted), $Z^\top = [1, \tilde{W}, \tilde{X}^\top, \tilde{W}\tilde{X}^\top]$ (+covariates + interactions), $Z^\top = [1, \tilde{W}, \tilde{X}^\top, \tilde{M}^\top, \tilde{W}\tilde{X}^\top, \tilde{W}\tilde{M}^\top]$ (+covariates + prognostic score + interactions), $Z^\top = [1, \tilde{W}, \tilde{X}^\top, \tilde{M}^\top]$ (+covariates + prognostic score), $Z^\top = \left[1, \tilde{W}, \tilde{W}\tilde{\mu}_0(X)^\top\right]$ (oracle).

and covariate-adjusted with prognostic score sans interactions.[9] We also repeated the prognostic-score-and-covariate-adjusted analysis using the true "oracle" prognostic score $\mathbb{E}[Y_0|X = x]$ to show a best case scenario in terms of the prognostic model (this estimator would not be feasible in practice). Results from additional regression specifications are available in the appendix.

The result was a set of five effect estimates. We calculated the squared-error of each estimate relative to the true treatment effect, $\mathbb{E}[Y_1 - Y_0]$, known from the data-generating counterfactual distribution, repeated this process 10,000 times, and averaged the squared-errors to obtain MSEs for each estimator. The results are shown in Table 2.

The simulation results support the theory in that the MSE of the estimator with prognostic covariate adjustment was always less than or equal to the MSE without it. The only two cases in which prognostic covariate adjustment did not substantially decrease the MSE were the linear and distributional shift scenarios. In the former, the linear prognostic relationship is already captured by the other baseline covariates so there is no additional benefit to adding the prognostic score, as expected. In the latter, the prognostic model may not generalize well to the study population, thereby losing a significant portion of its predictive power. All other scenarios, however, demonstrate that adding the prognostic score significantly reduced the MSE. Finally, most or all of the benefit was realized without including the interaction terms, except for a modest gain in the scenario with treatment effect heterogeneity.

5 Case study

In addition to the simulations presented above, we re-analyzed data from an existing trial to demonstrate how prognostic covariate adjustment decreases variance relative to a standard covariate-adjusted analysis. Our results show that prognostic covariate adjustment decreases the estimated standard errors relative to raw covariate adjustment.

Our demonstration trial, reported by Quinn et al. [32], was conducted to determine if 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

---

9 "Adjusted for" means by default that we included both the main effect of the covariate and its interaction with the treatment. We also report results for an estimator adjusted for the covariates and prognostic score, but excluding the interaction term. Regulatory guidance often recommends against the inclusion of interactions without strong *a-priori* evidence for their existence [17]. And although theoretically useful, the practical benefits from including the interactions may be negligible in practice. Indeed, it follows from Theorem A.5 that the benefit disappears when there is no heterogeneity of effect.

medical centers and private Alzheimer disease clinics funded by the National Institute on Aging to conduct clinical trials on Alzheimer disease.

Quinn et al. 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 [33]), and component scores of cognitive tests. A full list of the 37 covariates we used is available in the appendix (see Table 4). Any missing covariate values were mean-imputed in our reanalysis.

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) [34] over the duration of the trial (18 months). Decrease in an activities of daily living (ADL) score [35] and increase in Clinical dementia rating (CDR) [36] were also recorded in the study and we treated these as secondary endpoints to demonstrate the benefits of prognostic covariate adjustment for a surrogate outcome.

Before examining the trial data, we fit a prognostic model for the increase in ADAS-Cog 11 over 18 months conditional on the measured covariates. To train the model we used a large historical training dataset comprised of 6919 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 [37, 38], 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 clinical trials. These data also included the same baseline covariates as were measured in the DHA trial (imputed to a column mean where missing). We used a random forest with 1000 trees to learn the prognostic model; all other parameters were left to their defaults in the python sklearn package [31].

Once fit, we applied our prognostic model to generate a prognostic score for each subject in the trial dataset; that is, we used the trained random forest model to predict the change in ADAS-Cog11 for a particular patient under standard-of-care. We then estimated the treatment effect on each outcome using three different methods: (i) difference-in-means (i.e. unadjusted linear regression), (ii) linear regression adjusted for the baseline covariates, and (iii) linear regression adjusted for the baseline covariates and the ADAS-Cog 11 prognostic score corresponding to the appropriate timepoint for the trial, and (iv) the same estimator as case (iii) but with the interaction terms omitted. Note that in cases (iii) and (iv) the ADL and NPI outcomes were analyzed with our ADAS-Cog 11 prognostic score, not a separate ADL or CDR prognostic score. The purpose of this is to demonstrate how prognostic covariate adjustment works for a surrogate outcome. We report results in terms of an estimated effect and 1.96 times an estimated standard error (i.e. a 95% confidence interval) in Table 3.

Concordant with our simulation studies, the standard errors for the effect obtained using prognostic covariate adjustment were less than or equal to those obtained using unadjusted for standard covariate-adjusted analyses. This led to narrower confidence intervals (which are still theoretically guaranteed to have the correct frequentist coverage). Using prognostic covariate adjustment decreased the standard error for estimated treatment effects on ADAS-Cog11 and ADL, even though we adjusted for the predicted change in

Table 3: Results from the reanalysis of the DHA trial.

Outcome	Unadjusted	+Covariates + interactions	+Covariates + prognostic score + interactions	+Covariates + prognostic score
ADAS-Cog 11	−0.10 ± 2.02	0.58 ± 1.72	**0.56 ± 1.69**	**0.41 ± 1.69**
ADL	−0.31 ± 3.11	0.27 ± 2.58	0.34 ± 2.56	**−0.06 ± 2.43**
CDR	−0.01 ± 0.65	**0.03 ± 0.54**	**0.03 ± 0.54**	**0.04 ± 0.54**

Results are shown in terms of estimated effect ±1.96 × estimated standard deviation (Sandwich estimator). The result with the smallest estimated standard deviation is in bold for each outcome (row). Headings are the same as in Table 2.

ADAS-Cog11 for both outcomes. There was negligible impact of including the surrogate prognostic score for CDR. Similarly, including the interaction term in the prognostic-adjusted regressions had little, or deleterious, impact in this experiment.

Covariate and prognostic covariate adjustment did modify the point estimates for the treatment effects to some extent, but only minimally relative to the size of the estimated standard errors. Even though the point estimates for individual studies can change, adjusting for baseline covariates or a prognostic score does not add bias [16, 23, 24]. In this particular trial, none of the outcomes demonstrated statistically significant improvements regardless of the analysis used, consistent with the original analyses of these data [32].

5.1 Power analysis

In addition to the reanalysis, we considered how we would power this trial had we been involved prospectively. Since trials must be designed before they are run, we did not use any data collected from the DHA trial to do this.

Based on our historical training dataset, we estimated the marginal outcome variance to be $\sigma_0^2 = 61.76$ and the cross-validated prognostic model-outcome correlation to be $\rho_0 = 0.44$ (the maximum correlation between any single covariate and the outcome in the training data was 0.39, for baseline ADAS Cog 11 comprehension sub-score). Assuming common variances $\sigma_0^2 = \sigma_1^2$ the anticipated upper bound of the sampling variance of the unadjusted estimator (Eq. (4)) came out to 0.64 with 238 treated and 164 control subjects ($n = 402$, 3:2), giving a power of $\geq 80\%$ to detect an effect size of 2.25 points in ADAS-Cog11 via Eq. (5).

Assuming these same parameters and a common correlation $\rho_0 = \rho_1$, numerical optimization[10] of Eq. (3) composed with Eq. (5) at the same 3:2 randomization ratio showed that $\geq 80\%$ power was attainable with only 131 control and 190 treated subjects ($n = 321$) when knowledge of the prognostic model was leveraged in the design. This represents an impressive 20% reduction in the enrollment that would have been required.

6 Discussion

Our theoretical and empirical results demonstrate that linear adjustment for a prognostic score is an effective and safe method for leveraging historical data to reduce uncertainty in randomized trials. In fact, prognostic covariate adjustment is optimal (i.e. semiparametric efficient) when there is a constant treatment effect and the prognostic model accurately predicts the conditional mean of the trial control arm. These benefits may also be exploited to design smaller trials that maintain their power.

Heuristically, the reason that prognostic covariate adjustment via a linear model improves efficiency is that the prognostic score captures nonlinear relationships between the covariates and outcome that the linear model could not otherwise exploit. This helps to "explain away" some amount of previously unexplained variability in the outcome. It is easy to show that if the prognostic score is a linear function of the coefficients (i.e. $m(X) = X^\top \alpha$) then adjusting for it in addition to the covariates cannot improve efficiency (in fact, the linear model becomes indeterminate). The prognostic model must therefore be nonlinear to provide any benefit. Our simulation results bear out this conclusion; there was no gain from prognostic covariate adjustment in the linear scenario. The lack of improvement in the covariate shift scenario may also be due to a similar phenomenon – the smaller the support of the covariate distribution, the better a linear (or constant) approximation will capture the outcome-covariate relationship.

Together, these results can be summarized as follows: adjusting for a prognostic score obtained from a nonlinear predictive model trained on a large database of historical control arm data provides near optimal

---

10 Note that Eq. (6) does not apply here because the desired randomization ratio is not 1:1. Despite this, the obtained sample size reduction is still close to ρ^2.

treatment effect estimates in randomized trials with continuous outcomes (given previously stated conditions). There has recently been tremendous growth in the availability and performance of technology for nonlinear regression modeling (i.e. supervised machine learning), particularly in the area of deep learning. The intersection of this technological development with the creation of large historical control databases provides an opportunity to use prognostic covariate adjustment to substantially improve future clinical trials.

As an alternative, one might forego the historical data altogether and use a nonlinear analysis directly on the trial data in order to account for any nonlinearities in the relationship between covariates and outcome. Several semiparametric efficient procedures exist for this kind of nonlinear/adaptive covariate adjustment [39–41]. If there is effect heterogeneity, adaptive adjustment methods may have the upper hand because the prognostic score will be less predictive of the treatment-arm outcomes. One practical advantage of adaptive adjustment is that there is no need for historical data at all, but that also implies that no information from historical data sources is used to improve the estimate (e.g. it may be easier to capture nonlinearities using large historical datasets, which could include hundreds of thousands of samples, than it is in small trial datasets).

On the other hand, linear prognostic covariate adjustment has its own set of practical advantages. Perhaps the most important of these is that the analysis is a standard linear regression once the prognostic score has been calculated. This makes it easy to explain and interpret for trialists, simple to implement with existing software, and makes the approach suitable under current regulatory guidance [17]. The estimator can be further simplified by omitting the interaction terms if little effect heterogeneity is suspected. The procedure is also modular: construction of the prognostic model may be outsourced to a group of machine learning experts, which also makes it possible to separate access to the historical and trial datasets. In fact, the historical data can be used to train a prognostic model within a privacy preserving framework with guaranteed protection of private health information [42, 43]. Lastly, our conservative variance bound (Theorem 4) makes it easy to prospectively power a trial without estimating or assuming a large number of population parameters.

It is also possible that prognostic covariate adjustment retains a statistical advantage in finite samples relative to direct nonlinear adjustment, though we have left theoretical investigation of this question to future studies. The flexibility of our procedure with respect to the prognostic model also allows for the exploitation of large proxy datasets where a surrogate outcome was measured. This includes cases where the causal contrast of interest is between two non-baseline treatments (e.g. a trial with three or more treatment levels). The use of historical data could therefore serve as an effective regularizer for learning in small samples, but more research will be necessary to make that conclusion. It should also be possible to combine the advantages of multiple procedures, i.e. to perform adaptive adjustment for a fixed prognostic model trained on historical data.

Further, we show that efficiency is usually improved to some extent even if the prognostic model is imperfect or the treatment effect is not constant. In fact, prognostic covariate adjustment can never hurt asymptotic efficiency. For some studies, including a prognostic score as an adjustment covariate could mean the difference between a null result and a clear demonstration of efficacy or harm. Moreover, the unbiasedness, type-I error rate control, and correct confidence interval coverage of prognostic covariate adjustment are inherent properties of the procedure for any choice of prognostic model that only uses information from a subject's baseline covariates.

For our efficiency improvement to hold in practice, all that is necessary is for the trial to be large enough for the asymptotic variance to be a reasonable estimate and for mild technical regularity assumptions (which underpin most asymptotic theory) to hold.[11] These assumptions are not unique to prognostic covariate adjustment and are required for almost any meaningful frequentist analysis of the data.

It should also be possible to exploit prognostic covariate adjustment as a component in other kinds of estimators (repeated measures, binary outcomes, survival models etc.). We have limited our theoretical

---

11 These are technical conditions that might allow for the exchange of integration and differentiation or guarantee the existence of a mean value. They have no practical importance or violation in almost all real-world scenarios.

discussion here to the linear model since it is so common, but a prognostic score may be used as a covariate in any analysis that allows for covariate adjustment. It remains to be seen what optimality properties are satisfied by doing prognostic covariate adjustment in each kind of analysis and under what conditions.

prognostic covariate adjustment also presents a method for optimally reducing the dimensionality of the adjustment covariate set even when performing standard linear analyses. In our empirical demonstration we included 37 covariates (and their interactions with treatment) in the linear regression model in order to present a fair comparison with prognostic covariate adjustment, but including this many adjustment covariates is rarely if ever done in practice. The prognostic score offers an opportunity to exploit the information present in all of those covariates without necessarily having to include each of them in the analysis. However, including the raw covariates by themselves helps ensure some reduction in variance even if the prognostic score has a very poor correlation with the outcome. Similarly, including treatment-covariate interaction terms can only be beneficial (in large samples) but there may be a practical limit to the number of terms an analyst is willing to include in the analysis.

Regardless of theoretical considerations, our work shows that prognostic covariate adjustment offers practical advantages even when the assumptions that guarantee some forms of optimality are violated. This is borne out by the analyses of the secondary ADL and CDR outcomes using the ADAS-Cog 11 prognostic score in our empirical demonstration. Since there is no (asymptotic) harm in including a prognostic score, it may behoove trialists working with large-enough samples to amass and exploit a variety of prognostic scores for each analysis.

We used random forests as the prognostic models in our demonstration, but alternative methods may be preferred in practice. Missing covariates, multiple longitudinal outcomes, and high-dimensional covariates (e.g. a whole genome) may be present in real trial data. Deep learning methods (and in particular generative deep learning methods) are often well suited to handle these challenges [44–47]. Deep learning methods can also exploit transfer learning to improve performance when the relevant historical data are meager [48]. Moreover, most trials actually have many outcomes of interest (e.g. multiple primary or secondary endpoints, adverse events, and biomarkers monitored for safety). Often, each of these outcomes is measured at multiple timepoints during a trial. In principle, the development of comprehensive, longitudinal predictive models of patient outcomes under standard-of-care would enable prognostic covariate adjustment to be used for each of these analyses, thereby enabling the design of studies that require fewer subjects to achieve desired operating characteristics.

In comparison to other kinds of historical borrowing methods, prognostic covariate adjustment theoretically guarantees strict type-I error rate control and confidence interval coverage in general settings. In anything but the smallest of trials, there is no need for elaborate simulations to demonstrate the trial operating characteristics (as are usually required for methods that cannot theoretically guarantee control of type I error). Moreover, we provide a simple formula in theorem 4 that can be used to calculate power prospectively while accounting for the beneficial effect of prognostic covariate adjustment.

Data availability

The data used in this study are available from the following sources, subject to their discretion.

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.

Appendix A. Mathematical results

Throughout we assume enough regularity conditions for the asymptotic normality of M-estimators to hold. The details are found in chapter 5 (thm 5.23) of van der Vaart [49].

Lemma A.1. (Rosenblum). *The influence function for the linear regression treatment effect estimator we describe in* Section 3 *is* $\psi = \psi_1 - \psi_0$ *where*

$$\psi_w = \frac{W_w}{\pi_w}(Y - \hat{\mu}_w^*(X)) + (\hat{\mu}_w^*(X) - \hat{\mu}_w^*) \tag{7}$$

and $\hat{\mu}_w^*(X) = Z_w^\top \beta^*$ *and* $\hat{\mu}_w^* = \mathbb{E}\left[\hat{\mu}_w^*(X)\right]$. *The parameters* $\hat{\beta}^*$ *are those that maximize the (model-based) likelihood in expectation (under the true law of the data). In other words,* $\hat{\mu}_w^*(X)$ *characterizes the linear model that comes as close as possible to the true conditional mean function* $\mu_w(X) = \mathbb{E}[Y_w|X]$ *and* $\hat{\mu}_w^*$ *is its mean value (averaged over X).*

This follows from results in Robins et al. [50]. An accessible presentation for the case of generalized linear models is given in Rosenblum and Laan [51].

Definition A.1. (Difference-in-means). The "difference-in-means" (or "unadjusted") estimator of $\tau = \mu_1 - \mu_0$ is $\hat{\tau}_\Delta = \hat{\mathbb{E}}\left[Y|W_1\right] - \hat{\mathbb{E}}\left[Y|W_0\right]$.

Note that throughout the appendix we omit the subscript n on estimators. E.g. τ_Δ is shorthand for $\tau_{\Delta,n}$ and our asymptotic statements refer to the sequence of estimators as n becomes large.

Lemma A.2. *The difference-in-means estimator has asymptotic variance given by*

$$n\mathbb{V}\left[\hat{\tau}_\Delta\right] \xrightarrow{p} \frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} \tag{8}$$

where $\sigma_w = \mathbb{V}\left[Y_w\right]$.

Proof. This fact is well-known. One proof follows the outline of 7 below taking $Z^\top = [1, W]$. □

Definition A.2. (ANCOVA I). The "ANCOVA I" estimator of $\tau = \mu_1 - \mu_0$ (denoted $\hat{\tau}_\mathrm{I}$) is the effect estimated using a linear regression with predictors $Z^\top = [1, W, X^\top]$ and outcome Y.

Definition A.3. (ANCOVA II). The "ANCOVA II" estimator of $\tau = \mu_1 - \mu_0$ (denoted $\hat{\tau}_\mathrm{II}$) is the effect estimated using a linear regression with predictors $Z^\top = [1, \tilde{W}, \tilde{X}^\top, \tilde{W}\tilde{X}^\top]$ and outcome $\tilde{Y}$.

The following two Theorems A.3 and A.4 are mild generalizations of or follow closely from results stated in Leon et al. [24] and Yang and Tsiatis [16]. Details are provided here for the reader's convenience.

Theorem A.3. *The ANCOVA I estimator is asymptotically unbiased for* $\tau = \mu_1 - \mu_0$ *and has asymptotic variance given by*

$$n\mathbb{V}\left[\hat{\tau}_I\right] \xrightarrow{p} \frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} + \left(\frac{1}{\pi_0\pi_1}\right)\xi^\top V\xi - 2\left(\frac{1}{\pi_0\pi_1}\right)\xi_*^\top V\xi \tag{9}$$

where $\xi = \pi_0\mathbb{C}\left[Y_0, X\right] + \pi_1\mathbb{C}\left[Y_1, X\right]$, $\xi_* = \pi_0\mathbb{C}\left[Y_1, X\right] + \pi_1\mathbb{C}\left[Y_0, X\right]$, *and* $V = \mathbb{V}\left[X\right]^{-1}$.

Proof. We begin by applying Lemma A.1. Minimization of the expected log-likelihood shows that $\hat{\beta}^* = \mathbb{E}\left[ZZ^\top\right]^{-1}\mathbb{E}\left[ZY\right]$. Some algebra[12] demonstrates

$$\hat{\beta}^* = \left[\mu_0, \tau, (V\xi)^\top\right]^\top \tag{10}$$

---

12 The identity $\mathbb{E}\left[AB\right] = \mathbb{C}\left[A, B\right] + \mathbb{E}\left[A\right]\mathbb{E}\left[B\right]$ and the fact that $W \perp Y_w, X$ and $W_w Y = W_w Y_w$ by our structural assumption (Eq. (1)) may be used to show that

$$\mathbb{E}\left[ZZ^\top\right]^{-1} = \begin{bmatrix} \frac{1}{\pi_0} + \eta^\top V\eta & -\frac{1}{\pi_0} & -V\eta \\ -\frac{1}{\pi_0} & \frac{1}{\pi_0\pi_1} & 0 \\ -V\eta & 0 & V \end{bmatrix} \qquad \mathbb{E}\left[ZY\right] = \begin{bmatrix} \mu \\ \pi_1\mu_1 \\ \mu\eta + \xi \end{bmatrix}$$

where $\eta = \mathbb{E}\left[X\right]$, $\mu = \mathbb{E}\left[Y\right]$, $\xi = \mathbb{C}\left[X, Y\right]$, and $V = \mathbb{V}\left[X\right]^{-1}$. The inverse is easiest to verify by computing and multiplying by $\mathbb{E}\left[ZZ^\top\right]$.

where $V = \mathbb{V}[X]^{-1}$, $\xi = \mathbb{C}[X, Y]$, and $\tau = \mu_1 - \mu_0$. Thus $\hat{\mu}_w^*(X) = \mu_0 + w\tau + \tilde{X}^\top V \xi = \mu_w + \tilde{X}^\top V \xi$. In this equation and from here on, let $\tilde{X} = X - \mathbb{E}[X]$. So clearly $\hat{\mu}_w^* = \mu_w$. Then, from Eq. (7),

$$\psi_w = \frac{W_w}{\pi_w}(Y - \mu_w) - \frac{\tilde{W}_w}{\pi_w}\underbrace{(\tilde{X}^\top V \xi)}_{-h_w(X)} \tag{11}$$

Where $\tilde{W}_w = W_w - \pi_w$. An application of A.1 and some algebra gives

$$\psi_I = \underbrace{\underbrace{\frac{W_1}{\pi_1}(Y - \mu_1)}_{\psi_{1,\Delta}} - \underbrace{\frac{W_0}{\pi_0}(Y - \mu_0)}_{\psi_{0,\Delta}}}_{\psi_\Delta} - \underbrace{(W_1 - \pi_1)\underbrace{\frac{(\tilde{X}^\top V \xi)}{\pi_0 \pi_1}}_{-h(X)}}_{\phi} \tag{12}$$

It is known that all regular and asymptotically linear estimators of the treatment effect have an influence function of this form with $h(X)$ dependent on the choice of estimator [24, 26].

By the theory of influence functions, our estimator has a limiting distribution [26]

$$\sqrt{n}(\hat{\tau}_I - \tau) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left[\psi_I^2\right]\right) \tag{13}$$

The asymptotic variance of $\hat{\tau}_I$ is thus $\mathbb{E}\left[\psi_I^2\right] = \mathbb{E}\left[(\psi_\Delta - \phi)^2\right] = \mathbb{E}\left[\psi_\Delta^2\right] - 2\mathbb{E}\left[\psi_\Delta \phi\right] + \mathbb{E}\left[\phi^2\right]$. The first term is the variance of the influence function for the difference-in-means (also called "unadjusted") estimator. It may be verified that this evaluates to $\mathbb{E}\left[\psi_\Delta^2\right] = \frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1}$ where $\sigma_w^2 = \mathbb{V}[Y_w]$. The variance of ϕ is

$$\mathbb{E}\left[\phi^2\right] = \mathbb{E}\left[\left(\frac{W_1 - \pi_1}{\pi_0 \pi_1}\tilde{X}^\top V \xi\right)^2\right] \tag{14}$$

$$= \frac{\mathbb{E}\left[(W_1 - \pi_1)^2\right]}{\pi_0^2 \pi_1^2}\xi^\top V \mathbb{E}\left[\tilde{X}\tilde{X}^\top\right] V \xi \tag{15}$$

$$= \left(\frac{1}{\pi_0 \pi_1}\right)\xi^\top V \xi \tag{16}$$

The covariance of the two terms involves the expectations $\mathbb{E}\left[(Y_w - \mu_w)\tilde{X}\right] = \mathbb{C}[Y_w, X] = \xi_w$ (note that $\xi = \pi_0 \xi_0 + \pi_1 \xi_1$):

$$\mathbb{E}\left[\psi_\Delta \phi\right] = \mathbb{E}\left[\psi_{1,\Delta}\phi\right] - \mathbb{E}\left[\psi_{0,\Delta}\phi\right] \tag{17}$$

$$= \frac{1}{\pi_1}\xi_1^\top V \xi - \frac{-1}{\pi_0}\xi_0^\top V \xi \tag{18}$$

$$= \left(\frac{1}{\pi_0 \pi_1}\right)\xi_*^\top V \xi \tag{19}$$

where we have introduced $\xi_* = \pi_1 \xi_0 + \pi_0 \xi_1$. Assembling obtains the desired result.   □

Corollary A.3.1. *When $X \in R$ (a single covariate), a consistent estimate of the sampling variance $\mathbb{V}[\hat{\tau}_I]$ is*

$$\hat{v}_I^2 = \frac{\hat{\sigma}_0^2}{n_0} + \frac{\hat{\sigma}_1^2}{n_1} + \frac{n_0 n_1}{n}\left(\frac{\hat{\rho}_0 \hat{\sigma}_0}{n_1} + \frac{\hat{\rho}_1 \hat{\sigma}_1}{n_0}\right)^2 - 2\frac{n_0 n_1}{n}\left(\frac{\hat{\rho}_0 \hat{\sigma}_0}{n_1} + \frac{\hat{\rho}_1 \hat{\sigma}_1}{n_0}\right)\left(\frac{\hat{\rho}_0 \hat{\sigma}_0}{n_0} + \frac{\hat{\rho}_1 \hat{\sigma}_1}{n_1}\right) \tag{20}$$

where $\rho_w = \mathbb{C}[Y_w, X] / \sqrt{\mathbb{V}[X]\mathbb{V}[Y_w]}$ and the "hat" quantities are any consistent estimates of their respective population parameters.

Proof. This follows from the definitions and Slutsky's theorem. □

Corollary A.3.2. *If either $\pi_0 = \pi_1$ or $\xi_0 = \xi_1$, then*

$$n\mathbb{V}\left[\hat{\tau}_I\right] \xrightarrow{p} \frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} - \left(\frac{1}{\pi_0\pi_1}\right)\xi_*^\top V\xi_* \tag{21}$$

Theorem A.4. *The ANCOVA II estimator is asymptotically unbiased for $\tau = \mu_1 - \mu_0$ and has asymptotic variance given by*

$$n\mathbb{V}\left[\hat{\tau}_{II}\right] \xrightarrow{p} \frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} - \left(\frac{1}{\pi_0\pi_1}\right)\xi_*^\top V\xi_* \tag{22}$$

Proof. Arguments similar to those in Theorem A.3 show that the influence function for the GLM marginal effect estimator with this specification is identical to Eq. (12) except that $\xi = \pi_0\xi_0 + \pi_1\xi_1$ is replaced by $\xi_* = \pi_1\xi_0 + \pi_0\xi_1$. Specifically $\psi_{II} = \psi_{1,II} - \psi_{0,II}$ with

$$\psi_{w,II} = \frac{W_w}{\pi_w}(Y - \mu_w) - \frac{\tilde{W}_w}{\pi_w}\underbrace{\left(\tilde{X}^\top V\xi_*\right)}_{-h_w(X)} \tag{23}$$

The result follows from proceeding along the outline of Theorem A.3. □

Corollary A.4.1. *When $X \in R$ (a single covariate), a consistent estimate of the sampling variance $\mathbb{V}\left[\hat{\tau}_{II}\right]$ is*

$$\hat{v}_{II}^2 = \frac{\hat{\sigma}_0^2}{n_0} + \frac{\hat{\sigma}_1^2}{n_1} - \frac{n_0 n_1}{n}\left(\frac{\hat{\rho}_0\hat{\sigma}_0}{n_0} + \frac{\hat{\rho}_1\hat{\sigma}_1}{n_1}\right)^2 \tag{24}$$

Corollary A.4.2. *Adding covariates to the ANCOVA II estimator can only decrease its asymptotic variance.*

Proof. Consider using covariates X with variance Σ_x and covariance with Y_w of $\xi_{w,x}$ versus a set of covariates $[X, M]$ ($M \in \mathbb{R}$) such that M is not a linear combination of the variables in X. Let $\mathbb{C}[X, M] = \zeta$, $\mathbb{V}[M] = \sigma_m^2$ and $\mathbb{C}[Y_w, M] = \xi_{w,m}$. Let $\xi_{m*} = \pi_0\xi_{1,m} + \pi_1\xi_{0,m}$ and $\xi_{x*} = \pi_0\xi_{1,x} + \pi_1\xi_{0,x}$. From Eq. (22) and some matrix algebra the difference in asymptotic variance between these two estimators is

$$-\left(\frac{1}{\pi_0\pi_1}\right)\frac{\left(\xi_m - \xi_x^\top\Sigma_x^{-1}\zeta\right)^2}{\sigma_m^2 - \zeta^\top\Sigma_x^{-1}\zeta} \tag{25}$$

The denominator must be positive because $\mathbb{V}[X, M] \geq 0$, $\mathbb{V}[X] \geq 0$ implies $\det(\mathbb{V}[X, M]) = \det\left(\Sigma_x^{-1}\right)\left(\sigma_m^2 - \zeta^\top\Sigma_x^{-1}\zeta\right) \geq 0$. □

Theorem A.5. *ANCOVA II is a more efficient estimator than ANCOVA I or difference-in-means. ANCOVA I may or may not be more efficient than difference-in-means (unless $\pi_0 = \pi_1 = 0.5$ or $\xi_0 = \xi_1$, in which case it is as efficient as ANCOVA II). In a slight abuse of notation,*

$$\mathbb{V}\left[\hat{\tau}_{II}\right] \leq \mathbb{V}\left[\hat{\tau}_I\right] \tag{26}$$

$$\mathbb{V}\left[\hat{\tau}_{II}\right] \leq \mathbb{V}\left[\hat{\tau}_\Delta\right] \tag{27}$$

$$\mathbb{V}\left[\hat{\tau}_I\right] \not\leq \mathbb{V}\left[\hat{\tau}_\Delta\right] \tag{28}$$

$$\pi_0 = \pi_1 \Rightarrow \mathbb{V}\left[\hat{\tau}_I\right] = \mathbb{V}\left[\hat{\tau}_{II}\right] \tag{29}$$

Proof. $\mathbb{V}[\hat{\tau}_{\mathrm{II}}] \leq \mathbb{V}[\hat{\tau}_{\mathrm{I}}]$ because Eq. (22) subtracted from Eq. (9) is $(V^{1/2}(\xi - \xi_*))^2/(\pi_0\pi_1) \geq 0$. $\mathbb{V}[\hat{\tau}_{\mathrm{II}}] \leq \mathbb{V}[\hat{\tau}_\Delta]$ is self-evident from Eq. (22). To show $\mathbb{V}[\hat{\tau}_{\mathrm{I}}] \not\leq \mathbb{V}[\hat{\tau}_\Delta]$ we rely on an example: using $X \in R$ with $\pi_1 = 5/6$ (so $\pi_0 = 1/6$), $\xi_1 = 4$ and $\xi_0 = 1$ in Eq. (9) gives a positive addition to $\mathbb{V}[\hat{\tau}_\Delta]$. $\qquad\square$

Lemma A.6. *Consider using the ANCOVA II estimator with an arbitrary (multivariate) transformation of the covariates $f(X)$ in place of the raw covariates X. Among all fixed transformations $f(X)$, the transformation $[\mu_0(X), \mu_1(X)]^\top$ is optimal in terms of efficiency. Furthermore, the estimator is semiparametric efficient: the ANCOVA II estimator with $[\mu_0(X), \mu_1(X)]^\top$ used as the vector of covariates has the lowest possible asymptotic variance among all regular and asymptotically linear estimators with access to the covariates X.*

Consider replacing X in the interacted linear model (ANCOVA II) with an arbitrary fixed (possibly multivariate) function of the covariates $f(X)$. By Eq. (23) and our definitions of ξ_* and V the influence function for this estimator is $\psi = \psi_1 - \psi_0$ with

$$\psi_w = \frac{W_w}{\pi_w}(Y - \mu_w) - \frac{\tilde{W}_w}{\pi_w}\underbrace{\left((f(X) - \mathbb{E}[f(X)])^\top V_f \xi_{f*}\right)}_{-h_w(X)} \tag{30}$$

where $\xi_{f*} = \pi_1 \mathbb{C}[Y_0, f(X)] + \pi_0 \mathbb{C}[Y_1, f(X)]$ and $V_f = \mathbb{V}[f(X)]^{-1}$. Consider now using the special transformation $f(X)^\top = [\mu_0(X), \mu_1(X)]$ where $\mu_w(X) = \mathbb{E}[Y_w|X]$. Note that $\mathbb{C}[Y_w, \mu_w(X)] = \mathbb{V}[\mu_w(X)]$ and $\mathbb{C}[Y_1, \mu_0(X)] = \mathbb{C}[\mu_1(X), \mu_0(X)]$ by an orthogonal decomposition of Y_w.[13] Plugging these in and performing the appropriate algebra shows that $V_f \xi_{f*}$ in this case is $[\pi_1, \pi_0]^\top$ so $h_w(X)$ in 30 is $\pi_0(\mu_1(X) - \mu_1) + \pi_1(\mu_0(X) - \mu_0)$. A little algebra shows

$$\psi = \psi_1 - \psi_0 \tag{31}$$

$$= \frac{W_1}{\pi_1}(Y - \mu_1) - \frac{W_0}{\pi_0}(Y - \mu_0) - (W_1 - \pi_1)\left[\frac{\pi_0(\mu_1(X) - \mu_1) + \pi_1(\mu_0(X) - \mu_0)}{\pi_0\pi_1}\right] \tag{32}$$

---

[13] Let $R = Y_w - \mathbb{E}[Y_w|X]$ be the part of Y_w orthogonal to $\mathbb{E}[Y_w|X] = \mu_w(X)$ so that $Y_w = R + \mu_w(X)$. Note that

$$\mathbb{C}[Y_w, f(X)] = \mathbb{C}[R, f(X)] + \mathbb{C}[\mu_w(X), f(X)]$$

Now we prove a known result that $\mathbb{C}[R, f(X)] = 0$ for any function f:

$$\mathbb{C}[R, f(X)] = \mathbb{E}\left[(R - \underbrace{\mathbb{E}[R]}_{0})(\underbrace{f(X) - \mathbb{E}[f(X)]}_{\tilde{f}(X)})\right]$$

$$= \mathbb{E}[(Y_w - \mathbb{E}[Y_w|X])\tilde{f}(X)]$$

$$= \mathbb{E}[\mathbb{E}[(Y_w - \mathbb{E}[Y_w|X])\tilde{f}(X)|X]]$$

$$= \mathbb{E}[(\mathbb{E}[Y_w|X] - \mathbb{E}[Y_w|X])\tilde{f}(X)]$$

$$= 0$$

The result is precisely the *efficient influence function* for the treatment effect [24, 26]. It is known that no regular and asymptotically linear (RAL) estimator (which essentially all practical and reasonable estimators are) can be more efficient than any estimator with this influence function.

Corollary A.6.1. *Presume a constant treatment effect:* $\mu_1(X) = \mu_0(X) + \tau$. *Then the ANCOVA II analysis that uses* $\mu_0(X)$ *in the role of X has the lowest possible asymptotic variance among all regular and asymptotically linear estimators with access to the covariates X.*

Proof. $\mu_1(X) = \mu_0(X) + \tau$ implies $\mathbb{C}\left[\mu_0(X), \mu_1 X\right] = \mathbb{V}\left[\mu_0(X)\right] = \mathbb{V}\left[\mu_1(X)\right]$. Following the outline for the proof of Lemma A.6 above shows that the influence function for the ANCOVA II estimator with $\mu_0(X)$ as the single covariate is

$$\psi = \frac{W_1}{\pi_1}(Y - \mu_1) - \frac{W_0}{\pi_0}(Y - \mu_0) - (W_1 - \pi_1)\left[\frac{\mu_0(X) - \mu_0}{\pi_0 \pi_1}\right] \tag{33}$$

which is the same as the efficient influence function when $\mu_1(X) = \mu_0(X) + \tau$. □

Corollary A.6.2. Corollary A.6.1 *also holds when the ANCOVA II estimator is replaced by the ANCOVA I estimator.*

Proof. Theorem A.5 establishes that ANCOVA I is as efficient as ANCOVA II when $\mathbb{C}\left[m(X), Y_0\right] = \xi_0 = \xi_1 = \mathbb{C}\left[m(X), Y_1\right]$. A constant treatment effect means that $\mu_1(X) = \mu_0(X) + \tau$ and this ensures the equality of the covariances. □

The following lemma is required for the proof that proceeds it.

Lemma A.7. *Let* $f: \mathcal{X} \to \mathbb{R}$ *be a bounded function on a compact set* $\mathcal{X}$ *and let* $\hat{f}_n: \mathcal{X} \to \mathbb{R}$ *be a sequence of uniformly bounded random functions such that* $|f(X) - \hat{f}_n(X)| \xrightarrow{L_2} 0$. *Let* $X \in \mathcal{X}$ *be a random variable independent of* $\hat{f}_n$. *Then* $\mathbb{E}_X\left[\hat{f}_n(X)\right] \xrightarrow{p} \mathbb{E}\left[f(X)\right]$, $\mathbb{C}_X\left[f(X), \hat{f}_n(X)\right] \xrightarrow{p} \mathbb{V}\left[f(X)\right]$, *and* $\mathbb{V}_X\left[\hat{f}_n(X)\right] \xrightarrow{p} \mathbb{V}\left[f(X)\right]$.

Proof. $\hat{f}_n$ and X are independent, so let their joint distribution factor into P'_n and P. Now

$$\int \left(\mathbb{E}_X\left[\hat{f}_n(X)\right] - \mathbb{E}\left[f(X)\right]\right)^2 dP'_n = \int \left[\int \hat{f}_n(X)dP - \int f(X)dP\right]^2 dP'_n$$

$$= \int \left[\int \hat{f}_n(X) - f(X)dP\right]^2 dP'_n$$

$$\leq \int \int \left[\hat{f}_n(X) - f(X)\right]^2 dPdP'_n \quad \text{(Jensen's inequality)}$$

$$\to 0$$

The final convergence holds by our assumption that $|f(X) - \hat{f}_n(X)| \xrightarrow{L_2} 0$. This shows $\mathbb{E}_X\left[\hat{f}_n(X)\right] \xrightarrow{L_2} \mathbb{E}\left[f(X)\right]$ and convergence in probability follows.

Taking advantage of the fact that $|f|, |f_n| \leq b$ are bounded we can make similar arguments to show that $\mathbb{E}_X\left[f(X)\hat{f}_n(X)\right] \xrightarrow{p} \mathbb{E}_X\left[f(X)^2\right]$ and $\mathbb{E}_X\left[\hat{f}_n(X)^2\right] \xrightarrow{p} \mathbb{E}_X\left[f(X)^2\right]$. Slutsky's theorem and the definition of covariance and variance then imply $\mathbb{C}_X\left[f(X), \hat{f}_n(X)\right] \xrightarrow{p} \mathbb{C}\left[f(X), f(X)\right]$ and $\mathbb{V}_X\left[\hat{f}_n(X)\right] \xrightarrow{p} \mathbb{V}\left[f(X)\right]$ as desired. □

Corollary A.7.1. *Let* $\mathbb{V}_X\left[\hat{f}_n(X)\right] > \epsilon > 0$. *Under the conditions of the above lemma,* $\left| f(x) - \hat{f}_n(x) \frac{\mathbb{C}_X\left[f(X), \hat{f}_n(X)\right]}{\mathbb{V}_X\left[\hat{f}_n(X)\right]} \right| \xrightarrow{L_2} 0$.

Proof. Let $B_n = \frac{\mathbb{C}_X\left[f(X), \hat{f}_n(X)\right]}{\mathbb{V}_X\left[\hat{f}_n(X)\right]}$. By the above lemma, our assumption that $\mathbb{V}_X\left[\hat{f}_n(X)\right] > \epsilon > 0$, and Slutsky's theorem, $B_n \xrightarrow{p} 1$. Together with the uniform bound on $\mathbb{V}_X\left[\hat{f}_n(X)\right]$ and Cauchy-Schwarz this is also enough to ensure that $(1 - B_n) \xrightarrow{L_2} 0$.

Now note $\left|f(x) - \hat{f}_n(x) B_n\right| \leq \left|f(x) - \hat{f}_n(x)\right| + b\left|1 - B_n\right|$ by the triangle inequality and the fact that $|\hat{f}_n(x)| < b$. Thus

$$\mathbb{E}\left[(f(X) - \hat{f}_n(X) B_n)^2\right] = \underbrace{\mathbb{E}\left[(f(X) - \hat{f}_n(X))^2\right]}_{\to 0 \;\;\text{(by assumption)}} + \underbrace{b^2 \mathbb{E}\left[(1 - B_n)^2\right]}_{\to 0 \;\;\text{(shown above)}} + b\mathbb{E}\left[\underbrace{(f(X) - f_n(X))}_{\leq 2b}(1 - B_n)\right]$$

$$= o(1) + o(1) + 2b^2 \mathbb{E}\left[1 - B_n\right]$$

$$\to 0$$

as desired. $\qquad\square$

Theorem A.8. *Presume X has compact support and there is a constant treatment effect: $\mu_1(X) = \mu_0(X) + \tau$ with $|\mu_0(x)| < b$ bounded. Let $m(x)$ be a (random) function learned from the external data $(\mathbf{Y}', \mathbf{X}')_{n'}$ such that $|m(x)| < b$ is also bounded and $|m(X) - \mu_0(X)| \xrightarrow{L_2} 0$ so that the learned model approaches the truth in MSE as $n' \to \infty$. If the number of trial samples n grows in tandem with the size of the historical data n' (i.e. $n = O(n')$), then the ANCOVA II analysis that uses the learned model $m(X)$ in the role of X has the lowest possible asymptotic variance among all regular and asymptotically linear estimators with access to the covariates X.*

Proof. Define our estimator of interest as the ANCOVA II estimator that uses the learned model $m(X)$ in place of the covariates X if $m(X)$ is not numerically constant up to some machine precision and otherwise as the difference-in-means estimator. Denote this estimator $\hat{\tau}$ (omitting the II subscript for the duration of this proof). Define the "oracle" estimator as the equivalent estimator that uses the true conditional mean $\mu_0(X)$ instead of the estimate $m(X)$ and denote this estimator $\hat{\tau}^*$. The oracle estimator is obviously infeasible in practice because $\mu_0(\cdot)$ is not known. Corollary A.6.1 proves that the oracle estimator is semiparametric efficient (i.e. has the lowest possible asymptotic variance among regular and asymptotically linear estimators). Thus, letting v_*^2 denote the optimal asymptotic variance, we have that $\sqrt{n}(\hat{\tau}^* - \tau) \rightsquigarrow N\left(0, v_*^2\right)$. If we can show that $\sqrt{n}(\hat{\tau} - \hat{\tau}^*) \xrightarrow{p} 0$, then Slutsky's theorem and the delta method imply that $\hat{\tau}$ has the same asymptotic properties as $\hat{\tau}^*$, i.e. $\sqrt{n}(\hat{\tau} - \tau) \rightsquigarrow N\left(0, v_*^2\right)$. 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 because the two are asymptotically equivalent.

Showing $\sqrt{n}(\hat{\tau} - \hat{\tau}^*) \xrightarrow{p} 0$ requires an intermediate estimator that is asymptotically equivalent to $\hat{\tau}$. Using the assumption of the constant effect and Eq. (23) from Theorem A.4 we can show (with an application of the law of total variance) that the influence function for $\hat{\tau}$ using some fixed $m(\cdot)$ is $\psi = \psi_1 - \psi_0$ with

$$\psi_w = \frac{W_w}{\pi_w}(Y - \mu_w) - \frac{\tilde{W}_w}{\pi_w}\left((m(X) - \mathbb{E}_X[m(X)])\frac{\mathbb{C}_X[m(X), \mu_0(X)]}{\mathbb{V}_X[m(X)]}\right) \qquad (34)$$

where $\mathbb{E}_X\left[m(X)\right]$ denotes that the expectation (or variance or covariance) is taken only with respect to X, i.e. $m(\cdot)$ is considered fixed.

Let $\check{\tau} = \hat{\mathbb{E}}\left[\psi + \tau\right]$ and let $\check{\tau}^* = \hat{\mathbb{E}}\left[\psi^* + \tau\right]$ where ψ^* is the influence function above with $\mu_0(\cdot)$ substituted for $m(\cdot)$. Note that $\hat{\tau}$ and $\check{\tau}$ share the same influence function so we must have that $\sqrt{n}(\hat{\tau} - \check{\tau}) \xrightarrow{p} 0$. Similarly, $\sqrt{n}(\hat{\tau}^* - \check{\tau}^*) \xrightarrow{p} 0$. Therefore if $\sqrt{n}(\check{\tau} - \check{\tau}^*) \xrightarrow{p} 0$, then we have $\sqrt{n}(\hat{\tau} - \hat{\tau}^*) \xrightarrow{p} 0$ as desired. This is useful because the estimator $\check{\tau}$ and its oracle counterpart $\check{\tau}^*$ are easier to work with.

To wit, consider the difference $\check{\tau} - \check{\tau}^* = \widehat{\mathbb{E}}\left[(\psi_1 - \psi_0) - (\psi_1^* - \psi_0^*)\right]$. So all we need to show the desired convergence $\sqrt{n}(\check{\tau} - \check{\tau}^*) \xrightarrow{p} 0$ is to show $\sqrt{n}\widehat{\mathbb{E}}\left[\psi_w - \psi_w^*\right] \xrightarrow{p} 0$. Expanding,

$$\widehat{\mathbb{E}}\left[\psi_w - \psi_w^*\right] = \frac{1}{n}\sum_i^n \frac{\tilde{W}_{w,i}}{\pi_w} \left((\mu_0(X_i) - \mathbb{E}_X\left[\mu_0(X)\right]) \frac{\mathbb{C}_X\left[\mu_0(X), \mu_0(X)\right]}{\mathbb{V}_X\left[\mu_0(X)\right]} \right.$$

$$\left. - (m(X_i) - \mathbb{E}_X\left[m(X)\right]) \frac{\mathbb{C}_X\left[m(X), \mu_0(X)\right]}{\mathbb{V}_X\left[m(X)\right]} \right)$$

$$= \frac{1}{n}\sum_i^n \frac{\tilde{W}_{w,i}}{\pi_w} \left(\mu_0(X_i) - m(X_i)B \right) - \frac{1}{n}\sum_i^n \frac{\tilde{W}_{w,i}}{\pi_w} \left(\mu_0 - mB \right) \tag{35}$$

where we've abbreviated $B = \frac{\mathbb{C}_X\left[m(X), \mu_0(X)\right]}{\mathbb{V}_X\left[m(X)\right]}$ and $m = \mathbb{E}_X\left[m(X)\right]$. Our plan is to show that both of these terms L^2-converge to 0 at the $\sqrt{n}$ rate so that they both converge in probability in that rate, as does their sum (which is what we want). To show L^2 convergence for the first term, we must consider the expression

$$\mathbb{E}\left[\left(\sqrt{n}\frac{1}{n}\sum_i^n \frac{\tilde{W}_w}{\pi_w}(\mu_0(X) - m(X)B) \right)^2\right] \tag{36}$$

And show it converges to 0. Recalling that m itself is random (depends on the external data $(X'Y')$), but independent of the trial data (X, W, Y), note that we can treat $m(\cdot)$ as if it were a fixed function and B as a fixed constant if we condition on the external data. After conditioning, the quantity inside the parentheses is IID and has mean zero because its $\mu_0(X) - m(X)B$ and $\tilde{W}_w$ (by randomization) and because $\mathbb{E}\left[\tilde{W}_w\right] = 0$. Therefore the quantity above is

$$\mathbb{E}\left[n\mathbb{E}\left[\left(\frac{1}{n}\sum_i^n \frac{\tilde{W}_w}{\pi_w}(\mu_0(X) - m(X)B) \right)^2 \middle| X', Y'\right]\right] = \mathbb{E}\left[n\mathbb{V}\left[\left(\frac{1}{n}\sum_i^n \frac{\tilde{W}_w}{\pi_w}(\mu_0(X) - m(X)B) \right) \middle| X', Y'\right]\right]$$

$$= \mathbb{E}\left[\frac{n}{n}\mathbb{V}\left[\left(\frac{\tilde{W}_w}{\pi_w}(\mu_0(X) - m(X)B) \right) \middle| X', Y'\right]\right]$$

$$= \frac{1 - \pi_w}{\pi_w}\mathbb{E}\left[\left(\mu_0(X) - m(X)B \right)^2\right] \tag{37}$$

where we've used the fact that the summands are IID to pass the variance through the sum and effectively gain the $1/n$ required to cancel the n. The same argument shows that the equivalent for the second term in Eq. (35) is $\frac{1-\pi_w}{\pi_w}\mathbb{E}\left[(\mu_0 - mB)^2\right]$ (note m and B are random here).

To complete the proof we invoke Corollary A.7.1 in combination with our assumptions $|m(x)| < b, |\mu_0(x)| < b$ and $|m(X) - \mu_0(X)| \xrightarrow{L_2} 0$ to arrive at the fact that $|m(x)B - \mu_0(x)| \xrightarrow{L_2} 0$ and $|mB - \mu_0| \xrightarrow{L_2} 0$. The condition that $\mathbb{V}_X\left[\hat{f}_n(X)\right]$ in Corollary A.7.1 is automatically satisfied because we only include the prognostic score in the regression if it has nonzero variance. Thus the expectations $\frac{1-\pi_w}{\pi_w}\mathbb{E}\left[(\mu_0(X) - m(X)B)^2\right]$ and $\frac{1-\pi_w}{\pi_w}\mathbb{E}\left[(\mu_0 - mB)^2\right]$ converge to 0 as desired. □

Corollary A.8.1. Theorem A.8 *also holds for the ANCOVA I estimator.*

Proof. In the case of a constant treatment effect ANCOVA I and ANCOVA II have the same asymptotic variance (Theorem A.5). The result follows immediately. □

Appendix B. Estimating σ_w^2 and ρ_w for power calculations

One method for obtaining estimates for the marginal potential outcome variances (σ_w^2) and potential outcome-prognostic score correlations (ρ_w) is to use prior data, for example data from the placebo control arm of a previous trial performed on a similar population (separate from the data used to train the prognostic model). In this case we presume we have access to a vector $\boldsymbol{Y}'' = \begin{bmatrix} Y_1'' & \dots & Y_{n''}'' \end{bmatrix}$ of outcomes for these subjects and their corresponding prognostic scores $\boldsymbol{M}'' = \begin{bmatrix} M_1'' & \dots & M_{n''}'' \end{bmatrix}$, calculated by applying the prognostic model m to each subject's vector of baseline covariates X, i.e. $M_i'' = m\left(X_i''\right)$.

The control-arm marginal outcome variance σ_0^2 can be estimated with the usual estimator

$$\hat{\sigma}_0^2 = \frac{1}{n'' - 1} \sum \left(Y_i'' - \bar{Y}''\right)^2$$

The correlation ρ_0 between M'' and Y'' can be estimated by

$$\hat{\rho}_0 = \frac{\sum \left(Y_i'' - \bar{Y}''\right)\left(M_i'' - \bar{M}''\right)}{\sqrt{\sum \left(Y_i'' - \bar{Y}''\right)^2 \sum \left(M_i'' - \bar{M}''\right)^2}}$$

which is the usual sample correlation coefficient. These values may be inflated (σ_0^2) or deflated (ρ_0) in order to provide more conservative estimates of power.

The corresponding values for the treatment arm can rarely be estimated from data because treatment-arm data for the experimental treatment is likely to be scarce or unavailable. It is therefore prudent to assume $\sigma_0^2 = \sigma_1^2$ and $\rho_0 = \rho_1$, the latter which holds exactly if the effect of treatment is constant across the population. It may also be prudent (and conservative) to assume a slightly higher value for σ_1^2 and a slightly smaller value for ρ_1 relative to their control-arm counterparts in the absence of data to the contrary.

Appendix C. Additional simulation results

Here we detail a full set of simulation results using additional specifications for the regression estimators (Figure 1). "Covariates" indicates whether the raw covariates were adjusted for. "Prognostic score" indicates whether any prognostic score was used, and, if so, whether it was estimated from a training dataset or whether the true value was used. "Interactions" specifies whether treatment × (covariates and/or prognostic score) interactions were used. "SE" indicates the standard deviation of the mean squared error.

Scenario	Covariates	Prognostic score	Interaction	MSE	SE
Baseline	False	None	True	7.64×10^{-2}	1.08×10^{-3}
Baseline	False	None	False	7.64×10^{-2}	1.08×10^{-3}
Baseline	False	Estimated	True	1.76×10^{-2}	2.46×10^{-4}
Baseline	False	Estimated	False	1.75×10^{-2}	2.45×10^{-4}
Baseline	False	Oracle	True	7.69×10^{-3}	1.09×10^{-4}
Baseline	False	Oracle	False	7.69×10^{-3}	1.09×10^{-4}
Baseline	True	None	True	5.07×10^{-2}	7.18×10^{-4}
Baseline	True	None	False	5.04×10^{-2}	7.14×10^{-4}
Baseline	True	Estimated	True	1.74×10^{-2}	2.46×10^{-4}
Baseline	True	Estimated	False	1.73×10^{-2}	2.44×10^{-4}
Baseline	True	Oracle	True	7.85×10^{-3}	1.11×10^{-4}
Baseline	True	Oracle	False	7.85×10^{-3}	1.11×10^{-4}
Surrrogate	False	None	True	7.47×10^{-2}	1.05×10^{-3}
Surrrogate	False	None	False	7.47×10^{-2}	1.05×10^{-3}
Surrrogate	False	Estimated	True	4.05×10^{-2}	5.69×10^{-4}
Surrrogate	False	Estimated	False	4.03×10^{-2}	5.66×10^{-4}

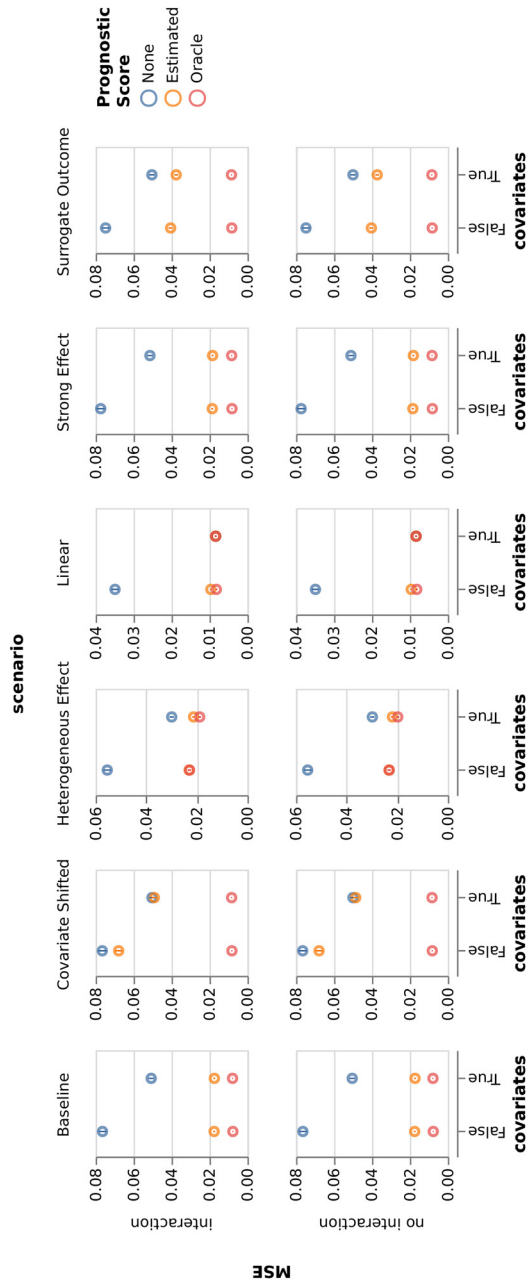| Scenario	Covariates	Prognostic score	Interaction	MSE	SE
Surrrogate	False	Oracle	True	8.25×10^{-3}	1.18×10^{-4}
Surrrogate	False	Oracle	False	8.24×10^{-3}	1.18×10^{-4}
Surrrogate	True	None	True	5.03×10^{-2}	7.09×10^{-4}
Surrrogate	True	None	False	5.00×10^{-2}	7.04×10^{-4}
Surrrogate	True	Estimated	True	3.75×10^{-2}	5.27×10^{-4}
Surrrogate	True	Estimated	False	3.72×10^{-2}	5.23×10^{-4}
Surrrogate	True	Oracle	True	8.41×10^{-3}	1.20×10^{-4}
Surrrogate	True	Oracle	False	8.41×10^{-3}	1.20×10^{-4}
Shifted	False	None	True	7.65×10^{-2}	1.10×10^{-3}
Shifted	False	None	False	7.65×10^{-2}	1.10×10^{-3}
Shifted	False	Estimated	True	6.79×10^{-2}	9.62×10^{-4}
Shifted	False	Estimated	False	6.79×10^{-2}	9.62×10^{-4}
Shifted	False	Oracle	True	8.20×10^{-3}	1.15×10^{-4}
Shifted	False	Oracle	False	8.20×10^{-3}	1.15×10^{-4}
Shifted	True	None	True	5.03×10^{-2}	7.11×10^{-4}
Shifted	True	None	False	5.00×10^{-2}	7.05×10^{-4}
Shifted	True	Estimated	True	4.91×10^{-2}	6.97×10^{-4}
Shifted	True	Estimated	False	4.86×10^{-2}	6.90×10^{-4}
Shifted	True	Oracle	True	8.34×10^{-3}	1.17×10^{-4}
Shifted	True	Oracle	False	8.34×10^{-3}	1.17×10^{-4}
Strong	False	None	True	7.73×10^{-2}	1.08×10^{-3}
Strong	False	None	False	7.73×10^{-2}	1.08×10^{-3}
Strong	False	Estimated	True	1.85×10^{-2}	2.65×10^{-4}
Strong	False	Estimated	False	1.85×10^{-2}	2.64×10^{-4}
Strong	False	Oracle	True	8.16×10^{-3}	1.16×10^{-4}
Strong	False	Oracle	False	8.16×10^{-3}	1.16×10^{-4}
Strong	True	None	True	5.14×10^{-2}	7.18×10^{-4}
Strong	True	None	False	5.11×10^{-2}	7.13×10^{-4}
Strong	True	Estimated	True	1.84×10^{-2}	2.62×10^{-4}
Strong	True	Estimated	False	1.82×10^{-2}	2.59×10^{-4}
Strong	True	Oracle	True	8.33×10^{-3}	1.18×10^{-4}
Strong	True	Oracle	False	8.32×10^{-3}	1.18×10^{-4}
Linear	False	None	True	3.49×10^{-2}	4.83×10^{-4}
Linear	False	None	False	3.49×10^{-2}	4.83×10^{-4}
Linear	False	Estimated	True	9.64×10^{-3}	1.38×10^{-4}
Linear	False	Estimated	False	9.64×10^{-3}	1.38×10^{-4}
Linear	False	Oracle	True	8.20×10^{-3}	1.16×10^{-4}
Linear	False	Oracle	False	8.20×10^{-3}	1.16×10^{-4}
Linear	True	None	True	8.37×10^{-3}	1.18×10^{-4}
Linear	True	None	False	8.37×10^{-3}	1.18×10^{-4}
Linear	True	Estimated	True	8.39×10^{-3}	1.19×10^{-4}
Linear	True	Estimated	False	8.39×10^{-3}	1.19×10^{-4}
Linear	True	Oracle	True	8.37×10^{-3}	1.18×10^{-4}
Linear	True	Oracle	False	8.37×10^{-3}	1.18×10^{-4}
Heterogeneous	False	None	True	5.54×10^{-2}	7.76×10^{-4}
Heterogeneous	False	None	False	5.54×10^{-2}	7.76×10^{-4}
Heterogeneous	False	Estimated	True	2.30×10^{-2}	3.23×10^{-4}
Heterogeneous	False	Estimated	False	2.32×10^{-2}	3.25×10^{-4}
Heterogeneous	False	Oracle	True	2.29×10^{-2}	3.20×10^{-4}
Heterogeneous	False	Oracle	False	2.32×10^{-2}	3.24×10^{-4}
Heterogeneous	True	None	True	2.99×10^{-2}	4.30×10^{-4}
Heterogeneous	True	None	False	2.98×10^{-2}	4.29×10^{-4}
Heterogeneous	True	Estimated	True	2.13×10^{-2}	3.01×10^{-4}
Heterogeneous	True	Estimated	False	2.19×10^{-2}	3.08×10^{-4}
Heterogeneous	True	Oracle	True	1.89×10^{-2}	2.69×10^{-4}
Heterogeneous	True	Oracle	False	1.98×10^{-2}	2.81×10^{-4}

Figure 1: Visualization of the simulation results presented in tabular form above.

Appendix D. Covariates in the empirical demonstration dataset

Table 4: Baseline covariates in the DHA study and ADNI/CPAD historical training data.

Covariate	Description
AChEI or memantine usage	Whether a subject is using a class of symptomatic Alzheimer's drugs
ADAS commands	Assesses the subject's ability to follow commands
ADAS comprehension	Assesses the subject's ability to understand spoken language
ADAS construction	Assesses the subject's ability to draw basic figures
ADAS ideational	Assesses the subject's ability to carry out a basic task
ADAS naming	Assesses the subject's ability to name common objects
ADAS orientation	Assesses the subject's knowledge of time and place
ADAS remember instructions	Assesses the subject's ability to remember test instructions
ADAS spoken language	Assesses the subject's ability to speak clearly
ADAS word finding	Assesses the subject's word finding in speech
ADAS word recall	Assesses the subject's ability to recall a list of words
ADAS word recognition	Assesses the subject's ability to remember and identify words
Age	Subject age at baseline
ApoE e4 Allele count	The number of ApoE e4 alleles a subject has (0, 1, or 2)
CDR community	Assesses the subject's engagement in community activities
CDR home and hobbies	Assesses the subject's engagement in home and personal activities
CDR judgement	Assesses the subject's judgement skills
CDR memory	Assesses the subject's memory
CDR orientation	Assesses the subject's knowledge of time and place
CDR personal care	Assesses the subject's ability to care for themselves
Diastolic blood pressure	The diastolic blood pressure of a subject
Education (Years)	The number of years of education of a subject
Heart rate	The resting heart rate of a subject
Height	The height of a subject
Indicator for clinical trial	1 if the subject is in an RCT, 0 if not
MMSE attention and calculation	Assesses the subject's attention and calculation skills
MMSE language	Assesses the subject's language skills
MMSE orientation	Assesses the subject's knowledge of place and time
MMSE recall	Assesses the subject's ability to remember prompts
MMSE registration	Assesses the subject's ability to repeat prompts
Region: Europe	1 if the subject lives in Europe, 0 otherwise
Region: Northern America	1 if the subject lives in the US or Canada, 0 otherwise
Region: Other	1 if the subject lives outside of Europe/US/Canada, 0 otherwise
Serious adverse events	The number of serious adverse events reported
Sex	1 if female, 0 if male
Systolic blood pressure	The systolic blood pressure of a subject
Weight	The weight of a subject

References

1. Maldonado G, Greenland S. Estimating causal effects. Int J Epidemiol 2002;31:422−9.
2. Sox HC, Goodman SN. The methods of comparative effectiveness research. Publ Health 2012;33:425−45.
3. Overhage JM, Ryan PB, Schuemie MJ, Stang PE. Desideratum for evidence based epidemiology. Drug Saf 2013;36:5−14.
4. Hannan EL Randomized clinical trials and observational studies guidelines for assessing respective strengths and limitations. JACC Cardiovasc Interv 2008;1:211−7.
5. Kopp-Schneider A, Calderazzo S, Wiesenfarth M. Power gains by using external information in clinical trials are typically not possible when requiring strict type I error control. Biom J 2020;62:361−74.

6. Ibrahim JG, Chen M-H, Gwon Y, Chen F. The power prior: theory and applications. Stat Med 2015;34:3724−49.
7. Lim J, Walley R, Yuan J, Liu J, Dabral A, Best N. Minimizing patient burden through the use of historical subject-level data in innovative confirmatory clinical trials. TIRS 2018;52:546−59.
8. Baker SG, Lindeman KS. Rethinking historical controls. Biostatistics 2001;2:383−96.
9. Ghadessi M, Tang R, Zhou J, Liu R, Wang C, Toyoizumi K, et al. A roadmap to using historical controls in clinical trials − by drug information association adaptive design scientific working group (DIA-ADSWG). Orphanet J Rare Dis 2020;15:69.
10. Hansen BB. The prognostic analogue of the propensity score. Biometrika 2008;95:481−8.
11. Aikens RC, Greaves D, Baiocchi M. A pilot design for observational studies: using abundant data thoughtfully. Stat Med 2020;39:4821−40.
12. Wyss R, Lunt M, Brookhart MA, Glynn RJ, Stürmer T. Reducing bias amplification in the presence of unmeasured confounding through out-of-sample estimation strategies for the disease risk score. J Causal Inference 2014;2:131−46.
13. Lin W. Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. Ann Appl Stat 2013;7:295−318.
14. Kahan BC, Jairath V, J Doré C, Morris TP. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. Trials 2014;15:139.
15. Raab GM, Day S, Sales J. How to select covariates to include in the analysis of a clinical trial. Contr Clin Trials 2000;21:330−42.
16. Yang L, Tsiatis AA. Efficiency study of estimators for a treatment effect in a pretest−posttest trial. Am Statistician 2001;55:314−21.
17. Committee for Medicinal Products for Human Use. Guideline on adjustment for baseline covariates in clinical trials. London: European Medicines Agency; 2015.
18. Cooney MT, Dudina AL, Graham IM. Value and limitations of existing scores for the assessment of cardiovascular risk: a review for clinicians. J Am Coll Cardiol 2009;54:1209−27.
19. Austin SR, Wong Y-N, Uzzo RG, Beck JR, Egleston BL. Why summary comorbidity measures such as the Charlson comorbidity index and elixhauser score work. Medical Care 2015;53:e65−72.
20. Ambrosius WT, Sink KM, Foy CG, Berlowitz DR, Cheung AK, Cushman WC, et al, The SPRINT Study Research Group. The design and rationale of a multicenter clinical trial comparing two strategies for control of systolic blood pressure: the systolic blood pressure intervention trial (SPRINT). Clin Trials 2014;11:532−46.
21. Borm GF, Fransen J, Lemmens WAJG. A simple sample size formula for analysis of covariance in randomized clinical trials. J Clin Epidemiol 2007;60:1234−8.
22. Rubin DB. Causal inference using potential outcomes. J Am Stat Assoc 2005;100:322−31.
23. Wang B, Ogburn EL, Rosenblum M. Analysis of covariance in randomized trials: more precision and valid confidence intervals, without model assumptions. Biometrics 2019;75:1391−400.
24. Leon S, Tsiatis AA, Davidian M. Semiparametric estimation of treatment effect in a pretest−posttest study. Biometrics 2003;59:1046−55.
25. Aronow PM, Miller BT. Foundations of agnostic statistics. New York: Cambridge University Press; 2019:286−7 pp.
26. Tsiatis A. Semiparametric theory and missing data. New York: Springer Science & Business Media; 2007.
27. Luo Y, Spindler M. High-dimensional L_2 boosting: rate of convergence. 2016 arXiv.
28. Belloni A, Chernozhukov V. Least squares after model selection in high-dimensional sparse models. Bernoulli 2013;19:521−47.
29. Farrell MH, Liang T, Misra S. Deep neural networks for estimation and inference. 2018 arXiv.
30. Syrgkanis V, Zampetakis M. Estimation and inference with trees and forests in high dimensions. 2020 arXiv.
31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. 2012 arXiv.
32. Quinn JF, Raman R, Thomas RG, Yurko-Mauro K, Nelson EB, Van Dyck C, et al. Docosahexaenoic acid supplementation and cognitive decline in alzheimer disease: a randomized trial. J Am Med Assoc 2010;304:1903−11.
33. Coon KD, Myers AJ, Craig DW, Webster JA, Pearson JV, Lince DH, et al. A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset alzheimer's disease. J Clin Psychiatr 2007;68:613−8.
34. Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. Am J Psychiatr 1984;141:1356−64.
35. Galasko D, Bennett D, Sano M, Ernesto C, Thomas R, Grundman M, et al. An inventory to assess activities of daily living for clinical trials in Alzheimer's disease. The Alzheimer's disease cooperative study. Alzheimer Dis Assoc Disord 1997;11:S33−9.
36. Morris JC. The clinical dementia rating (CDR): current version and scoring rules. Neurology 1993;43:2412−4.
37. Neville J, Kopko S, Broadbent S, Avilés E, Stafford R, Solinsky CM, et al, Coalition Against Major Diseases. Development of a unified clinical trial database for Alzheimer's disease. Alzheimer's Dementia 2015;11:1212−21.
38. Romero K, Mars M, Frank D, Anthony M, Neville J, Kirby L, et al. The coalition against major diseases: developing tools for an integrated drug development process for Alzheimer's and Parkinson's diseases. Clin Pharmacol Ther 2009;86:365−7.

39. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al. Double/debiased machine learning for treatment and structural parameters. Econom J 2018;21:C1−68.
40. Wager S, Du W, Taylor J, Tibshirani RJ. High-dimensional regression adjustments in randomized experiments. Proc Natl Acad Sci Unit States Am 2016;113:12673−8.
41. Rothe C. Flexible covariate adjustments in randomized experiments, Working Paper; 2018.
42. Dankar FK, El Emam K. The application of differential privacy to health data. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops on − EDBT-ICDT '12; 2012. pp. 158−66.
43. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated electronic health records. Int J Med Inf 2018;112:59−67.
44. Coalition Against Major Diseases, Organiza, Abbott, Alliance for Aging Research, Alzheimer's Association, Alzheimer's Foundation of America, AstraZeneca Pharmaceuticals LP, Bristol-Myers Squibb Company, Critical Path Institute, CHDI Foundation Inc, Eli Lilly and Company, F Hoffmann-La Roche Ltd, Forest Research Institute, Genentech Inc, GlaxoSmithKline, Johnson & Johnson, National Health Council, Novartis Pharmaceuticals Corporation, Parkinson's Action Network, Parkinson's Disease Foundation, Pfizer Inc, sanofi-aventis Collaborating, Fisher CK, Smith AM, Walsh JR. Machine learning for comprehensive forecasting of Alzheimer's disease progression. Sci Rep 2019;9:13622.
45. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine 2018;1:18.
46. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436.
47. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Briefings Bioinf 2018;19:1236−46.
48. Dubois S, Romano N, Jung K, Shah N, Kale D. The effectiveness of transfer learning in electronic health records data. In: Workshop Track - ICLR; 2017.
49. van der Vaart AW. Asymptotic statistics. Cambridge: Cambridge University Press; 2000.
50. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 1994;89:846.
51. Rosenblum M, van der Laan MJ. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. Int J Biostat 2010;6:13.
52. Freedman DA. On regression adjustments to experimental data. Adv Appl Math 2008;40:180−93.
53. Long JS, Ervin LH. Using heteroscedasticity consistent standard errors in the linear regression model. Am Statistician 2012;54:217−24.