Pharmaceutical Statistics

WILEY

TUTORIAL

A Tutorial on Improving RCT Power Using Prognostic Score Adjustment for Linear Models

Emilie Højbjerre-Frandsen[1,2] | Mathias Lerbech Jeppesen[1] | Rasmus Kuhr Jensen[1] | Claus Dethlefsen[1,2] | Rasmus Waagepetersen[2]

[1]Biostatistics, Novo Nordisk A/S, Søborg, Denmark | [2]Department of Mathematical Sciences, Aalborg University, Aalborg Øst, Denmark

Correspondence: Emilie Højbjerre-Frandsen (ehfd@novonordisk.com)

Received: 2 September 2024 | **Revised:** 27 June 2025 | **Accepted:** 5 July 2025

Funding: This work was supported by the Innovationsfonden, 2052-00044B.

Keywords: causal inference | diabetes | historical data | prognostic score | randomized trials

ABSTRACT

The use of historical data to increase power in clinical trials has been a topic of interest for many years. A recent approach adjusts linearly for a prognostic score. This is supported by asymptotic optimality results using influence functions for asymptotically linear estimators as well as finite sample optimality results. We review plug-in and linear estimators of average treatment effect in randomized clinical trials, sample size determination, and linear adjustment for a prognostic score. Guidelines and recommendations for the implementation of linear adjustment for a prognostic score are given including curation of historical data and construction of a prognostic score based on the historical data. A simulation study is conducted to investigate the performance in finite samples, comparing it to standard procedures such as propensity score matching for RCTs (PSM-RCT) and ANCOVA using simple baseline adjustment. Unlike PSM-RCT, linear adjustment for a prognostic score avoids biased treatment effect estimates and maintains control of type I error probability. The simulation study shows that the method is robust against deviations from method assumptions and poor performance of the prognostic model. A case study demonstrates an increase in prospective power using linear adjustment with a prognostic score in a phase IIIb clinical trial for type 2 diabetes. A final discussion considers limitations of the method for example in regard to subgroup analysis and the existence of already known prognostic baseline covariates.

1 | Introduction

A randomized clinical trial (RCT) is a vital tool for testing the efficacy and safety of new treatments. For instance, for Novo Nordisk[1] alone, more than 25,000 individuals participate in clinical trials each year. This comes with large economic costs and long timelines, as illustrated by the review Bentley et al. [1] on the costs, impact, and value of conducting clinical trials, where the costs are due to (1) infrastructure and (2) patient accrual and management. The overall cost of a trial can thus be reduced, and the trial process accelerated by having fewer participants in the trial. Furthermore, recruitment in itself constitutes a major challenge for several disease areas, and in some scenarios, it may not be ethically acceptable to conduct large placebo-controlled studies, as discussed by Temple and Ellenberg [2]. Overall, even a slight reduction in sample size will enable faster development of medicines at a lower cost, ultimately bringing new effective drugs faster to the patients.

Lowering the number of participants comes at a cost of loss in power; i.e., less ability to detect a treatment effect that truly exists. An approach to reduce the number of participants without compromising power is to leverage historical data, i.e., data from previous RCTs, observational studies, or other evidence sources. Lim et al. [3] proposed a partially external control arm method in the setting of an RCT, where the control group is populated with

historical controls by matching participants based on known confounders. This method is a type of propensity score matching (PSM) which we will refer to as PSM-RCT. Although the procedure proposed by Lim et al. may be effective, it is susceptible to bias since it disrupts the randomization of the trial, not all confounders may be accessible, and the covariate distribution may differ between the historical and current data. Thus, the procedure might increase the risk of a type I error, i.e., falsely declaring a non-effective treatment beneficial. Even when Pocock [4] six criteria are fulfilled for the historical data to be used as a synthetic control arm, there is no guarantee of type I error control. Another usage of historical data is through Bayesian methods that rely on specified prior beliefs about the parameters in the model used to estimate the treatment effect. As new data become available, these prior beliefs are updated. Descriptions of some Bayesian methods for causal inference are given in [5, 6]. However, these methods also lack strict type I error control in the frequentist sense.

Adjusting for measured baseline covariates may reduce the variance of the treatment effect estimate and thus offers an alternative to increasing power by increasing sample size [7–10]. This was demonstrated by Moore and van der Laan [8] in the context of logistic regression treatment effect estimation and more generally in Rosenblum and van der Laan [9] for a large class of generalized linear models (GLMs). However, adjusting for multiple not prespecified covariates may result in overfitting and an increase in type I error rates. Furthermore, the ad hoc selection of the adjustment set raises concerns about data dredging further elevating the risk of inflating the type I error rates. To mitigate these concerns, regulatory agencies such as the US Food and Drug Administration (FDA) and the European Medicines Agency (EMA) have issued guidelines on covariate adjustment [11, 12], which are conservative in regards to the number of covariates that may be adjusted for. Balzer et al. [7, 13] suggest to use the Adaptive Prespecification (APS) method combined with targeted maximum likelihood estimation (TMLE) and demonstrate great gains in study power. However, custom practice is still to use standard linear models without any targeting step for the primary analysis. It is thus worth exploring whether power can be increased within the realm of linear models by combining the use of historical data with covariate adjustment.

Schuler et al. [14] proposed linear adjustment with a prognostic score also known as PROCOVA[2]. In this paper, we will refer to this method as linear prognostic score adjustment. This method does not increase the risk of conducting a type I error, and Schuler et al. [14] showed that the average treatment effect (ATE) estimate obtained using this method is efficient under the assumption of homogeneous treatment effect (to be defined in Section 2), i.e., has the smallest possible asymptotic variance among a large group of estimators locally under the homogeneous treatment effect assumption. Many of the previously mentioned methods for covariate adjustment are also locally efficient [7–10]. For instance, using a GLM as a working model for the outcome, Rosenblum and van der Laan [9] show local efficiency of a marginal effect estimand under an RCT, where "local" refers to the condition that the working GLM model is correctly specified. Also, TMLE is locally semi-parametric efficient for many types of data [7, 9, 13, 15–19]. The method proposed by Schuler et al. [14] thus adds to the existing toolbox of locally semi-parametric efficient estimators. In addition to the attractive theoretical properties of the method, the method is appealing by being quite easy to comprehend, as conveyed in Section 4. Using linear prognostic score adjustment with a prespecified power, Unlearn.AI [20] and Schuler et al. [14] demonstrated large reductions in the control arm size for phase III trials. In September 2022, the Committee for Medicinal Products for Human Use [21] at the EMA issued a qualification opinion for linear prognostic score adjustment, expressing a generally favorable assessment of the method, highlighting the ability to control the type I error rate.

In this paper we initially set the theoretical framework for randomization that enables causal estimation of clinical trial estimands. We next describe simple plug-in and ordinary least squares approaches to ATE estimation and give a guide to prospective sample size determination in the design phase of a trial. We provide a practical account of linear adjustment with a prognostic score while a theoretical discussion of asymptotic and finite-sample efficiency is given in an Appendix. We outline the practicalities and recommendations for the method step by step, including curation and cleaning of the historical data as well as the training of the prognostic model.

The sensitivity to method assumptions and finite sample properties of linear prognostic score adjustment are examined through a simulation study. The study evaluates the performance of linear prognostic score adjustment under various scenarios and compares it to PSM-RCT [3] and standard ANCOVA methodologies. An R software package, called PostCard[3], was developed with functionalities for implementation and for deployment of a simulation study using prognostic score adjustment.

A prospective sample size calculation and post hoc analysis using linear prognostic score adjustment is conducted for a Novo Nordisk A/S phase IIIb trial examining a new drug in people with type 2 diabetes and with historical data from 16 previously finalized trials provided by Novo Nordisk A/S. In the field of diabetes, phase IIIb studies play a pivotal role in improving outcomes for people affected by the disease. Furthermore, there is a lot of data available both from previously conducted RCTs and real-world data. Hence, utilizing linear prognostic score adjustment in phase IIIb studies within the field of diabetes is an excellent use case that may shed light on the potential and possible limitations of the method.

The paper is concluded by a discussion of points of consideration regarding the use of adjustment for a prognostic score.

2 | Setting and Notation

The aim of RCTs is to collect data enabling the estimation of the effect of an intervention (such as a drug, device or other procedures) compared to a placebo, standard of care, or active comparator. Hatswell et al. [22] conducted a review of pharmaceutical approvals by EMA and FDA from 1999 to 2014 and demonstrated that RCTs form the foundation of regulatory approval. In RCTs, participants are randomly assigned to different groups: some receiving a new treatment and the others receiving the control. Randomization ensures that the groups can be expected to be (statistically) similar in terms of observed and

unobserved baseline characteristics, thereby minimizing confounding that could undermine the validity and reliability of the results. Randomization thus plays a large role in ensuring fair and unbiased decision making in regard to the causal effect of an intervention [23]. In this paper, we consider complete non-stratified randomization, but the proposed method is also applicable to other types of randomization.

We consider the setting of a two-armed trial with n participants, where the observational units, $O_i = (W_i, A_i, Y_i)$, are independent and identically distributed for $i = 1,2,3, \ldots, n$. Since the observations are i.i.d. we use the notation $O = (W, A, Y)$ without index i for a generic observation. Here Y represents a continuous primary endpoint variable while W is a vector of p baseline covariates collected at the first visit of the participant. Once the necessary baseline information is collected, participants are assigned to their respective treatment groups through randomization. This is indicated by the variable A, which is 1 if the participant is randomized to the new intervention and 0 if the participant is randomized to the control group. We make no parametric assumptions of the distribution of Y given (A, W). The trial data set is denoted $(\mathbb{W}, \mathbb{A}, \mathbb{Y}) \in \mathcal{W}^n \times \{0,1\}^n \times \mathbb{R}^n$, where $\mathcal{W}$ is the sample space of the W_i's, allowing covariates to be continuous, binary and categorical. The sizes of the two treatment groups are denoted n_1 and n_0 for treatment and control, respectively. For linear model analyses, we denote the design matrix $\mathbb{X}$, specifying the relevant form in each case.

To estimate the treatment effect, we follow the causal inference framework and roadmap from Petersen and van der Laan [24]. We use a Rubin causal model from [25, 26]. Each participant has two potential outcomes: $Y(1)$ under the new treatment and $Y(0)$ with the control treatment. The estimand of interest is the causal average treatment effect (ATE):

$$\Psi^* = \mathbb{E}[Y(1) - Y(0)] \tag{1}$$

We say that there is a homogeneous treatment effect when $\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1) - Y(0) \mid W]$, i.e., the effect of treatment is the same across covariate values. We only observe $Y = Y(0)(1 - A) + Y(1)A$, i.e., the potential outcome corresponding to the actual treatment allocation, which leads to a type of missing data problem. However, by randomization, the potential outcomes $Y(0)$ and $Y(1)$ are independent of the treatment allocation A and $P(A = a) = \pi_a$ with $0 < \pi_a < 1$ for $a \in \{0,1\}$. It then follows that there is no causal gap, i.e., the causal estimand (1) coincides with the statistical estimand,

$$\Psi = \mathbb{E}[Y \mid A=1] - \mathbb{E}[Y \mid A=0] = \mathbb{E}[\mathbb{E}[Y|A=1, W] - \mathbb{E}[Y|A=0, W]] = \mathbb{E}[\mu(1, W) - \mu(0, W)] \tag{2}$$

where $\mu(a, W) = \mathbb{E}[Y|A = a, W]$ is the conditional mean function.

3 | Estimators of the ATE

For continuous outcomes, the ATE is usually estimated using a linear model including an intercept and a treatment term. Specifically, ordinary least squares (OLS) estimation is employed for β_0 and β with the mean vector of the outcome vector $\mathbb{Y}$ modeled as $\beta_0 \mathbb{1}_n + \mathbb{X}\beta$ where 1_n is a vector of ones and $\mathbb{X}$ is an $n \times (1 + q)$ design matrix. The first column in $\mathbb{X}$ is the treatment indicator vector, $\mathbb{A}$. For the remaining columns we consider three scenarios: $q = 0$ meaning that $\mathbb{X}$ only consists of $\mathbb{A}$, $q = p$ with $\mathbb{X} = [\mathbb{A}\ \mathbb{W}]$, and $q = 2p$ with $\mathbb{X} = [\mathbb{A}\ \mathbb{W}\ \mathbb{A} * \mathbb{W}]$ where $\mathbb{A} * \mathbb{W}$ is the matrix obtained by multiplying each row in $\mathbb{W}$ with the corresponding component of $\mathbb{A}$. With $q = 0$ an ANOVA estimator is obtained known as the difference-in-means or unadjusted estimator. The cases with $q > 0$ yield Analysis of Covariance (ANCOVA) estimators. Following [14], we call the last two estimators ANCOVA I and ANCOVA II, respectively. We denote by $\widehat{\beta}_0$, and $\widehat{\beta}$ the OLS estimates of β_0 and β where $\widehat{\beta}$ has components $\widehat{\beta}_A$, $\widehat{\beta}_W$ and $\widehat{\beta}_{A*W}$ depending on the model.

3.1 | Plug-In Based ATE Estimator

To estimate the ATE using a linear model we consider the following straightforward plug-in method due to Rosenblum and van der Laan [9]: start by fitting ANOVA, ANCOVA I or ANCOVA II to estimate the conditional mean functions $\widehat{\mu}(a, w) = \widehat{\beta}_0 + x\widehat{\beta}$ and plug-in the result to obtain the estimator

$$\widehat{\Psi} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\mu}(1, w_i) - \widehat{\mu}(0, w_i) \tag{3}$$

We thus extract the counterfactual predictions from the linear model assuming everyone in the sample was actually treated ($a = 1$) as well as the opposite ($a = 0$), and replace expectation in (2) by a sample average over all covariate values.

Rosenblum and van der Laan [9] show that the plug-in estimator based on the linear model is a regular and asymptotically linear (RAL) estimator of the ATE under the assumption that $W \perp\!\!\!\perp A$ and $0 < \mathbb{P}(A) = \pi_a < 1$, which is fulfilled under an RCT. This means that the estimator is consistent and asymptotically normal regardless of the type of misspecification of both the linear model and distribution of the error term. Any RAL estimator has an influence function (IF) φ, which determines the asymptotic variance of the estimator [18], Chapter 3.1. This result outlined in [27, 28] shows that any RAL estimator $\widehat{\Psi}_n$ based on the n observations $O_1, \ldots, O_n$ has the limiting distribution

$$\sqrt{n}\left(\widehat{\Psi}_n - \Psi\right) \xrightarrow{d} N(0, \mathbb{V}\mathrm{ar}(\varphi(O))) \tag{4}$$

Hence the asymptotic variance of a RAL estimator is given by the variance of the IF φ. In [18], Chapter 3.3 the IF for the ATE is shown to be

$$\phi(O) = \frac{A}{\pi_1}(Y - \mu(1, W)) - \frac{1 - A}{\pi_0}(Y - \mu(0, W)) + \mu(1, W) - \mu(0, W) - \Psi \tag{5}$$

In practice we consistently estimate $\mathbb{V}\mathrm{ar}(\varphi(O))$ by the empirical variance of $\widehat{\phi}(O_i), i = 1, \ldots, n$, where $\widehat{\phi}$ is obtained by replacing μ and Ψ by their estimates. We thereby obtain valid confidence intervals (CI) and hypothesis testing even when the model is misspecified. This is further explored in Section 3.3.

In [14, 29, 30], the IF is used to determine the relation between the asymptotic variance of the ATE estimate found from the three counterfactual mean models (difference-in-means, ANCOVA I and ANCOVA II), under the assumption of independence of the observations. Specifically, when $\pi_1 \neq \pi_0$ and there is a heterogeneous (non-constant) treatment effect, the asymptotic variance of the ANCOVA I estimator could be larger than for the difference-in-means estimator. However, the ANCOVA II estimator yields the smallest asymptotic variance compared to the two other estimators, except when $\pi_1 = \pi_0$ or the treatment effect is constant, in which case ANCOVA I and II give the same asymptotic variance. Rosenblum and van der Laan [9] show that the ATE estimate is efficient if the linear model is correctly specified.

3.2 | Relation Between Plug-In and OLS Treatment Effect Estimation

For the ANOVA and ANCOVA I models, the plug-in estimator is simply the OLS estimate $\widehat{\beta}_A$ since for each i we have $\widehat{\mu}(1, w_i) - \widehat{\mu}(0, w_i) = \widehat{\beta}_A$. This also holds for ANCOVA II in the case where $\mathbb{W}$ is centered by subtracting its sample average. For an RCT it is in fact easy to show that the large sample limit of the OLS parameter estimate is equal to Ψ for all three models (see [10, 14, 29, 30]) even when the models are misspecified (assuming centered $\mathbb{W}$ for ANCOVA II). A self-contained account of this and of the asymptotic distribution of $\widehat{\beta}_A$ under a misspecified model is given in Appendix A. For ANOVA and ANCOVA I, the variance estimate found from the IF is equivalent to the variance estimate that can be extracted from White's [31] *heteroskedasticity consistent* (HC) variance estimator

$$\widehat{\mathbb{V}\mathrm{ar}}_{HC}\left[\left(\widehat{\beta}_0, \widehat{\beta}\right)^\top\right] = \left((1, \mathbb{X})^\top (1, \mathbb{X})\right)^{-1} \left((1, \mathbb{X})^\top \mathrm{diag}\left(\widehat{\varepsilon}_1^2, \ldots, \widehat{\varepsilon}_n^2\right)(1, \mathbb{X})\right)\left((1, \mathbb{X})^\top (1, \mathbb{X})\right)^{-1}$$

where $\widehat{\varepsilon}_i = Y_i - \widehat{\beta}_0 - X_i\widehat{\beta}$ is the estimated error term for the i'th subject. This is a consistent estimator of the asymptotic variance for $\left(\widehat{\beta}_0, \widehat{\beta}\right)$ even in presence of heteroskedasticity provided centering is not applied for $\mathbb{W}$ in case of ANCOVA I. To adjust for the finite sample size when using the HC estimator in practice, MacKinnon and White [32] proposed different correction factors for the estimated residuals. One of these is the HC3 correction, recommended for practical use by Long and Ervin [33]. Thus for ANOVA and ANCOVA I we can use the parameter estimate $\widehat{\beta}_A$ and the HC variance estimate to conduct valid hypothesis testing. For ANCOVA II, centering of $\mathbb{W}$ gives additional variation that is not covered by the HC estimator. This problem is pointed out by Center for Drug Evaluation and Research & Center for Biologics Evaluation and Research [11] and Ye et al. [34]. Ye et al. [34] suggest another variance estimator for the ANCOVA II $\widehat{\beta}_A$ estimator obtained with centered $\mathbb{W}$. However, in this paper, when using the ANCOVA II linear model, we refrain from centering $\mathbb{W}$ and instead use the suggested plug-in estimator and estimate its variance by the variance of the IF. This is implemented in the R software package PostCard. Schuler and van der Laan [18], Chapter 4.4 discuss some general advantages of using a plug-in based estimator compared to using model based parameter estimates.

3.3 | Sample Size Determination

An appropriate determination of the sample size is crucial during the planning phase of a trial to have sufficient precision of the subsequently estimated treatment effect. It is also pertinent to avoid unnecessary exposure of subjects to a potentially harmful treatment. To determine the sample size for a trial using the difference-in-mean, ANCOVA I or ANCOVA II estimator, we will formally state the hypothesis of the trial. In case of a superiority trial with superiority margin $\Delta > 0$ the $\mathscr{H}_0$ and $\mathscr{H}_1$-hypotheses are

$$\mathscr{H}_0 : \Psi - \Delta \leq 0 \qquad \text{and} \qquad \mathscr{H}_1 : \Psi - \Delta > 0 \tag{6}$$

A similar hypothesis can be formulated in case of a non-inferiority (with $\Delta < 0$) or equivalence trial. Using the plug-in ATE estimate $\widehat{\Psi}$ we can use the test statistic

$$t = \frac{\sqrt{n}\left(\widehat{\Psi} - \Delta\right)}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}}\left(\widehat{\varphi}\right)}} = \frac{\sqrt{n}\left(\widehat{\Psi} - \Delta\right)}{\sqrt{\mathbb{V}\mathrm{ar}(\varphi)}} \frac{\sqrt{\mathbb{V}\mathrm{ar}(\varphi)}}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}}\left(\widehat{\varphi}\right)}} \tag{7}$$

with $\mathbb{V}\mathrm{ar}(\varphi)$ short for $\mathbb{V}\mathrm{ar}(\varphi(O))$. By the consistency of the variance estimate, the last factor in (7) converges in probability toward 1. Therefore, by Slutsky's theorem, the asymptotic distribution of the test statistic coincides with the asymptotic distribution of

$$\frac{\sqrt{n}\left(\widehat{\Psi} - \Delta\right)}{\sqrt{\mathbb{V}\mathrm{ar}(\varphi)}} = \frac{\sqrt{n}\left(\widehat{\Psi} - \Psi\right)}{\sqrt{\mathbb{V}\mathrm{ar}(\varphi)}} + \frac{\sqrt{n}(\Psi - \Delta)}{\sqrt{\mathbb{V}\mathrm{ar}(\varphi)}} \tag{8}$$

Under the null hypothesis closest to $\mathscr{H}_1$, namely $\Psi = \Delta$, the last term is 0 and by (4),

$$t = \frac{\sqrt{n}\left(\widehat{\Psi} - \Psi\right)}{\sqrt{\mathbb{V}\mathrm{ar}(\varphi)}} \xrightarrow{d} \mathcal{N}(0, 1)$$

Suppose we reject when t exceeds the critical value given by the $1 - \alpha$ quantile of $\mathcal{N}(0, 1)$. Then, asymptotically, the significance level is α when $\Psi = \Delta$. Consider any other value $\Psi < \Delta$ under $\mathscr{H}_0$. Then the distribution of t is shifted to the left meaning that the significance level becomes smaller than α. Thus, for the chosen critical value, we have asymptotic type I error control.

To determine the power, we need to consider the distribution of t under $\mathscr{H}_1$ where the last term in (8) now moves the distribution to the right. We call the last term the non-centrality parameter, since it determines the mean shift under $\mathscr{H}_1$. [14, 35] show that for the ANCOVA II estimator adjusting for only one covariate W, the non-centrality parameter collapses to

$$nc = \sqrt{n}(\Psi - \Delta)\left(\frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} - \pi_1\pi_0\left(\frac{\rho_1\sigma_1}{\pi_1} + \frac{\rho_0\sigma_0}{\pi_0}\right)^2\right)^{-1/2} \tag{9}$$

where $\sigma_a^2 = \mathbb{V}\mathrm{ar}(Y(a))$ and $\rho_a = \mathbb{C}\mathrm{orr}(Y(a), W)$ for $a \in \{0, 1\}$. The parameters σ_0^2 and ρ_0 can be estimated from control arm historical data. Data for estimating σ_1^2 and ρ_1 are often unavailable. Therefore it is assumed that $\sigma_0^2 = \sigma_1^2$ coincides with the marginal

variance $\mathbb{V}\text{ar}(Y) = \sigma_Y^2$ and $\rho_0 = \rho_1 = \rho$. In this case the expression in (9) reduces to

$$\frac{\Psi - \Delta}{\sigma_Y \sqrt{(1-\rho^2)}} \sqrt{\frac{r}{(1+r)^2} n}$$

where $r = n_1 / n_0$ is the allocation ratio. A similar non-centrality parameter can be determined [36] when adjusting for more than one covariate. In this case ρ is replaced by

$$R^2 = \frac{\sigma_{WY}^\top \Sigma_W^{-1} \sigma_{WY}}{\sigma_Y^2} \tag{10}$$

where Σ_W denotes the covariance matrix of the covariates, and σ_{WY} is the q-dimensional column vector consisting of the covariances between the outcome variable and each covariate.

For a significance level α, the critical value is $F_0^{-1}(1-\alpha)$ where F_0 is the distribution function of t under H_0 where $t \sim \mathcal{N}(0,1)$. The power is the probability that t exceeds the critical value under H_1 where $t \sim \mathcal{N}(nc, 1)$. Specifically, the power is $1 - F_1\left(F_0^{-1}(1-\alpha)\right)$ where F_1 is the distribution function of $\mathcal{N}(nc, 1)$. The required sample size can be determined by increasing n until the power reaches the desired value. The significance level α may be replaced by $\alpha/2$ in accordance with the ICH E9 Guideline ([37], 27) stating that: "*The approach of setting type I errors for one-sided tests at half the conventional type I error used in two-sided tests is preferable in regulatory settings.*" This ensures consistency between the one-sided tests and the corresponding two-sided tests and therefore the same sample size is required regardless of whether a one-sided or two-sided test is conducted. Approximation formulas for the sample size are given in Appendix B. The formulas show that sample size is decreasing as a function of $1 - \rho^2$. That is, it is beneficial to adjust for covariates that are highly prognostic. Finally, when determining the sample size a sensitivity analysis is often conducted by inflating the estimate of σ_Y^2 and deflating the estimate of ρ using inflation and deflation factors.

4 | Linear Adjustment With A Prognostic Score

As shown in several publications [7–10] and motivated by the previous section, adjusting for highly prognostic covariates can decrease the standard deviation and thus increase the power (or decrease the sample size of a trial maintaining a prespecified power level). However, care should be taken in order not to increase the type I error by adjusting for too many non-prognostic covariates when using standard models like ANCOVA or GLMs. This is because, in a finite sample setting with n fixed, adding covariates (i.e., increasing p) may decrease the estimated squared error $\hat{\varepsilon}^2 = \left(Y - \hat{\beta}_0 - X\hat{\beta}\right)^2$ even though the added covariates are not correlated with Y. This problem of overfitting may cause the variance estimator to be biased downwards, which leads to increased type I error rates and invalid tests and confidence intervals. The Center for Drug Evaluation and Research & Center for Biologics Evaluation and Research [11] at FDA and the Committee for Medicinal Products for Human Use [12] at EMA therefore provide guidelines on covariate adjustment. According to the guidelines only a few highly prognostic baseline covariates should be included and they should be prespecified in the protocol or the statistical analysis plan (SAP) before any unblinding of data. No covariates measured after randomization should be included, as these could have been affected by the treatment allocation. Stratification variables and baseline values for continuous outcomes should always be included as adjustment covariates.

One way to increase power while using a prespecified set of adjustment covariates is to use linear adjustment with a prognostic score which is included among the prespecified adjustment covariates. Inspired by Hansen [38] we can use historical data to construct a highly prognostic covariate. Define the stochastic variable D to be one if a generic observation comes from the new trial and zero if the observation is from the historical data. The prognostic score is defined as the expected observed outcome conditional on the covariates and that the observation comes from the historical control data (where $A = 0$):

$$\rho(W) = \mathbb{E}[Y \mid W, A = 0, D = 0] \tag{11}$$

We will denote this as the oracle prognostic score. An estimator $\hat{\rho}(W)$ of the prognostic score is obtained by applying a machine learning algorithm to historical data $\left(\widetilde{\mathbb{W}}, \widetilde{\mathbb{Y}}\right) \in \mathcal{W}^{\tilde{n}} \times \mathbb{R}^{\tilde{n}}$ obtained for $\tilde{n}$ control participants. For an observation in the current trial with covariate W, $\hat{\rho}(W)$ is used as an additional covariate which may reduce the variance of the ATE estimator without compromising its consistency. Specifically, we augment the new RCT data with an additional column that consists of the estimated prognostic score for each participant, i.e., row i has an additional entry $\hat{\rho}(w_i)$, and we then use the plug-in estimator with ANOVA, ANCOVA I or ANCOVA II including in addition the prognostic score. The design matrices for this can be seen in Appendix C.

Intuitively, by constructing a prognostic score that explains much of the variation in the outcome Y, there is scope for considerably decreasing residual variance and increasing power. Specifically, we saw in Section 3.3 that the power increases if the correlation between the outcome and the adjustment covariate increases. In this respect, prognostic score adjustment is superior in several ways to mere linear adjustment for covariates. First, when fitting a prognostic score to historical data we implicitly perform a variable selection not influenced by the new trial data. This decreases the risk of overfitting arising from adjustment for covariates that by chance appear to be related to the outcome in the new trial. Second, as illustrated in the simple Example 4.1, through the application of machine learning models, we are able to detect non-linear and subgroup effects, thereby capturing nuanced relationships that may remain undetected when relying solely on linear adjustments. A detailed discussion on efficiency of prognostic score adjustment is given in Appendix D. Briefly, Schuler et al. [14] demonstrated that the estimator obtained from the method is semi-parametrically efficient when there is a homogeneous treatment effect and the prognostic model is an L_2-consistent estimate of the oracle prognostic score. Under the assumption of a homogeneous treatment effect and constant conditional variance, Theorem 2 in Appendix D states that the method is also optimal from a finite sample perspective although within

a smaller class of estimators. The assumption of a homogeneous treatment effect implies that the effect of treatment is the same across covariate values, which may not be realistic. However, as shown in Section 5, improvements may also be obtained in case of heterogeneous treatment effect.

Example 4.1. This example is based on the following structural causal model

$$\begin{aligned} W_1 &\sim \text{Unif}(-2,2) \\ A &\sim \text{Bern}(0.5) \\ Y(1)|W_1 &= 1.5 + 2\ \sin\left(|\ W_1|\ \right) + N(0,0.4) \\ Y(0)|W_1 &= 0.5 + 2\ \sin\left(|\ W_1|\ \right) + N(0,0.4) \\ Y &= AY(1) + (1-A)Y(0) \end{aligned} \tag{12}$$

We simulate a new RCT data set of size $n = 500$ and a historical data set of size $\widetilde{n} = 3000$. The historical data is simulated from the same structural causal model but with $Y = Y(0)$ for all participants. In this data generating scenario the covariate effect is not linear. This implies that no matter the size of the data set the ANCOVA I model is unable to capture the relationship between Y and W_1. Specifically, Figure 1A shows that the differences between the two groups is not explained by W_1 meaning that the model collapses to a difference-in-means model. Using the historical data to fit a prognostic model as an additional adjustment covariate for the ANCOVA I model, we can more accurately detect the treatment difference. This is illustrated in Figure 1B. The prognostic model was obtained using the Discrete Super Learner [39] with a library specified as in Appendix E.1. The model chosen by the Discrete Super Learner was a multivariate adaptive regression spline.

The data is simulated such that the true ATE equals 1. The model only adjusting for W_1 gives an CI of $\left[1.053 - t_{0.975,n-3} \cdot 0.0651; .0530 + t_{0.975,n-3} \cdot 0.065\right] = [0.9251; .182]$, whereas the model that additionally adjusts for the estimate prognostic scores yields $\left[0.977 - t_{0.975,n-4} \cdot 0.0380; .977 + t_{0.975,n-4} \cdot 0.038\right] = [0.9031; .050]$. This illustrates that we obtain narrower confidence intervals by including the estimated prognostic scores. Even though the relationship between Y and W_1 is poorly modeled by the linear model, we still obtain a consistent estimate of the ATE. Intuitively, this can be explained by the randomization process, which ensures an equal number of participants in both groups for all values of W_1. As a result, the ATE estimate effectively reduces to the difference-in-means estimator. Using the Frison-Pocock approximation formulas in Appendix B, the reduction in standard deviation would lead to a 42.5% reduction in sample size when adjustment for the prognostic score is used.

4.1 | Practical Implementation of Prognostic Score Adjustment for Linear Models

In September 2022, EMA issued a qualification opinion [21] on linear adjustment with a prognostic score. The assessment was, in general, favorable due to the method being a special case of the standard ANCOVA method. Thus, the method inherits properties that allow for establishing causal inference and asymptotic control of the type I error probability under randomization, even if the model is misspecified; see Section 3. This implies that, given suitable historical data, the method can be used for any clinical trial where it is decided to use an ANCOVA model for ATE estimation. In the following, we go through step by step the practical considerations and recommendations for implementing the method. This is based on the guidelines [40] by EMA.

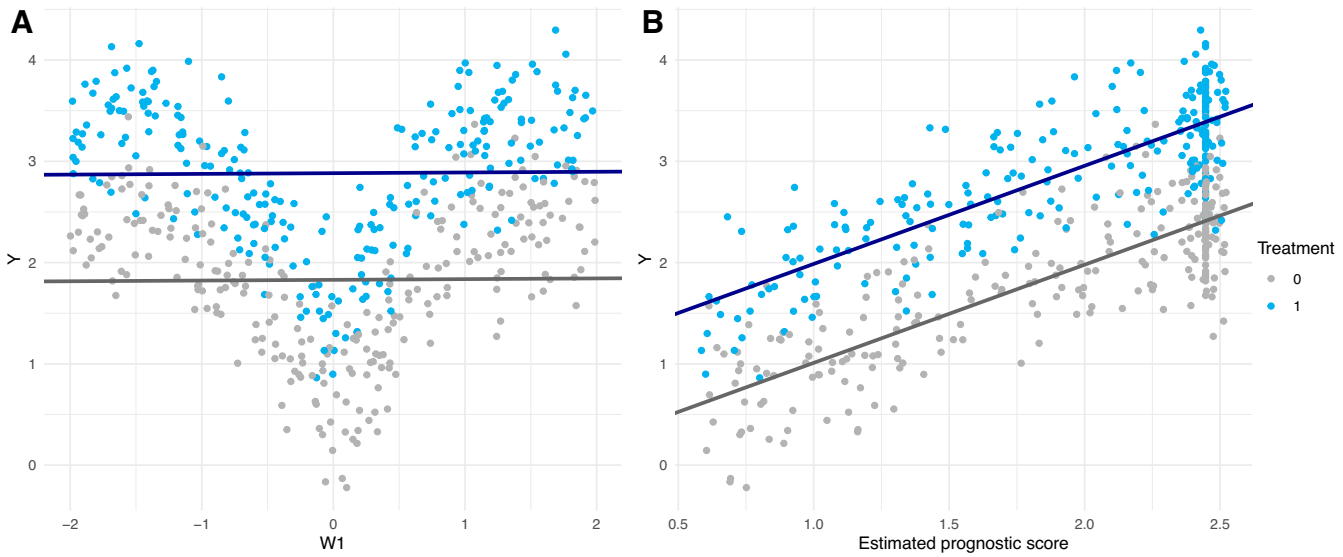

FIGURE 1 | Dots represent the data points with colors corresponding to the treatment groups: gray for $A = 0$ and blue for $A = 1$. (A) Relationship between W_1 and Y stratified by treatment A. The lines illustrate the fitted regression lines for the ANCOVA I model adjusting only for W_1. (B) Relationship between estimated prognostic score and Y stratified by treatment A. The lines illustrate the fitted regression lines for the ANCOVA I model adjusting for both W_1 and the estimated prognostic score.

4.1.1 | Curation of Historical Data

The theoretical optimality of the method is not always materialized in practice. For the method to be beneficial in a practical setup, we first need a sufficient amount of high quality historical data that are independent of the new study. One challenge is to ensure that the historical data are representative of the new study, both in terms of population and the type of data being collected. It is essential to fulfill Pocock's criteria to ensure the prognostic model's adequacy [4]. Additionally, the historical and the current data should be in the same format to avoid difficulties with collecting and structuring the data to perform prognostic model building. It is therefore important to allocate sufficient resources early in the trial process to structure the historical data into one large subject level data set that can be used for model building. To make the method effective, large integrated databases, potentially with shared data between different pharmaceutical companies, may be needed. Even if the data quality is good, we still need enough data to build a model using cross-validation to reduce variance and avoid overfitting. It is also crucial that there is enough historical data to partition it into a training and a test data set, which can be used to estimate the population and prognostic model performance parameters used for sample size determination.

4.1.2 | Prognostic Score Construction and Adjustment

It is recommended to construct the prognostic score using a highly adaptive machine learning model like the Discrete Super Learner [39] that encompasses both flexible models like multivariate adaptive splines and regression trees as well as a simple linear prognostic model. The Discrete Super Learner is a powerful tool in predictive modeling, known for its ability to select the best-performing model from a pool of candidate models. Its robustness against overfitting and flexibility in accommodating a wide array of base learner algorithms make it a powerful choice for capturing complex relationships within the data. The Discrete Super Learner has the oracle property of performing as well as the best machine learning algorithm in the library of models [39]. Even when there is strong prognostic baseline covariates available, the historical data can be used for selection of non-standard prognostic baseline covariates. Also, adjusting for a prognostic score built on several variables helps include non-standard prognostic baseline covariates without challenging the limitations on the number of covariates to adjust for in regulatory guidelines by FDA and EMA.

As we will see in Section 5 and illustrated by Example 4.1, a main benefit of using prognostic score adjustment comes from the ability to capture non-linear effects of the data. It is also recommended to include missingness indicators as input for the prognostic model. In addition, methods may be needed to handle missingness of the covariates as in any other RCT analysis. Again, it is important to allocate sufficient resources early in the trial process to build and validate the prognostic model based on good machine learning practices [41]. The decisions on model selection and tuning parameters must be prespecified in the SAP, and the prognostic model should be finalized before unblinding. The data scientists providing the prognostic scores should be blinded to the randomization code in the current study. Inclusion of other adjustment covariates and choice of variance estimator should also be specified in the SAP, in accordance with the recommendations of regulatory authorities as in [11]. Linear adjustment for a prognostic score can easily be used in combination with multiple imputation using the estimand framework [42], using Rubin's rules as usual. This should also be prespecified in the SAP. The sponsor should conduct the same sensitivity analysis as for any other trial regarding recruitment bias, complete losses to follow-up, and treatment compliance.

4.1.3 | Evaluation of Prognostic Model Performance

Validation of the model involves estimating the correlation coefficient (ρ or R in (10)) between the prognostic score and actual outcomes. This must be done on an out-of-sample (OOS) test data set similar to the current study in terms of duration, data collection, inclusion and exclusion criteria etc. Failing to use a representative OOS dataset can result in an underpowered study, which is unethical, since participants are unnecessarily exposed to a potentially harmful treatment. This would be at the risk of the sponsor. The assessment of representativity lies with the sponsor, who must also convince the ethics committee that the study is not underpowered due to the use of linear prognostic score adjustment. If an adequate OOS dataset is not available, it is not recommended to use linear adjustment with a prognostic score to decrease the sample size of the target study.

4.1.4 | Prospective Sample Size Determination With Prognostic Score Adjustment

Sample size determination can be performed based on the estimated correlation coefficient, target effect size, standard deviation of Y, randomization ratio r, expected dropout rate, significance, and power level. This involves a sensitivity analysis where the correlation is deflated and the standard deviation is inflated. The handbook [40] presents a rule-of-thumb approach to determine the deflation parameter: start with a deflation parameter of 0.95 if the in sample correlation coefficient is similar to the correlation coefficient obtained from two separate trial OOS data sets. If only one OOS data set is available it is set to 0.9. The deflation parameter is adjusted down by 0.05 if there are changes in the standard of care or different patterns of missingness in the current study compared to the historical data (such as some covariate values having a higher tendency to be missing). When including missingness indicators in the building of the prognostic model as an addition to imputation of missing data, different patterns of missingness may not be a concern. Also, the sponsor should be attentive to the inclusion of predictive rather than prognostic biomarkers in the prognostic model. This is because a predictive biomarker identifies the participants in

the control group that respond well to the control medicine, but does not predict how patients respond to the new treatment. Thus, if the prognostic model includes predictive biomarkers, the correlation between the actual outcome and the prognostic scores may be weaker for the new treatment arm than for the control arm. In this case the deflation parameter is adjusted down by 0.05 for the treatment arm and one should conduct the sample size determination using (9). Contrary to predictive biomarkers, a prognostic biomarker identifies the responders equally well under the two treatments. The handbook [40] recommends comparing the sample size determination to that of an ANCOVA I that adjusts for a few baseline covariates. It recommends using the correlation coefficient ρ for adjustment in the prognostic model and not R. However, R is relevant when we also directly adjust for some covariates. We therefore suggest to use both ρ and R to further assess sensitivity when determining the sample size. Considerations on parameter choices for the sample size determination must be prespecified in the SAP.

5 | Simulation Study

The simulation study examines the finite sample properties of linear prognostic score adjustment and how sensitive the method is to deviations from method assumptions such as the assumption of homogeneous treatment effect. Specifically, we examine how the method performs in different data generating scenarios including the presence of a distributional shift in the covariates between the historical and current RCT data as well as different data set sizes. We simulate the current and historical data using a complex mean structure. For conducting the simulation study, the R software package PostCard was developed. The code is available here.

5.1 | Setup

5.1.1 | Data Generation

The simulation study utilizes the structural causal model presented in (13) for generating the current trial data. It has a total of 7 observed covariates of diverse types, alongside one *unobserved* covariate U. For the historical data we use variations of the same model.

$$\begin{aligned} W_1 &\sim \text{Unif}(-2,1) \\ W_2 &\sim \text{Unif}(-2,1) \\ W_3 &\sim \mathcal{N}(0,3) \\ W_4 &\sim \text{Exp}(0.8) \\ W_5 &\sim \Gamma(5,10) \\ W_6, W_7 &\sim \text{Unif}(1,2) \\ U &\sim \text{Unif}(0,1) \\ A &\sim \text{Bern}(0.5) \\ Y(a)|W,U &= m_a(W,U) + \mathcal{N}(0,1.1) \\ Y &= AY(1) + (1-A)Y(0) \end{aligned} \tag{13}$$

This indicates that the outcome Y is simulated using the conditional mean $m_a(W,U)$. As the mean is conditional on both observed and unobserved covariates, it follows that our observable conditional means are $\mu_a(W) = E\left[m_a(W,U)|\,W\right]$. In the homogeneous treatment effect scenario we let,

$$\begin{aligned} m_0(W,U) = {} & 4.1 \cdot \sin\left(|\,W_2|\,\right) + 1.5 \cdot I\left(|\,W_4| > 0.25\right) + 1.5 \cdot \sin\left(|\,W_5|\,\right) + 1.4 \cdot I\left(|\,W_3| > 2.5\right) \\ & - 4.1 \cdot I\left(W_1 < -4.1\right) \cdot \sin\left(|\,W_2|\,\right) - 4.1 \cdot I\left(W_1 < -6.1\right) \cdot \sin\left(|\,W_2|\,\right) \\ & - 4.1 \cdot I(U > 1.1) \cdot \sin\left(|\,W_2|\,\right) - 4.1 \cdot I(U > 1.55) \cdot \sin\left(|\,W_2|\,\right) \end{aligned}$$

and

$$m_1(W,U) = \text{ATE} + m_0(W,U)$$

with ATE = 0.84. For the heterogeneous treatment effect scenario we use the same definition of the mean function m_0 and,

$$\begin{aligned} m_1(W,U) = {} & 4.3 \cdot \sin\left(|W_2|\right)^2 + 1.3 \cdot I\left(|W_4| > 0.25\right) + 4.1 \cdot I\left(W_2 > 0\right) \cdot \sin\left(|W_5|\right) + 1.6 \cdot \sin\left(|W_6|\right) + 1.4 \cdot I\left(|W_3| > 2.5\right) \\ & - 4.1 \cdot I\left(W_1 < -4.1\right) \cdot \sin\left(|W_2|\right) - 4.1 \cdot I\left(W_1 < -6.1\right) \cdot \sin\left(|W_2|\right) \\ & - 4.1 \cdot I(U > 1.1) \cdot \sin\left(|W_2|\right) - 4.1 \cdot I(U > 1.55) \cdot \sin\left(|W_2|\right) \end{aligned}$$

The heterogeneity is introduced in the first four terms. We find the ATE for this data generating scenario using the law of large numbers by simulating a large sample data set and using the difference-in-means estimator to determine the value 0.84 for the ATE. The oracle standard error (SE) is found from the EIF in (5) now using m_0 and m_1. The oracle SE for sample size $n = 200$ is 0.212 in the homogeneous treatment effect scenario and 0.221 in the heterogeneous scenario.

5.1.2 | Simulation Study Scenarios

Initially, we examine the scenario with no distributional shift between the current trial data and the historical data. Thus, the same data generating distribution is used for both data sets except that $A = 0$ deterministically for the historical data set. We next relax this assumption by modifying the historical data generating distribution, introducing varying degrees of observed

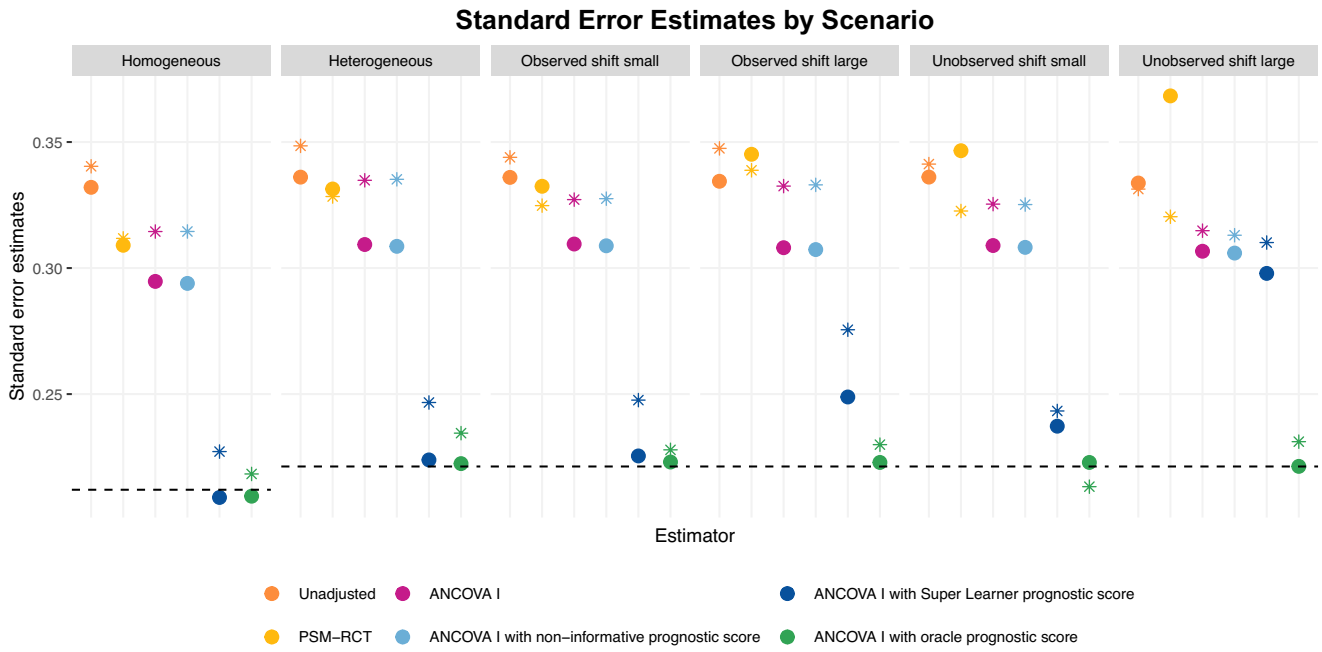

FIGURE 2 | Standard error estimates for different scenarios (distributional shifts only considered in the case of heterogeneous treatment effect). Dots represent the means of the estimated SEs while stars represent the empirically estimated SEs of the ATE estimates across the 500 simulated pairs of datasets. The dashed lines at 0.21 and 0.22 represent the oracle SEs in the homogeneous and heterogeneous cases, respectively.

and unobserved covariate shifts. For all scenarios, we simulate $N = 500$ pairs of historical and current trial data.

We start by fixing the current trial sample size $n = 200$ and the historical sample size $\widetilde{n} = 4000$. For these sample sizes, we investigate the homogeneous treatment effect and the heterogeneous data generating scenario in (13). For the heterogeneous case we further investigate scenarios with small and large distributional shifts between historical and trial populations. Specifically we sample the historical data with a small observable shift by using $W_1 \mid D = 0 \sim \text{Unif}(-4, -1)$ and a large observable shift by $W_1 \mid D = 0 \sim \text{Unif}(-7, -4)$. For the unobservable shift we use $U \mid D = 0 \sim \text{Unif}(0.5,\ 1.5)$ and $U \mid D = 0 \sim \text{Unif}(1.5,\ 2.5)$, respectively.

We also examine the effect of varying the historical and current RCT sample sizes both simultaneously and separately under the heterogeneous treatment effect scenario. First, we examine the effect of increasing the amount of current and historical data by setting $n = 50,\ 60,\ 70,\ \ldots,\ 200,\ 225,\ 250,\ 275,\ 300$ and setting $\widetilde{n} = 10n$ to align with the assumption of $n = \mathcal{O}(\widetilde{n})$ from Schuler et al. [14], Thm. 2. We also considered varying the current sample size as specified while fixing $\widetilde{n} = 4000$. Similarly we fixed $n = 100$ while varying $\widetilde{n}$ as specified before. When fixing the size of one of the data sets while varying the other we violate the assumption of [14], Thm. 2.

5.1.3 | Models for ATE Estimation

For ATE estimation, we consider the plug-in method described in Section 3.1 using ANCOVA models with and without linear prognostic score adjustment. In all ANCOVA models used for ATE estimation, we adjusted for all the observable covariates.

For the practically relevant example of linear prognostic score adjustment we consider the Discrete Super Learner specified in Appendix E.1 trained on the historical data using the observed covariates to estimate the prognostic score. This accommodates non-linear and interaction effects. We further benchmark against the optimal (but practically infeasible) oracle prognostic score, adjusting for $\mathbb{E}[Y(0) \mid W]$. In addition we consider a non-informative prognostic score that outputs a random value from the uniform distribution on the range of outcomes in the current RCT control group to test the robustness of the method in case the prognostic score does not have any predictive effect. We finally compare with PSM-RCT, for which we utilize a simple logistic regression model to estimate the propensity scores used for matching.

5.2 | Results

5.2.1 | Results in Different Data Generation Scenarios

Table S1 shows the empirical means of ATE and SE estimates across the 500 simulated data sets for all the data generating scenarios. The table further shows empirical SE, RMSE, power and 0.95 coverage. For all methods there is an approximate coverage of 0.95. The results indicate that the ATE estimates are unbiased except for the PSM-RCT method which shows a small positive bias in the large observable shift scenario. In all scenarios the PSM-RCT results in accurate or too large coverage because of an overly conservative SE estimate. This also yields a loss of power compared to the standard ANCOVA method.

Figure 2 presents a comparison of the SE estimates across the various scenarios. The filled points represent the means of the SE estimates of the ATE, while the empirically estimated SE is indicated by an asterisk (*). In general, we see that the empirically estimated SE is a bit underestimated for all of the

ANCOVA I plug-in estimators across all scenarios. However, as seen in Table S1, the coverage is still relatively close to 0.95. The underestimation of the SE could be eliminated by adjusting for fewer baseline covariates, which aligns with the guidelines from FDA and EMA [11, 12]. Another way to reduce this problem would be to use an out-of-sample estimate of the IF using a cross-validation procedure as described in Balzer et al. [7].

We see that the greatest advantage of using linear adjustment for a prognostic score is observed in the homogeneous treatment effect scenario. This is consistent with the findings of Schuler et al. [14] and the result in Theorem 2. However, in the heterogeneous treatment effect scenario, the relative benefit of using linear adjustment for a prognostic score compared to standard ANCOVA estimation is very similar, even though there are no analytical results that ensure asymptotic efficiency in this case.

In the covariate shifted cases, linear adjustment with a Super Learner prognostic score yields great improvements when the shift (observed or unobserved) is small. When the shift is large and observable there is less improvement. This can be explained by the need to extrapolate to predict on the current RCT data. When the shift is large and unobservable only small improvements in performance is observed. In this case the estimated prognostic score only contributes with noise to the ANCOVA I model similar to the non-informative prognostic score, but crucially this does not increase SE relative to the other ANCOVA I estimators in this case. This is in accordance with the asymptotic value of the corresponding β parameter being 0 (see (A8)), when the correlation between the adjustment covariate and the endpoint is 0, which is the case when the prognostic model does not carry any information on the new trial data.

The performance of the PSM-RCT method is volatile, and it even inflates the estimated SE compared to the unadjusted estimator in some scenarios. In contrast, linear adjustment for a prognostic score generally avoids inflation of the empirical SE. Moreover, in the scenarios where PSM-RCT enhances performance, linear adjustment for a Super Learner prognostic score still performs better. When there are unobserved shifts in the covariates, the SE estimate of the RCT-PSM method is highly overestimated, resulting in an overly conservative coverage.

Overall, we can conclude from Figure 2 and Table S1 that the linear adjustment with a prognostic score estimated by Super Learner exhibits the best feasible performance in terms of RMSE, average SE, and thus power in all scenarios. Furthermore, we observe that linear prognostic score adjustment is robust against poor-performing prognostic models, as adjusting for a randomly generated prognostic score produces similar results as for the corresponding ANCOVA estimators without prognostic adjustment.

5.2.2 | Results With Varying Sample Size

Figure 3 shows for the heterogeneous scenario the empirically estimated power (A) and 0.95 coverage (B) as a function of n with $\tilde{n} = 10n$. Overall, the oracle estimator has the fastest increase in

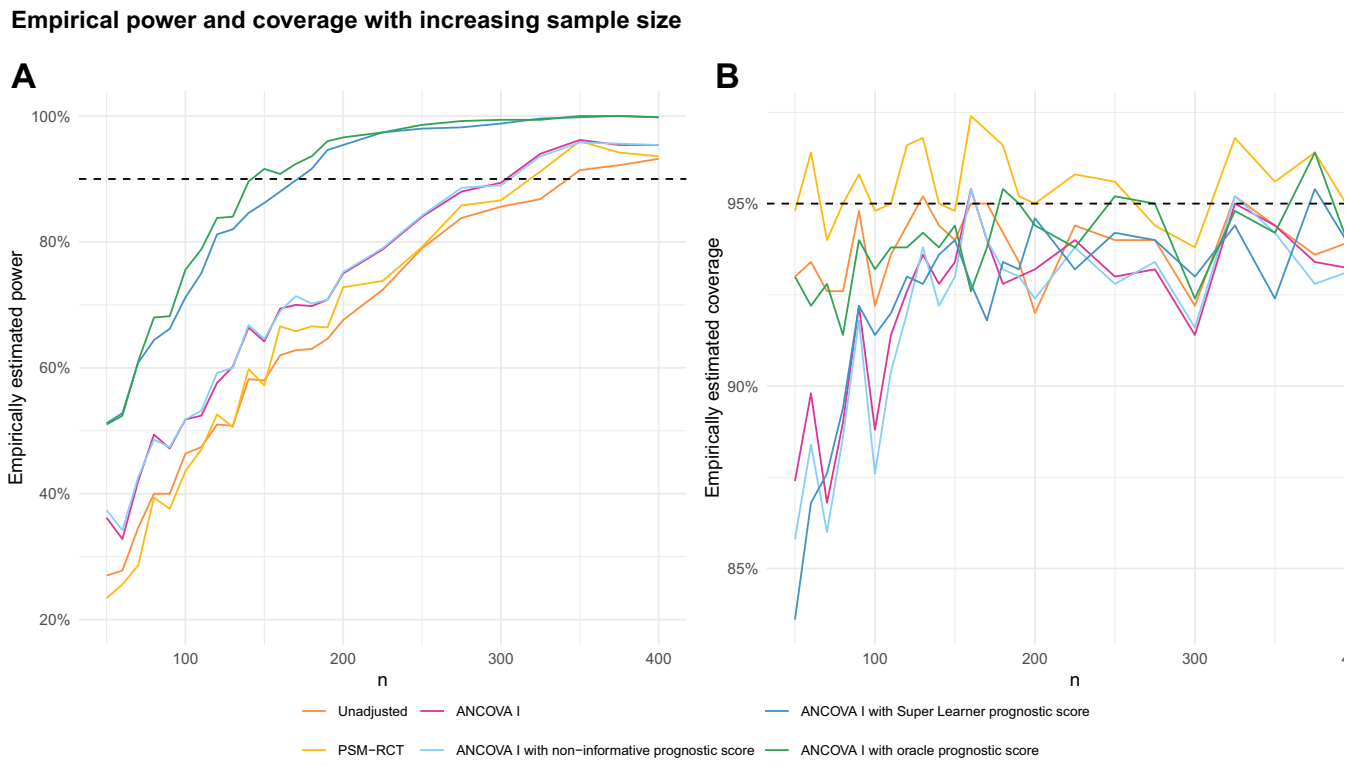

FIGURE 3 | Empirically estimated power (A) and 0.95 coverage (B) for each model in the heterogeneous scenario with varying n and $\tilde{n} = 10n$.

power, followed by the linear adjustment with a prognostic score estimated with the Super Learner. For these methods the coverage and equivalently the type I error probability is generally controlled. PSM-RCT yields a very slight increase in power compared to the difference-in-means estimator for $n > 150$. Once again, we note the robustness against poorly performing prognostic models since the curves for the ANCOVA I and linear adjustment for a random prognostic score align. This suggests that in the case where a strong prognostic model cannot be created the power will at most decrease down to the standard ANCOVA model, which is the current practice for continuous outcome analysis. In Section 4.1 we discussed how to conservatively estimate the performance of the prognostic model to perform sample size determination.

Additional results from the simulation study are presented in Appendix E.2. These findings suggest that varying only n produces similar results to varying both n and $\widetilde{n}$ simultaneously, indicating that n primarily controls the rate at which performance improves, see Figure E1. The use of a Super Learner prognostic model further does not appears to significantly enhance performance as more historical data becomes available. However, the effectiveness of a Super Learner prognostic model may vary depending on the specific circumstances of the study, such as more complex data generating processes that could yield different results for this plateau value.

6 | Phase IIIB Case Study

In this case study, we investigate the effect of using linear adjustment with a prognostic score for a phase IIIb RCT involving people diagnosed with type 2 diabetes (T2D). This is a chronic disease with a gradual decline in the regulation of glucose control. The measurement of hemoglobin A1C (HbA1C) is typically used to assess long-term blood glucose levels as an indicator of glucose control.

Specifically, we investigate the potential of reducing the prospective sample size for an upcoming phase IIIb RCT conducted by Novo Nordisk A/S by utilizing linear adjustment with a prognostic score. This trial will be referred to as the new RCT. The analysis will utilize data from 16 previously conducted RCTs within the field of diabetes, which were provided by Novo Nordisk A/S; see Appendix F. For conducting the prospective sample size determination, trial NN9068-4228 will be used as a test data set for estimating the population parameters and prognostic model performance required for the prospective sample size determination. This trial was also used for the original sample size calculation that was conducted for the new RCT. The remaining 15 trials are used to build the prognostic model and will be referred to as the historical data set. This means that we have three data sources, as illustrated in Appendix F. Summaries of the baseline covariates can be found in Tables S2 and S3.

The same data was used by Liao et al. [43] for another analysis using prognostic score adjustment for efficient estimators. However, here the goal is to conduct a prospective sample size calculation, whereas the purpose for [43] was to validate a novel method for prognostic score adjustment.

6.1 | Study Design

The study is an open-label, parallel group, and treat-to-target trial. The study objective was to confirm the efficacy (superiority for HbA1C) for a new type of basal insulin (referred to as new treatment) compared with daily existing insulin treatment, with or without oral anti-diabetic drugs (OADs) in insulin naive participants with T2D inadequately controlled with OADs. Inadequately controlled was defined as having $\text{HbA1C} \geq 8.0\%$. The goal was to obtain a product label expansion. The primary endpoint was defined as the change in HbA1C from baseline to week 40. The primary estimand was defined as a treatment policy estimand, i.e., the treatment effect of the new treatment against existing daily insulin treatment comparing change in HbA1c from baseline to week 40 in participants with T2D *regardless of* discontinuation of randomized treatment for any reason and *regardless of* initiation of non-randomized insulin treatment or additional anti-diabetic treatments for more than 2 weeks. For details on data preparation see Appendix F.2.

6.2 | Prognostic Score Estimation

For prognostic score estimation, a Discrete Super learner is built following the guidelines on good machine learning practices from [41]. This has the oracle property of performing as well as the best machine learning algorithm in the library of models [39]. The Lasso machine learning model was selected by the Discrete Super learner. The model provided an RMSE of 1.08 for the test data and 0.866 for the historical data, which indicates some degree of overfitting. For details on the prognostic score estimation, see Appendix F.3.

We chose to include baseline HbA1C in the prognostic model to more adequately model the prognostic scores even though this is also included directly in the ANCOVA model. This means that we cannot interpret the parameter associated with the covariate in the ANCOVA model in the usual way. However, only the parameter associated with the treatment will be used in the analysis, so this does not affect any conclusions.

6.3 | Prospective Power Estimation

For the prospective power calculation we use an allocation ratio of 1, assumed effect size -0.299, superiority margin of 0, and a significance level of $\alpha/2 = 0.025$. In the original prospective sample size determination made for the new RCT, the conditional variance was set to $\sigma^2 = 1$. In Section 3.3 the approximation formulas use the marginal variance σ_Y^2 and the correlation ρ. These quantities are interrelated by $\sigma^2 = \sigma_Y^2\left(1-\rho^2\right)$ [44]. To determine σ_Y^2 and ρ from $\sigma^2 = 1$, we calculated the marginal variance of change in HbA1C using the standard variance estimate and data from study NN9068-4228 and inflated this by 1.25 to be conservative, yielding $\sigma_Y^2 = 1.42$ and $\rho^2 = 0.30$. Using these population parameters resulted in a sample size of 474 participants for the new RCT without use of historical data. For the remaining sample size determinations we also used $\sigma_Y^2 = 1.42$ and determined R^2 based on the data from NN9068-4228, see Tables F1 and F2 in Appendix F. We compared the standard ANCOVA I method adjusting only for HbA1C with linear adjustment with a

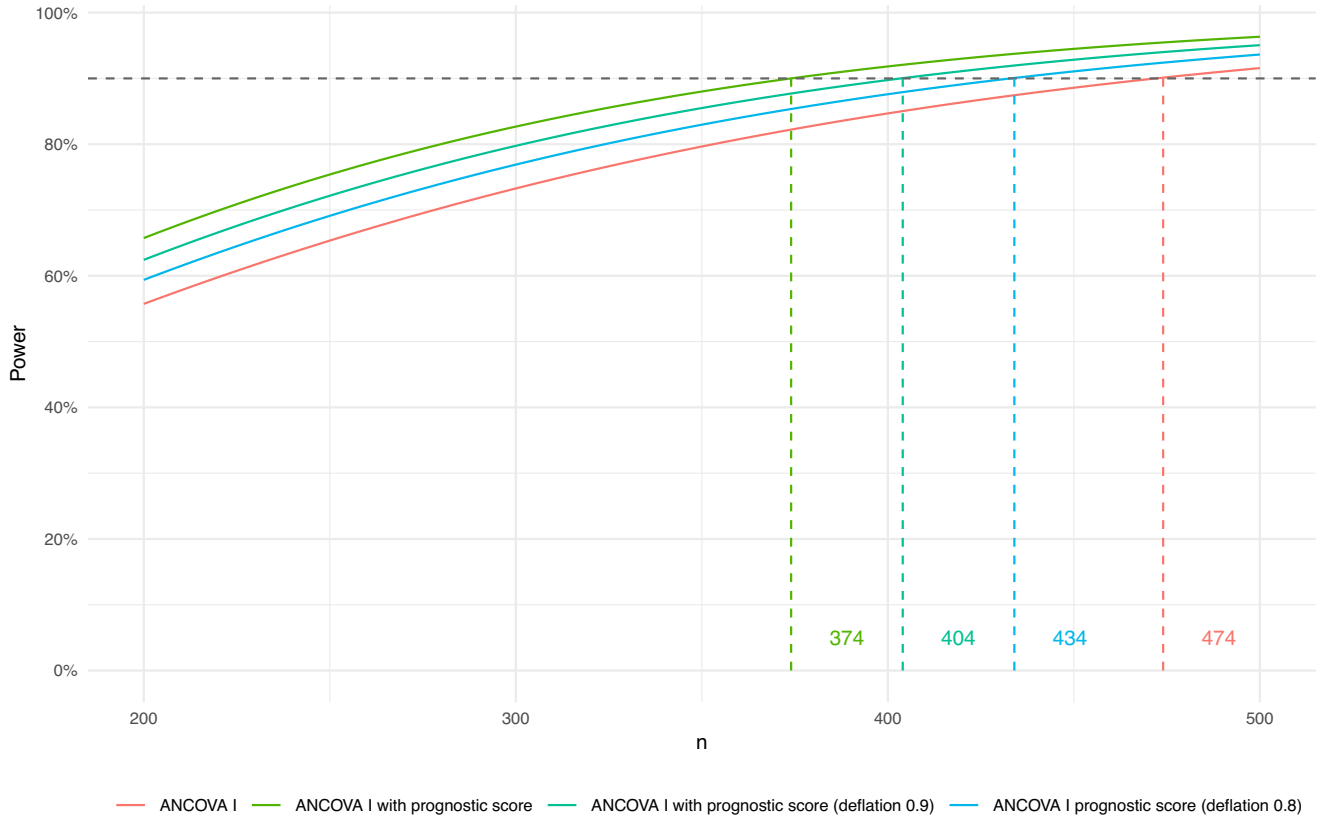

FIGURE 4 | Prospective estimation of power using the Guenther-Schouten approximation (see Appendix B) obtained from three different models for ATE estimation. Horizontal dashed line indicates the 90% power and the vertical dashed lines give the estimated sample size that gives a prospective power of 90%.

prognostic score with different deflation parameters on the estimated correlation.

The results of the prospective power estimation presented in Figure 4 show that linear adjustment with a prognostic score can be effective for increasing power or reducing sample size. For this particular trial, the business goal was to reduce the sample size by 40 participants, which could have been achieved through the use of linear adjustment with a prognostic score, even with a conservative deflation of the correlation parameter. In this particular case study, the goal of a 40-participant reduction could have been achieved using a deflation of 0.8, which is more conservative than the rule of thumb given in [40] (see Section 4.1).

7 | Discussion

In this section we discuss issues to consider before committing to using the method. The method is only theoretically validated in the case of a homogeneous treatment effect, see Appendix D. However, the simulation study in Section 5 suggests that the methodology is also beneficial in the heterogeneous treatment effect case. Moreover, the method can also still be effective when the historical sample is drawn from a different population than the current study population.

There should be enough historical data, so that this can be split into a training and testing data set. This poses a dilemma with using linear adjustment with a prognostic score for rare disease studies. These studies are often the ones for which we seek to decrease sample size, but for rare diseases there may be a limited pool of historical data available.

Adjusting the sample size for the primary endpoint may result in a decrease in power for secondary analyses. This is especially important in disease areas where a minimum number of participants need to be exposed. However, the problem could be eliminated by including a prognostic score for secondary endpoints to increase power for these analyses. Additionally, the method should not be used for subgroup analysis if the subgroup effect is already captured through the prognostic model, since this would bias the parameter associated with the subgroup effect. If the method is used to evaluate the impact of treatment on a specific subgroup of individuals, a new prognostic score should be constructed for that particular subgroup. However, creating distinct models for each analysis is logistically challenging and resource-intensive, especially when dealing with a plethora of secondary endpoints and numerous subgroups. We therefore suggest only using the method for certain secondary endpoints of high clinical importance. The sponsor should take the risk of underpowered secondary and subgroup analyses into account. Alternatively, the sponsor could consider keeping the same sample size but increasing the power by using the method, which may be more easily accepted.

Using linear adjustment with a prognostic score poses a business risk if the prognostic model is not as good as concluded

during the design phase of a trial. Underperformance in the new study would result in a decreased power, but with a limit down to the power gained from using a standard ANCOVA as seen in Section 5. Underpowered studies have the risk of producing false negatives, thus being a waste of resources for both the participants and the pharmaceutical company conducting the study. Furthermore, the effect of using linear adjustment with a prognostic score may be limited if highly prognostic baseline covariates are already directly adjusted for. However, including these highly prognostic covariates directly as adjustment covariates as well as in the prognostic model does not invalidate the analysis which seems to be in alignment with the opinion of the FDA ([11], 3): "*Covariate adjustment is acceptable even if baseline covariates are strongly associated with each other (e.g., body weight and body mass index). However, adjusting for less correlated baseline covariates generally provides greater efficiency gains*."

An ideal use case could be a phase IIIb clinical study. This type of study is crucial for broadening the understanding of new clinical treatments. The primary purpose is to expand the drug profile, e.g., the safety- or efficacy profile, or to obtain product label expansion. It represents a development stage of high importance, since it is initiated prior to regulatory approval but is not required for receiving the approval. However, the results should be ready before the drug is launched to be widely available in the market. Thereby, the study can directly impact patient care and benefit overall health outcomes by informing physicians and giving access to beneficial treatments earlier in the disease course. Also, by speeding up this type of study, the pharmaceutical companies can respond faster to market demands and remain at the forefront of innovation. Furthermore, the studies are important for strengthening the product's position and differentiating it from competitors; thereby potentially securing a larger market share.

Acknowledgments

We are grateful to Alejandro Schuler for introducing us to the method and for sharing his expertise. The framework and results from the simulation study are based on the master's thesis [35]. The guidance and support of the thesis supervisors Mikkel Meyer Andersen and Steffen Falgreen Larsen are gratefully acknowledged. We thank Martin Theil Hansen for curating the case study dataset and Kristoffer Segerstrøm Mørk and Andreas Dyreborg Christoffersen for assistance and helpful discussions related to the case study. We would like to express our sincere gratitude to the reviewers of Pharmaceutical Statistics for their invaluable and insightful comments, which greatly contributed to improving this article. This research was supported by Innovation Fund Denmark (grant number 2052-00044B).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Research data are not shared.

Endnotes

[1] Novo Nordisk Trials [webpage, 03/08/2023]. Available at: https://www.novonordisk-trials.com/.

[2] The authors of the article were all employed at Unlearn.AI[3] at the time of publication.

[3] The package is available on GitHub [webpage, 01/04/2024] at: https://github.com/NNpackages/PostCard.

References

1. C. Bentley, S. Cressman, K. van der Hoek, K. Arts, J. Dancey, and S. Peacock, "Conducting Clinical Trials—Costs, Impacts, and the Value of Clinical Trials Networks: A Scoping Review," *Clinical Trials* 16, no. 2 (2019): 183–193, https://doi.org/10.1177/1740774518820060.

2. R. Temple and S. S. Ellenberg, "Placebo-Controlled Trials and Active-Control Trials in the Evaluation of New Treatments. Part 1: Ethical and Scientific Issues," *Annals of Internal Medicine* 133, no. 6 (2000): 455–463, https://doi.org/10.7326/0003-4819-133-6-200009190-00014.

3. J. Lim, R. Walley, J. Yuan, et al., "Minimizing Patient Burden Through the Use of Historical Subject-Level Data in Innovative Confirmatory Clinical Trials: Review of Methods and Opportunities," *Therapeutic Innovation & Regulatory Science* 52, no. 5 (2018): 546–559.

4. S. J. Pocock, "The Combination of Randomized and Historical Controls in Clinical Trials," *Journal of Chronic Diseases* 29, no. 3 (1976): 175–188.

5. J. L. Hill, "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics* 20, no. 1 (2011): 217–240, https://doi.org/10.1198/jcgs.2010.08162.

6. D. Kaplan, J. Chen, S. Yavuz, and W. Lyu, "Bayesian Dynamic Borrowing of Historical Information With Applications to the Analysis of Large-Scale Assessments," *Psychometrika* 88, no. 1 (2023): 1–30, https://doi.org/10.1007/s11336-022-09869-3.

7. L. B. Balzer, M. J. van der Laan, and M. L. Petersen, "Adaptive Pre-Specification in Randomized Trials With and Without Pair-Matching," *Statistics in Medicine* 35, no. 25 (2016): 4528–4545, https://doi.org/10.1002/sim.7023.

8. K. L. Moore and M. J. van der Laan, "Covariate Adjustment in Randomized Trials With Binary Outcomes: Targeted Maximum Likelihood Estimation," *Statistics in Medicine* 28, no. 1 (2009): 39–64, https://doi.org/10.1002/sim.3445.

9. M. Rosenblum and M. J. van der Laan, "Simple, Efficient Estimators of Treatment Effects in Randomized Trials Using Generalized Linear Models to Leverage Baseline Variables," *International Journal of Biostatistics* 6, no. 1 (2010): 13, https://doi.org/10.2202/1557-4679.1138.

10. A. A. Tsiatis, M. Davidian, M. Zhang, and X. Lu, "Covariate Adjustment for Two-Sample Treatment Comparisons in Randomized Clinical Trials: A Principled Yet Flexible Approach," *Statistics in Medicine* 27, no. 23 (2008): 4658–4677, https://doi.org/10.1002/sim.3113.

11. Center for Drug Evaluation and Research & Center for Biologics Evaluation and Research, "Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products—Guidance for Industry," accessed July 11, 2023, https://www.fda.gov/media/148910/download.

12. Committee for Medicinal Products for Human Use, "Guideline on Adjustment for Baseline Covariates in Clinical Trials," accessed July 11, 2023, https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials_en.pdf.

13. L. B. Balzer, E. Cai, L. Godoy Garraza, and P. Amaranath, "Adaptive Selection of the Optimal Strategy to Improve Precision and Power in Randomized Trials," *Biometrics* 80, no. 1 (2024): ujad034, https://doi.org/10.1093/biomtc/ujad034.

14. A. Schuler, D. Walsh, D. Hall, J. Walsh, C. fisher, and 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," *International Journal of Biostatistics* 18, no. 2 (2022): 329–356, https://doi.org/10.1515/ijb-2021-0072.

15. W. Cai and M. J. van der Laan, "One-Step Targeted Maximum Likelihood Estimation for Time-To-Event Outcomes," *Biometrics* 76, no. 3 (2020): 722–733, https://doi.org/10.1111/biom.13172.

16. D. Chen, M. L. Petersen, H. C. Rytgaard, et al., "Beyond the Cox Hazard Ratio: A Targeted Learning Approach to Survival Analysis in a Cardiovascular Outcome Trial Application," *Statistics in Biopharmaceutical Research* 15, no. 3 (2023): 524–539, https://doi.org/10.1080/19466315.2023.2173644.

17. S. Gruber and M. J. van der Laan, "A Targeted Maximum Likelihood Estimator of a Causal Effect on a Bounded Continuous Outcome," *International Journal of Biostatistics* 6, no. 1 (2010): 26, https://doi.org/10.2202/1557-4679.1260.

18. A. Schuler and M. van der Laan, "Introduction to Modern Causal Inference," accessed November 15, 2023, https://alejandroschuler.github.io/mci.

19. M. J. van der Laan and S. Gruber, "Collaborative Double Robust Targeted Maximum Likelihood Estimation," *International Journal of Biostatistics* 6, no. 1 (2010): 1–71, https://doi.org/10.2202/1557-4679.1181.

20. "Case Study: Twinrcts Reduce Control Arm Sizes for Rare Neurodegenerative Diseases," accessed June 30, 2023, https://www.unlearn.ai/resources/case-study-twinrcts-tm-reduce-control-arm-sizes-for-rare-neurodegenerative-diseases#gated-file.

21. Committee for Medicinal Products for Human Use, "Qualification Opinion for Prognostic Covariate Adjustment," accessed June 30, 2023, https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/qualification-opinion-prognostic-covariate-adjustment-procovatm_en.pdf.

22. A. J. Hatswell, G. Baio, J. A. Berlin, A. Irs, and N. Freemantle, "Regulatory Approval of Pharmaceuticals Without a Randomised Controlled Study: Analysis of ema and Fda Approvals 1999–2014," *BMJ Open* 6, no. 6 (2016): e011666, https://doi.org/10.1136/bmjopen-2016-011666.

23. J. Pearl, "Causal Diagrams for Empirical Research," *Biometrika* 82, no. 4 (1995): 669–710.

24. M. L. Petersen and M. J. van der Laan, "Causal Models and Learning From Data: Integrating Causal Modeling and Statistical Estimation," *Epidemiology* 25, no. 3 (2014): 418–426, http://www.jstor.org/stable/24759134.

25. G. W. Imbens, "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics* 86, no. 1 (2004): 4–29.

26. J. S. Sekhon, "The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods," in *The Oxford Handbook of Political Methodology*, ed. J. M. Box-Steffensmeier, H. E. Brady, and D. Collier (Oxford University Press, 2008), 271–299.

27. O. Hines, O. Dukes, K. Diaz-Ordaz, and S. Vansteelandt, "Demystifying Statistical Learning Based on Efficient Influence Functions," *American Statistician* 76, no. 3 (2022): 292–304, https://doi.org/10.1080/00031305.2021.2021984.

28. A. A. Tsiatis, *Semiparametric Theory and Missing Data* (Springer, 2006).

29. W. Lin, "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique," *Annals of Applied Statistics* 7, no. 1 (2013): 295–318, https://doi.org/10.1214/12-AOAS583.

30. L. Yang and A. A. Tsiatis, "Efficiency Study of Estimators for a Treatment Effect in a Pretest–Posttest Trial," *American Statistician* 55, no. 4 (2001): 314–321, https://doi.org/10.1198/000313001753272466.

31. H. White, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica* 48, no. 4 (1980): 817–838, https://doi.org/10.2307/1912934.

32. J. G. MacKinnon and H. White, "Some Heteroskedasticity-Consistent Covariance Matrix Estimators With Improved Finite Sample Properties," *Journal of Econometrics* 29, no. 3 (1985): 305–325, https://doi.org/10.1016/0304-4076(85)90158-7.

33. J. S. Long and L. H. Ervin, "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model," *American Statistician* 54, no. 3 (2000): 217–224, https://doi.org/10.1080/00031305.2000.10474549.

34. T. Ye, J. Shao, Y. Yi, and Q. Zhao, "Toward Better Practice of Covariate Adjustment in Analyzing Randomized Clinical Trials," *Journal of the American Statistical Association* 118, no. 544 (2023): 2370–2382, https://doi.org/10.1080/01621459.2022.2049278.

35. E. Højbjerre-Frandsen, M. L. Jeppesen, and R. K. Jensen, "Increasing the Power in Randomised Clinical Trials Using Digital Twins," accessed November 19, 2023, https://projekter.aau.dk/projekter/da/studentthesis/increasing-the-power-in-randomised-clinical-trials-using-digital-twins(55a2b94f-d0a6-4ae3-a0dd-cf13e1f05cea).html.

36. G. Zimmermann, M. Kieser, and A. C. Bathke, "Sample Size Calculation and Blinded Recalculation for Analysis of Covariance Models With Multiple Random Covariates," *Journal of Biopharmaceutical Statistics* 30, no. 1 (2020): 143–159, https://doi.org/10.1080/10543406.2019.1632871.

37. Committee for Medicinal Products for Human Use, "ICH E9; Note for Guidance on Statistical Principles for Clinical Trials," accessed December 15, 2023, https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf.

38. B. B. Hansen, "The Prognostic Analogue of the Propensity Score," *Biometrika* 95, no. 2 (2008): 481–488.

39. M. J. van der Laan, E. C. Polley, and A. E. Hubbard, "Super Learner," *Statistical Applications in Genetics and Molecular Biology* 6, no. 1 (2007): 25, https://doi.org/10.2202/1544-6115.1309.

40. Unlearn.AI, "PROCOVA Handbook for the Target Trial Statistician," accessed December 18, 2023, https://www.ema.europa.eu/en/documents/other/procovatm-handbook_en.pdf.

41. The U.S. Food & Drug Administration and Health Canada and the United Kingdom's Medicines & Healthcare Products Regulatory Agency, "Good Machine Learning Practice for Medical Device Development: Guiding Principles," accessed January 3, 2024, https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles.

42. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, "Addendum on Estimands and Sensitivity Analysis in Clinical Trials," accessed August 13, 2024, https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf.

43. L. D. Liao, E. Højbjerre-Frandsen, A. E. Hubbard, and A. Schuler, "Prognostic Adjustment With Efficient Estimators to Unbiasedly Leverage Historical Data in Randomized Trials," accessed January 1, 2024, https://arxiv.org/ftp/arxiv/papers/2305/2305.19180.pdf.

44. G. Shieh, "Power and Sample Size Calculations for Contrast Analysis in ANCOVA," *Multivariate Behavioral Research* 52, no. 1 (2017): 1–11, https://doi.org/10.1080/00273171.2016.1219841.

45. R. Davison and J. G. MacKinnon, *Econometric Theory and Methods* (Oxford University Press, 2004).

46. L. Frison and S. J. Pocock, "Repeated Measures in Clinical Trials: Analysis Using Mean Summary Statistics and Its Implications for Design," *Statistics in Medicine* 11, no. 13 (1992): 1685–1704, https://doi.org/10.1002/sim.4780111304.

47. W. C. Guenther, "Sample Size Formulas for Normal Theory *T* Tests," *American Statistician* 35, no. 4 (1981): 243–244, https://doi.org/10.1080/00031305.1981.10479363.

48. M. Kieser, *Methods and Applications of Sample Size Calculation and Recalculation in Clinical Trials* (Springer, 2020).

49. H. J. A. Schouten, "Sample Size Formula With a Continuous Outcome for Unequal Group Sizes and Unequal Variances," *Statistics in Medicine* 18, no. 1 (1999): 87–91.

50. A. W. van der Vaart, *Asymptotic Statistics* (Cambridge University Press, 2012).

51. D. J. Stekhoven and P. Bühlmann, "MissForest—Non-Parametric Missing Value Imputation for Mixed-Type Data," *Bioinformatics* 28, no. 1 (2011): 112–118, https://doi.org/10.1093/bioinformatics/btr597.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** Supporting Information.

Appendix A

Determination of Large Sample Limit of Parameter Coefficients

In this section we assume that A and W are centered, which we will denote by A_c and W_c. This is not in general needed for the ANCOVA I estimator, but we do it here for ease of the calculations and by the Frisch-Waugh-Lovell (FWL) theorem ([45], 69) we obtain equivalent results for the linear models with or without demeaning. Firstly, by the law of large numbers, the difference-in-means estimator is obviously a consistent estimator of the statistical estimand (2). It also follows by the law of large numbers that $\widehat{\beta}_0$ and $\widehat{\beta}$ are consistent estimators of the best linear unbiased predictor (BLUP) coefficients

$$\beta_0^* = \mathbb{E}[Y] \quad \beta^* = \mathbb{E}\left[X^T X\right]^{-1}\mathbb{E}\left[X^T Y\right] \tag{A1}$$

where $X = \left(A_c, W\right)$ for ANCOVA I or $X = \left(A_c, W_c, A_c W_c\right)$ for ANCOVA II. More precisely, $\left(\beta_0^*, \beta^*\right)$ satisfies

$$\left(\beta_0^*, \beta^*\right) = \arg\min_{(\beta_0,\beta)} \mathbb{E}\left(Y - \beta_0 - X\beta\right)^2 \tag{A2}$$

We first show that $\beta_A^* = \Psi$ for the ANCOVA I. We start with determining $\mathbb{E}\left[(1,X)^\top Y\right]$ as

$$\mathbb{E}\left[(1,X)^\top Y\right] = \mathbb{E}\left[\begin{pmatrix} Y \\ A_c Y \\ W_c^\top Y \end{pmatrix}\right] = \begin{pmatrix} \mathbb{E}[Y] \\ \mathbb{E}\left[A_c Y\right] \\ \mathbb{E}\left[W_c^\top Y\right] \end{pmatrix} = \begin{pmatrix} \mathbb{E}[Y] \\ \mathbb{E}\left[A_c Y\right] \\ \mathbb{C}\mathrm{ov}\left(W_c^\top, Y\right) \end{pmatrix} \tag{A3}$$

where we use that in an RCT, we have independence between the potential outcomes and the covariates. We can express the second entry as

$$\mathbb{E}\left[A_c Y\right] = \sum_{a=0}^{1} \mathbb{E}\left[A_c Y \mid A = a\right] P(A = a) = \sum_{a=0}^{1} \mathbb{E}[Y \mid A = a]\left(a - \pi_1\right) P(A = a)$$
$$= E[Y \mid A = 1]\left(1 - \pi_1\right)\pi_1 - E[Y \mid A = 0]\pi_1\left(1 - \pi_1\right) = \pi_1\pi_0(\mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0]) \tag{A4}$$

We now wish to determine the first factor $\mathbb{E}\left[(1,X)^\top(1,X)\right]^{-1}$. We begin by writing

$$(1,X)^\top(1,X) = \begin{bmatrix} 1 & A_c & W_c \\ A_c & A_c^2 & A_c W_c \\ W_c^\top & W_c^\top A_c & W_c^\top W_c \end{bmatrix} \tag{A5}$$

so

$$\mathbb{E}\left[(1,X)^\top(1,X)\right] = \begin{bmatrix} 1 & \mathbb{E}\left[A_c\right] & \mathbb{E}\left[W_c\right] \\ \mathbb{E}\left[A_c\right] & \mathbb{E}\left[A_c^2\right] & \mathbb{E}\left[A_c W_c\right] \\ \mathbb{E}\left[W_c^\top\right] & \mathbb{E}\left[W_c^\top A_c\right] & \mathbb{E}\left[W_c^\top W_c\right] \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \mathbb{V}\mathrm{ar}\left(A_c\right) & 0 \\ 0 & 0 & \mathbb{E}\left[W_c^\top W_c\right] \end{bmatrix} \tag{A6}$$

using $\mathbb{C}\mathrm{ov}\left(A_c, W_c\right) = 0$ for an RCT and afterwards that A_c and W_c are centered. Thus the inverse is

$$\mathbb{E}\left[(1,X)^\top(1,X)\right]^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \dfrac{1}{\pi_1\pi_0} & 0 \\ 0 & 0 & \dfrac{1}{\mathbb{E}\left[W_c^\top W_c\right]} \end{bmatrix} \tag{A7}$$

Multiplying the expressions (A7) and (A3) finally yields

$$\beta^* = \begin{pmatrix} \mathbb{E}[Y] \\ \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\ \mathbb{V}\mathrm{ar}\left(W_c\right)^{-1}\mathbb{C}\mathrm{ov}\left(W_c^\top, Y\right) \end{pmatrix} \tag{A8}$$

Using a similar argument and in addition using that $\mathbb{C}\mathrm{ov}\left(A_c, A_c W_c\right) = \mathbb{E}\left[A_c^2\right]\mathbb{E}\left[W_c\right] = 0$ the same can be shown for ANCOVA II. But note here that we specifically use the centered version of W to obtain the result, whereas for ANCOVA I using an uncentered version of W would only change the first and third entry of β^*.

We conclude by deriving the asymptotic normal distribution of the OLS parameter estimate

$$\left(\widehat{\beta}_0, \widehat{\beta}^\top\right)^\top = \left(\left[1_n \mathbb{X}\right]^\top\left[1_n \mathbb{X}\right]\right)^{-1}\left[1_n \mathbb{X}\right]^\top \mathbb{Y}$$

The estimation error scaled by $\sqrt{n}$ is

$$\sqrt{n}\left(\widehat{\beta}_0 - \beta_0^*, \left(\widehat{\beta} - \beta^*\right)^\top\right)^\top = \left(n^{-1}\left[1_n \mathbb{X}\right]^\top\left[1_n \mathbb{X}\right]\right)^{-1} n^{-1/2}\left[1_n \mathbb{X}\right]^\top\left(\mathbb{Y} - \beta_0^* 1_n - \mathbb{X}\beta^*\right)$$

Here $n^{-1}\left[\not{s}_n \mathbb{X}\right]^\top\left[\not{s}_n \mathbb{X}\right]$ converges to $S = \mathbb{E}\left[(1,X)^\top(1,X)\right]$ by the law of large numbers. Further, by the central limit theorem,

$$n^{-1/2}\left[1_n \mathbb{X}\right]^\top\left(\mathbb{Y} - \beta_0^*\not{s}_n - \mathbb{X}\beta^*\right) = n^{-1/2}\sum_{i=1}^{n}\left(1, X_i\right)^\top \varepsilon_i$$

converges in distribution to $\mathcal{N}(0, \Sigma)$ with $\Sigma = \mathbb{V}\mathrm{ar}\left(\varepsilon(1,X)^\top\right)$, where $\varepsilon = Y - \beta_0^* - X\beta^*$ is the BLUP prediction error. It follows that $\sqrt{n}\left(\widehat{\beta}_0 - \beta_0^*, \left(\widehat{\beta} - \beta^*\right)^\top\right)^\top$ converges in distribution to $N\left(0, S^{-1}\Sigma S^{-1}\right)$.

Appendix B

Approximation Formulas for Sample Size Determination

Following [36] we state some sample size approximation formulas for a one-sided t test with significance level $\alpha/2$. A simple approximation formula is the multivariate version of the Frison-Pocock [46] formula given by

$$n_{FP} = \frac{(1+r)^2}{r}\frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2\sigma_Y^2\left(1 - R^2\right)}{(\Psi - \Delta)^2} \tag{B9}$$

where $z_{1-\alpha/2}$ denotes the $100 \cdot 1 - \alpha/2$-quantile of the normal distribution. This formula is derived using that the t-distribution is

approximately normal. A correction can be used when approximating the t-distribution by a normal distribution. This yields the Guenther-Schouten [47–49] approximation

$$n_{GS} = n_{FP} + \frac{\left(z_{1-\alpha/2}\right)^2}{2} \tag{B10}$$

When using the difference-in-mean estimator the term $\left(1 - R^2\right)$ is equal to one and when adjusting for only one covariate it equals $\left(1 - \rho^2\right)$.

Appendix C

Design Matrices for Ate Estimation With A Prognostic Score

Following the notation from Section 3 we define the $\mathbb{X}$ matrices as follows. For linear adjustment with a prognostic score, we construct the prognostic model $\widehat{\rho}$ and define the $\mathbb{X}$ matrix as follows. The two ANCOVA I estimators have

$$\mathbb{X} = \left[\mathbb{A} \quad \widehat{\rho}(\mathbb{W})\right], \quad \mathbb{X} = \left[\mathbb{A} \quad \widehat{\rho}(\mathbb{W}) \quad \mathbb{W}\right] \tag{C11}$$

respectively. The ANCOVA II counterparts for these two, have

$$\mathbb{X} = \left[\mathbb{A} \quad \widehat{\rho}(\mathbb{W}) \quad \mathbb{A} * \widehat{\rho}(\mathbb{W})\right], \quad \mathbb{X} = \left[\mathbb{A} \quad \mathbb{W} \quad \widehat{\rho}(\mathbb{W}) \quad \mathbb{A} * \mathbb{W} \quad \mathbb{A} * \widehat{\rho}(\mathbb{W})\right] \tag{C12}$$

respectively.

Appendix D

Efficiency of Linear Adjustment With A Prognostic Score

Schuler et al. [14] describe and theoretically validate the concept of linear adjustment with a prognostic score build from data from previously conducted trials or real world evidence. Further details of the theoretical derivations can be found in [35]. In this section we explain the method and give a more easily comprehensible argument for why it works.

Schuler et al. [14] use the concept of influence functions from efficiency theory to show that using $\widehat{\rho}$ for covariate adjustment indeed leads to semi-parametric efficient estimators under homogeneous treatment effect. Specifically, they use the IF that gives the smallest asymptotic variance, which is called the efficient influence function (EIF). An estimator whose IF is equal to the EIF is called an oracle estimator. The following theorem due to Schuler et al. [14] is restricted to regular asymptotic linear (RAL) estimators. However, by the Hájak-Le Cam convolution theorem [50] the most efficient regular estimator is guaranteed to be asymptotically linear. For a discussion on the assumption of regularity see [18], Section 3.1.

Theorem 1. (Oracle_{I} *estimator). Assume that* $\mathbb{E}[Y(1)|\,W] = \mathbb{E}[Y(0)|\,W] + \text{ATE}$. *Then the ANCOVA I ATE estimator with* $\mathbb{E}[Y(0)|\,W]$ *as covariate in place of* W *has the lowest possible asymptotic variance among all RAL estimators with access to* W [14].

The assumption of a homogeneous treatment effect implies that the effect of treatment is the same across covariate values, such that the conditional ATE (CATE) is equal to the ATE. In practice this may not be valid but as shown in Section 5, improvements may still be obtained. Furthermore, a similar result can be shown without the homogeneous treatment effect assumption. However, in this case we would need to use the ANCOVA II estimator with covariate vector $\left(\mathbb{E}[Y(0)|\,W], \mathbb{E}[Y(1)|\,W]\right)$, which is then called the $\text{Oracle}_{\text{II}}$ estimator. In practice it is usually not feasible to estimate $\mathbb{E}[Y(1)|\,W]$ and therefore we will focus on the homogeneous treatment effect set up in Theorem 1.

In the following we consider a less abstract approach to optimality and show that the model in Theorem 1 gives a best linear unbiased estimator (BLUE) of the ATE. This is a finite sample argument that further supports the use of prognostic score adjustment for linear models. However, we need further assumptions for this result to hold.

Theorem 2. *(Optimal ATE estimator under homogeneous treatment effect). Assume that* $\mathbb{E}[Y(1)|\,W] = \mathbb{E}[Y(0)|\,W] + \text{ATE}$. *Also assume that the conditional variance* $\mathbb{V}\text{ar}(Y|A,W) = \sigma^2$ *does not depend on* (A,W). *Then the OLS estimate* $\widehat{\beta}_A$ *obtained from an ANCOVA model with design matrix* $\mathbb{X} = [\mathbb{A} \quad \mathbb{E}[Y(0)|\,\mathbb{W}]]$ *is an unbiased estimator of* ATE *and has the lowest possible variance among all estimators of* ATE *that are conditionally unbiased given* (W,A) *and of the linear form.*

$$B(\mathbb{W},\mathbb{A})\mathbb{Y} \tag{D13}$$

where the $1 \times n$ matrix $B(\mathbb{W},\mathbb{A})$ is some function of $\mathbb{W}$ and $\mathbb{A}$.

Proof of Theorem 2. The ANCOVA procedure fits the linear model

$$\mathbb{E}[Y|W,A] = \beta_0 + \beta_A A + \beta_1 \mathbb{E}[Y(0)|\,W] \tag{D14}$$

where $\beta = \left(\beta_0, \beta_A, \beta_1\right)^\top \in \mathbb{R}^3$. The ANCOVA OLS estimate is $\widehat{\beta} = \mathbb{M}\mathbb{Y}$ where $\mathbb{M} = \left(\left[1_n\ \mathbb{X}\right]^\top \left[1_n\ \mathbb{X}\right]\right)^{-1} \left[1_n\ \mathbb{X}\right]^\top$. Under, D14 $\widehat{\beta}$ is conditionally unbiased (and hence unbiased) for any $\beta = \left(\beta_0, \beta_A, \beta_1\right)^\top \in \mathbb{R}^3$ since $\mathbb{E}[\mathbb{M}\mathbb{Y}|\mathbb{W},\mathbb{A}] = \mathbb{M}\left[1_n\ \mathbb{X}\right]\left(\beta_0, \beta_A, \beta_1\right)^\top = \left(\beta_0, \beta_A, \beta_1\right)^\top$. Under homogeneous treatment effect, $\widehat{\beta}_A$ is an unbiased estimator of ATE since then

$$\begin{aligned}
&\mathbb{E}[Y|W,A] = \mathbb{E}[AY(1) + (1-A)Y(0)\,|\,W,A] \\
&= A\mathbb{E}[Y(1)|\,W] + (1-A)\mathbb{E}[Y(0)|\,W] \\
&= \mathbb{E}[Y(0)|\,W] + A \cdot ATE
\end{aligned} \tag{D15}$$

which is the special case of (D14) with $\beta_0 = 0$, $\beta_1 = 1$ and $\beta_A = \text{ATE}$. Thus in case of homogeneous treatment effect the conditional expected value of Y in fact follows the model (D14).

We now show that $\widehat{\beta}_A = [0\,1\,0]\mathbb{M}\mathbb{Y}$ is optimal under the model (D14) with $\mathbb{V}\text{ar}(\mathbb{Y}|A,W) = \sigma^2 I$. Specifically, we show that $\mathbb{V}\text{ar}\left(\widetilde{\beta}_A\right) \geq \mathbb{V}\text{ar}\left(\widehat{\beta}_A\right)$ for all estimators $\widetilde{\beta}_A$ of the form (D13). This follows if we show that $\mathbb{C}\text{ov}\left(\widetilde{\beta}_A - \widehat{\beta}_A, \widehat{\beta}_A\right) = 0$ because then $\mathbb{V}\text{ar}\left(\widetilde{\beta}_A\right) - \mathbb{V}\text{ar}\left(\widehat{\beta}_A\right) = \mathbb{V}\text{ar}\left(\widetilde{\beta}_A - \widehat{\beta}_A\right) \geq 0$. To formally show this we use the law of total covariance to obtain

$$\begin{aligned}
&\mathbb{C}\text{ov}\left(\widetilde{\beta}_A - \widehat{\beta}_A, \widehat{\beta}_A\right) = \mathbb{C}\text{ov}\left(\mathbb{E}\left[\widetilde{\beta}_A - \widehat{\beta}_A|\,\mathbb{W},\mathbb{A}\right], \mathbb{E}\left[\widehat{\beta}_A|\,\mathbb{W},\mathbb{A}\right]\right) \\
&+ \mathbb{E}\left[\mathbb{C}\text{ov}\left(\widetilde{\beta}_A - \widehat{\beta}_A, \widehat{\beta}_A|\,\mathbb{W},\mathbb{A}\right)\right]
\end{aligned}$$

The first term is zero because both estimators are conditionally unbiased given $\mathbb{W}$ and $\mathbb{A}$. Considering the last term,

$$\begin{aligned}
&\mathbb{C}ov\left(\widetilde{\beta}_A - \widehat{\beta}_A, \widehat{\beta}_A\,|\,\mathbb{W},\mathbb{A}\right) = \mathbb{C}ov((B(\mathbb{W},\mathbb{A}) - [0\,1\,0]\,\mathbb{M})\mathbb{Y}, [0\,1\,0]\mathbb{M}\mathbb{Y}\,|\,\mathbb{W},\mathbb{A}) \\
&= (B(\mathbb{W},\mathbb{A}) - [0\,1\,0]\,\mathbb{M})\mathbb{V}\text{ar}(\mathbb{Y}|\mathbb{W},\mathbb{A})\mathbb{M}^\top [0\,1\,0]^\top = \sigma^2 (B(\mathbb{W},\mathbb{A}) - [0\,1\,0]\,\mathbb{M})P\mathbb{M}^\top [0\,1\,0]^\top
\end{aligned}$$

where $P = \left[1_n\ \mathbb{X}\right]\mathbb{M}$ is the projection onto the span of $\left\{1_n,\ \mathbb{A},\ \mathbb{E}[Y(0)|\,\mathbb{W}]\right\}$. To show that this is equal to 0, we will show that $(B(\mathbb{W},\mathbb{A}) - [0\,1\,0]\mathbb{M})Px = 0$ for all $x \in \mathbb{R}^n$. Letting $\mu = \mathbb{E}[\mathbb{Y}|\mathbb{W},\mathbb{A}]$ we have

$$\mathbb{E}[B(\mathbb{W},\mathbb{A})Y\,|\,\mathbb{W},\mathbb{A}] = B(\mathbb{W},\mathbb{A})\mu = \beta_A = [0\,1\,0]\mathbb{M}\mu \Rightarrow (B(\mathbb{W},\mathbb{A}) - [0\,1\,0]\mathbb{M})\mu = 0$$

since both $\widehat{\beta}_A$ and $\widetilde{\beta}_A$ are conditionally unbiased given $\mathbb{W}$ and $\mathbb{A}$. It follows that $(B(\mathbb{W},\mathbb{A}) - [0\,1\,0]\mathbb{M})\mu = 0$ for any μ in the span of $\left\{1_n, \mathbb{A}, \mathbb{E}[Y(0)|\,\mathbb{W}]\right\}$ and hence for all $x \in \mathbb{R}^n$, $(B(\mathbb{W},\mathbb{A}) - [0\,1\,0]\mathbb{M})Px = 0$. □

Schuler et al. [14] showed in an asymptotic setting using the broad class of RAL estimators that linear adjustment with a prognostic score gives the most efficient estimate of the ATE, whereas our result above supports the method in a finite sample setting but with more strict assumptions. Specifically, our result regards a subclass of the RAL estimators namely the linear and conditionally unbiased estimators. In practice, unbiasedness is typically obtained from conditional unbiasedness, so assuming conditional unbiasedness does not seem restrictive. Assuming constant

conditional variance is more restrictive. This implies that the conditional covariance matrix of $\mathbb{Y}$ is diagonal $\sigma^2 I$ since the observations are assumed to be independent. It is possible to relax this assumption by assuming a non-diagonal conditional covariance matrix $C(\mathbb{A}, \mathbb{W})$. Then the proof can be modified to show that the weighted least squares estimator

$$[0\,1\,0]\widetilde{\mathbb{M}}\mathbb{Y} \quad \text{with} \quad \widetilde{\mathbb{M}} = \left([1_n\ \mathbb{X}]^\top C(\mathbb{A}, \mathbb{W})^{-1} [1_n\ \mathbb{X}] \right)^{-1} [1_n\ \mathbb{X}]^\top C(\mathbb{A}, \mathbb{W})^{-1}$$

is optimal for β_A among all conditionally unbiased estimators of the form (D13). This perspective is relevant for example in case of repeated measurements with correlations between observations for the same subject. From a practical point of view one might obtain a working estimate of $C(\mathbb{A}, \mathbb{W})$ using e.g., linear mixed model software.

Appendix E

Simulation Study Specification

Discrete Super Learner

Number of folds: Cross-validation is used to select the best candidate learner in the library for the historical sample. A 3-fold scheme is used when the historical sample size is over 5000, 5-fold scheme when it is over 4000, and 10-fold when it is less than 1000.

Library of learners:

- Multivariate Adaptive Regression Spline with the highest interaction to be to the 3rd degree
- Linear regression
- Extreme gradient boosting with specifications: learning rate 0.1, tree depth 3, crossed with number of trees specified from 25 to 500 by 25 increments

Loss function: Mean square error loss.

Varying Sample Size

Figure E1 displays the power curves obtained when varying the sample sizes separately. We do not include the coverage plots in this case since these are similar to the results in Figure 3 with approximate control of the type I error.

Appendix F

Phase IIIB Case Study

Summary of Case Study Data

Data Missingness

None of the 15 studies contained week 40 HbA1c observations. If available, week 38 observations were imputed as surrogates for the week 40 observations. If week 38 observations were not available, week 42 was used instead. If neither 38 nor 42 week observations were available, the subject mean between week 36 and week 44 value was used. This imputation strategy is reasonable since HbA1C normally stabilizes around week 12–16. For the remaining missing values, we imputed using an ANCOVA with adjustment covariates: last observed HbA1C measurement before the landmark visit, time point of last measurement, baseline HbA1C, discontinuation prior to week 40 indicator, and study-id. This was only done on the historical data set.

After imputing the primary endpoint, a total of 94.7% of the participants had complete data in the pooled testing and historical data for the baseline covariates. A missingness pattern plot for the covariates can be seen in Figure F2. The missing covariates for 5.3% of the participants were imputed using an RF [51] separately on the historical and testing data. In the historical data sample, the normalized root mean square error for continuous covariates was 0.219 and the proportion of falsely classified data was 0.005. For each covariate that had missing values, a missingness indicator was constructed as an additional covariate used for model building. However, for the new trial data, the normalized root mean square error was 0.19, and there were no falsely classified data.

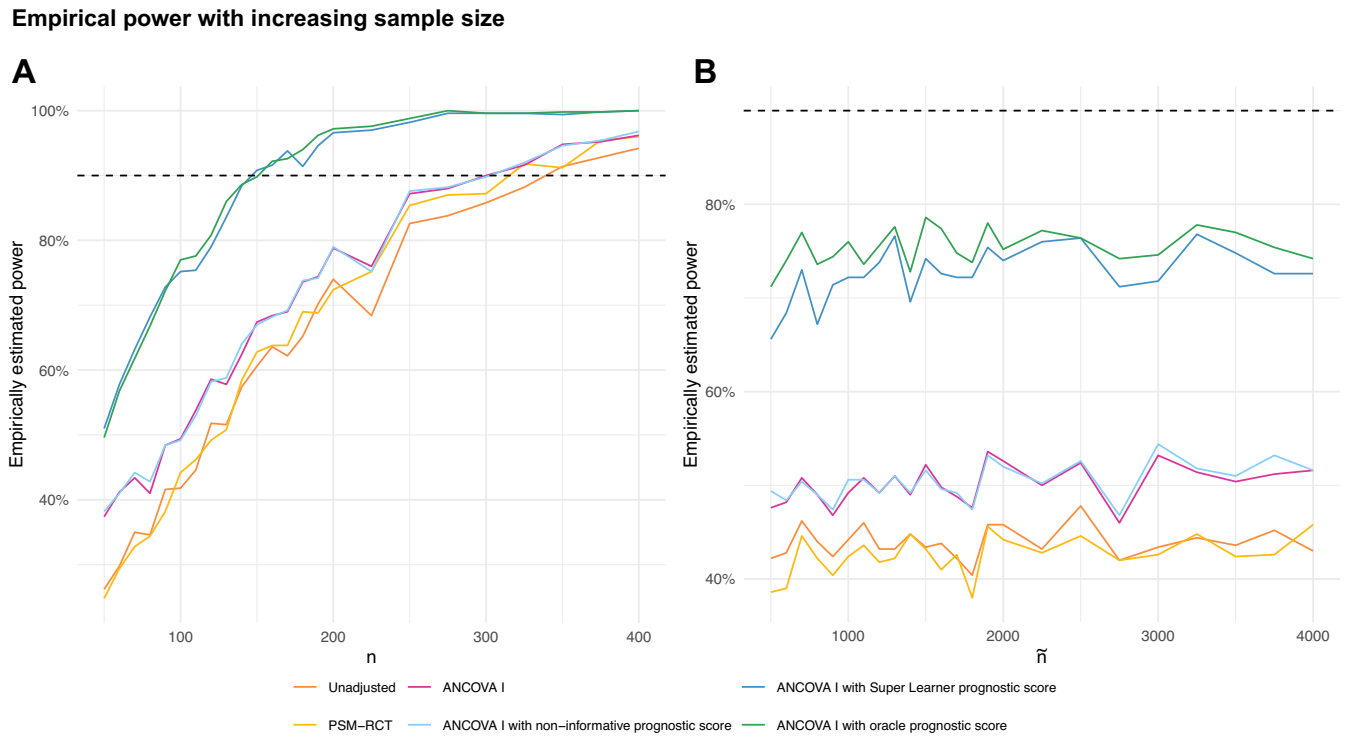

FIGURE E1 | Empirically estimated power for each model in the heterogeneous scenario. (A) varying only n with fixed $\tilde{n} = 4{,}000$ and (B) varying only $\tilde{n}$ with fixed $n = 100$.

TABLE F1 | Summary of case study data provided by Novo Nordisk A/S.

Data name	Trial ID	Duration	Titration target (mmol/L)	Blinding type	Number of participants	
					Randomized	Completed
New RCT		40 weeks	3.9–5.0	Open-label	TBD	
Test	NN9068-4228	104 weeks	4.0–5.0	Open-label	504	481
	NN9068-4229	26 weeks	4.0–5.0	Open-label	210	206
	NN1250-3579	52 weeks	4.0–5.0	Open-label	257	197
	NN1250-3586	26 weeks	4.0–5.0	Open-label	146	136
	NN1250-3672	26 weeks	4.0–5.0	Open-label	230	201
	NN1250-3718	26 weeks	4.0–5.0	Open-label	234	209
	NN1250-3724	26 weeks	4.0–5.0	Open-label	230	206
	NN1250-3587	26 weeks	4.0–5.0	Open-label	278	254
Historical	NN9535-3625	30 weeks	4.0–5.5	Open-label	365	343
	NN2211-1697	26 weeks	≤5.0	Double-blinded	34	219
	NN5401-3590	26 weeks	3.9–5.0	Open-label	264	232
	NN5401-3726	26 weeks	3.9–5.0	Open-label	Extension of 3590	209
	NN5401-3896	26 weeks	3.9–5.0	Open-label	149	137
	NN1436-4383	26 weeks	4.4–7.2	Double-blinded	122	119
	NN1436-4465	16 weeks	4.4–7.2	Open-label	51	51
	NN1436-4477	78 weeks	4.4–7.2	Open-label	492	477

Note: The new RCT data is highlighted in blue. The test data set used to determine the prospective power is highlighted in gray. The historical data consists of all the data sets that are not highlighted. The number of participants refers to the number of participants receiving the existing daily insulin treatment. TBD is short for to be determined.

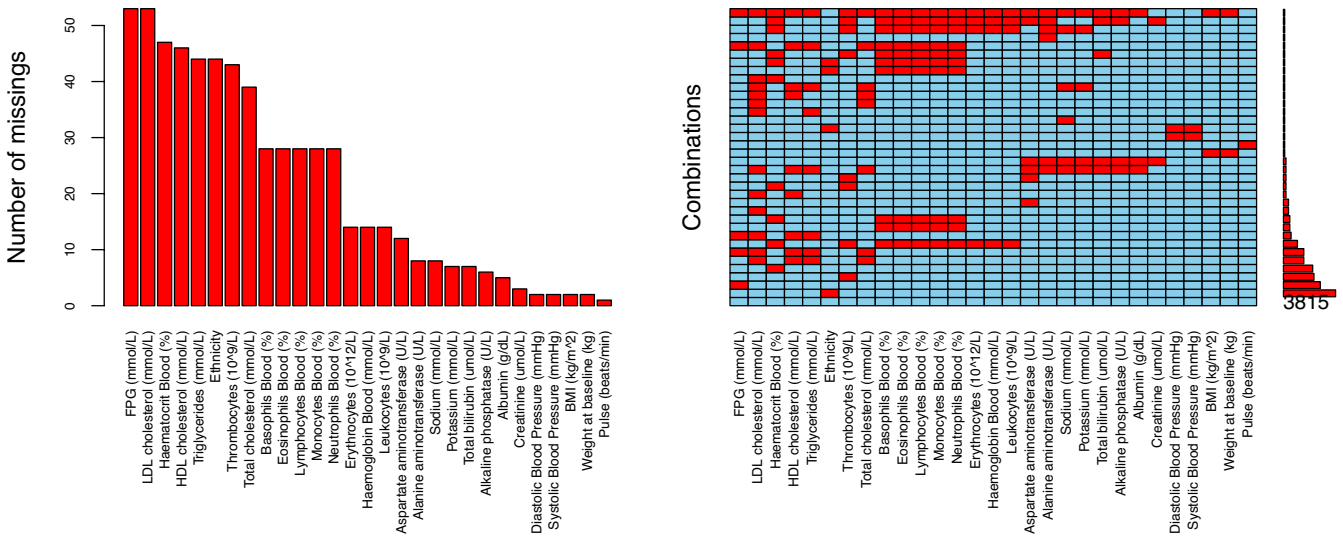

FIGURE F2 | Left: Total number of missing values for each covariate. Right: Combination pattern of missingness.

Discrete Super Learner

Number of folds: Cross-validation is used to select the best candidate learner in the library for the historical sample. A 3-fold scheme is used when the historical sample size is over 5000, a 5-fold scheme when it is over 4000, and a 10-fold scheme when it is less than 1000.

Library of learners:

- Multivariate Adaptive Regression Splines with the highest interaction of the 3rd degree
- Linear regression

- Extreme gradient boosting with specifications: Learning rate 0.1, tree depth 3, crossed with trees specified 25–500 by 25 increments
- A random forest with the number of trees found by cross validation from 25 to 500 by 25 increments
- K-nearest neighbors with number of neighbors between 3, 4, 5, 7, and 9 found by cross validation
- Lasso regression with penalty found by cross validation

Loss function: Mean square error loss.

Population Parameters for Prospective Sample Size Determination

TABLE F2 | Population parameters used in prospective sample size determination for five different models for ATE estimation.

Model	Baseline adjustment	$\sigma_y^2 \cdot 1.25$	ρ^2 or R^2
ANCOVA I	HbA1C	1.42	0.30
ANCOVA I with Super Learner prognostic score	HbA1C	1.42	0.44
ANCOVA I with Super Learner prognostic score (0.9 deflation)	HbA1C	1.42	0.40
ANCOVA I with Super Learner prognostic score (0.8 deflation)	HbA1C	1.42	0.35