Request for CHMP Qualification for Prognostic Covariate Adjustment (PROCOVA™) as an Efficient Statistical Methodology, Intended to Improve the Efficiency of Phase 2 and 3 Clinical Trials by Using Trial Subjects' Predicted Control Outcomes (Prognostic Scores) in Linear Covariate Adjustment

Unlearn.AI, Inc
75 Hawthorne Street, Suite 560
San Francisco, CA 94105

Version 2.0: 8 June 2021

Table of Contents

List of Tables

Abbreviations

AD	Alzheimer's Disease
ADAS-Cog11	Alzheimer's Disease Assessment Scale-Cognitive Subscale
ADCS	Alzheimer's Disease Cooperative Study
ADNI	Alzheimer's Disease Neuroimaging
ANCOVA	Analysis of Covariance
CRBM	Conditional Restricted Boltzmann Machine
CHMP	Committee for Medicinal Products for Human Use
CDR	Clinical Dementia Rating
CDR-SB	Clinical Dementia Rating Sum of Boxes
CODR-AD	C-Path Online Data Repository for Alzheimer’s Disease
CPAD	Critical Path for Alzheimer’s Disease
DHA	Docosahexaenoic Acid
EMA	European Medicines Agency
FDA	Food and Drug Administration
MCI	Mild Cognitive Impairment
ML	Machine Learning
MMSE	Mini-Mental State Examination
NIA	National Institute on Aging
NINDS	National Institute of Neurological Disorders and Stroke
PROCOVA™	Prognostic Covariate Adjustment
RCT	Randomized Controlled Trial
SOC	Standard of Care

Glossary

Baseline covariates	A variable which is measured or observed before randomization/start of treatment in a clinical trial, and which may influence the outcome.
Bayesian [framework]	A theory in the field of statistics based on the Bayesian interpretation of probability as a degree of belief in an event; the degree of belief may be based on prior knowledge about the event, such as the results of previous experiments.
Binary variable	A categorical variable that can only take one of two values.
Continuous variable	A numeric variable that can have an infinite number of values between any two values.
Corollary	A theorem that can be easily derived from another theorem
Deep learning model	Part of a broader family of machine learning methods based on artificial neural networks.
Estimator	A rule for calculating an estimate of a given quantity based on observed data.
Frequentist [framework]	A type of statistical inference that draws conclusions from sample data by emphasizing the frequency of events; this is the inference framework on which the well-established methodologies of statistical hypothesis testing and confidence intervals are based.
Historical data	Patient-level data collected in past/completed trials.
Pearson correlation coefficient	Measure of linear correlation between two variables.
Power	The probability of detecting a difference when one exists; the converse of the type-II error rate.
Prognostic model	A formal combination of multiple predictors from which the prediction of an outcome (prognostic score) can be calculated for individual patients.
Random forest model	Part of a broader family of machine learning methods that operates by constructing a multitude of decision trees during training.
Theorem	A statement that can be demonstrated to be true by accepted mathematical operations and arguments.
Time-to-event outcome	An outcome that takes into account whether an event takes place and also the time at which the event occurs, also known generically as survival data.
Type-I error rate	The probability of erroneously approving an ineffective or unsafe drug or device.
Type-II error rate	The probability or erroneously disapproving a safe and effective drug or device.

1. Executive Summary

Our objective is to seek CHMP qualification for the proposed statistical methodology intended to improve the efficiency of Phase 2 and 3 clinical trials, by using trial subjects' predicted outcomes on placebo (prognostic scores) in linear covariate adjustment; such prognostic scores can be generated using a predictive model trained on historical data. Our approach is efficient in the sense that it uses historical data to reduce variance of the treatment response estimates (and thus reduce the minimum sample size required to achieve the desired level of confidence) better than other available approaches.

Our proposed statistical methodology, called prognostic covariate adjustment or PROCOVA™, leverages historical data (from control arms of clinical trials and from observational studies) and predictive modeling to decrease the uncertainty in treatment effect estimates from Phase 2 and 3 Randomized Controlled Trials (RCTs) measuring continuous responses, in the large-sample setting. Our methodology comprises these three steps:

Step 1: Training and evaluating a prognostic model to predict control outcomes (generate prognostic score).

Step 2: Accounting for the prognostic score while estimating the sample size required for a prospective study.

Step 3: Estimating the treatment effect from the completed study using a linear model while adjusting for the control outcomes predicted by the prognostic model.

The last step amounts to adding a single (constructed) adjustment covariate into an adjusted analysis. As such, it poses no additional statistical risk over any other pre-specified adjusted analyses.

This methodology is recommended for use in trials with continuous variables for which there is historical data on the patient population in question, such that one can build a prognostic model to predict control outcomes (generate prognostic score) with sufficient accuracy, given the subjects' measured baseline covariates. Therefore, the variables used by the prognostic model must be measured at baseline for all subjects (and a missing data imputation scheme should be pre-specified).

Our procedure can utilize a prognostic score generated by any prognostic model, including mechanistic models, linear statistical models, as well as machine-learning-based methods as described in this submission. The latter are particularly useful as the machine-learning-based methods can learn non-linear predictive models from large databases. In addition, the construction of the prognostic model may be outsourced to machine learning experts, with access to the historical but not the trial dataset. In fact, the historical data can be used to train the prognostic model with guaranteed protection of private health information[1,2].

PROCOVA™ represents a special case of Analysis of Covariance (ANCOVA), in that once the prognostic score has been calculated, the analysis is a standard linear regression. This makes it simple to implement with existing software, and easy to explain, interpret, and incorporate into various analysis plans. We provide a simple formula that can be used to calculate power prospectively while accounting for the beneficial effect of prognostic score adjustment.

We show that PROCOVA™ is optimal if the prognostic model attains the maximal possible correlation with the actual outcomes of subjects under control conditions. However, one can realize gains in efficiency even with imperfect prognostic models. The other important advantage of PROCOVA™ is that it involves an adjustment for a single covariate derived from a larger set of variables that constitute the input of a prognostic model, providing a

substantial dimensionality reduction. Even if the input to the prognostic model is high-dimensional in nature (e.g., a brain image, or a whole transcriptome), PROCOVA™ still represents an adjustment for a single covariate. One only has to measure the Pearson correlation of this single covariate with the actual outcome in a similar historical population in order to account for the prognostic score in a prospective sample size estimation for a planned trial. We present mathematical proof and an actual demonstration of a prospective application of PROCOVA™ to power a trial without estimating or assuming a large number of population parameters.

In summary, our method is scientifically sound since it only adjusts for a single covariate derived from information collected at baseline/prior to randomization; produces unbiased estimates for treatment effects; controls the type-I error rate; and leads to correct confidence interval coverage. It is also consistent with current FDA and EMA regulatory guidance.

We demonstrate that PROCOVA™ is a robust methodology to optimize both the design and analysis of RCTs with continuous responses, in prospective context-of-use represented by the following two empirical examples:

Experiment 1. Pre-specified primary analysis of Phase 2 and 3 trials, to deliver higher power/confidence in the results compared to unadjusted analyses.

Experiment 2. Prospective design/sample size estimation for Phase 2 and 3 trials, to attain the desired level of power/level of confidence with a smaller sample size compared to unadjusted trials.

To demonstrate the flexibility of our approach with regard to the prognostic model, we utilize prognostic scores generated by two different models: a random forest and a deep learning model trained on historical data from clinical trials and observational studies.

While our methodology is applicable to in-scope trials in any therapeutic area where historical control data are available, we have chosen Alzheimer's Disease (AD) as our initial target. The predictive models described in this submission were constructed on historical data from AD trials contained in two different AD databases, and our empirical demonstrations involve re-analysis of a Phase 3 trial in patients with AD.

2. Statement of the Need for and Impact of the Proposed Novel Methodologies in Clinical Drug Development

2.1 Background

The goal of much clinical research is to estimate the effect of a treatment on an outcome of interest (causal inference). The RCT is the gold standard for causal inference because randomization cancels out the effects of any unobserved confounders in expectation. However, clinical research must still contend with the statistical uncertainty inherent to finite samples. Because of this, methods for the analysis of trial data are chosen to safely minimize this statistical uncertainty about the causal effect.

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

As most clinical trials compare an active treatment to a placebo (often against the background of standard-of-care (SOC), which all trial participants receive), there is a possibility to use existing historical control arm data from completed trials to reduce variance and decrease sample size. Even in the case of an active control, data from patients receiving the active control can often be obtained from historical or real-world sources. Such "historical borrowing" methods are becoming increasingly attractive especially with the recent creation of large, electronic patient datasets that can make it easier to find a suitably matched historical population.

Various approaches to historical borrowing have been proposed and their properties extensively evaluated, ranging from directly inserting subjects from previous studies into the current sample, to using previous studies to derive prior distributions for Bayesian analyses [3–6]. Although such methods do generally increase power, they cannot strictly control the type-I error rate [3,5,7] reducing the relevance of such methods, particularly for pivotal/ confirmatory/ Phase 3 RCTs [8]. A common approach to addressing the risk of type-I error rate inflation when information is borrowed is to carry out multiple simulation studies to quantify this effect.

2.2 The Novel Methodology

We propose a novel approach that leverages historical control arm data and predictive modeling to decrease the uncertainty in treatment effect estimates from RCTs without compromising strict type-I error rate control in the large-sample setting. Our methodology comprises these three steps:

Step 1: Training and evaluating a prognostic model to predict control outcomes. We define a prognostic model as a mathematical function of a subject's baseline covariates that predicts the subject's expected outcome if that subject were to receive the control treatment in the planned trial (e.g., placebo). The output of the prognostic model for a given subject is called that subject's prognostic score.

Step 2: Accounting for the prognostic model while estimating the sample size required for a prospective study.

Step 3: Estimating the treatment effect from the completed study using a linear model while adjusting for the control outcomes predicted by the prognostic model.

The last step amounts to adding a single (constructed) adjustment covariate into an adjusted analysis. As such, it poses no additional statistical risk over any other pre-specified adjusted analyses (which are preferable to unadjusted analyses in almost every case [9–12]). Our approach is entirely pre-specifiable, is generic enough to be integrated into many analysis plans and is supported by regulatory guidance [13,14].

Our procedure is flexible with respect to the prognostic model used to generate predicted control outcomes (e.g., on placebo) for the trial subjects and maintains type-I error rate control regardless of the type of such model. In this submission, we present results employing two different predictive models - random forests and a deep learning model [18–21] (Appendix 6). Deep learning models are particularly well suited to handle such common clinical trial challenges as missing covariates, multiple longitudinal outcomes, and high-dimensional covariates (e.g., a whole genome). Deep learning methods can also combine data from multiple sources to improve performance when the relevant historical data are meager [22]. In addition, the construction of the prognostic model may be outsourced to a group of machine-learning experts, which also makes it possible to separate access to the historical and trial datasets. In fact, the historical data can be used to train a prognostic model within a privacy preserving framework with guaranteed protection of private health information [1,2,23].

Adjustment for composite or computed covariates such as body mass index, Charlson comorbidity index, or Framingham risk score, is not new [9,11,15–17]. These “indices” or “scores” are usually the output of a simplified prognostic model derived from historical data. For instance, the Framingham cardiovascular risk score was developed by training Cox and logistic regression models using a large community-based cohort to obtain a single covariate that is highly predictive of cardiovascular outcomes. From that perspective, our proposed approach is a formalization of what has previously been an ad-hoc procedure.

A number of recent technological developments have led to substantial improvements in the ability to train highly accurate prognostic models. First, large databases of longitudinal patient data from control arms of historical clinical trials, observational and natural history studies, and real-world sources have become widely available. Second, high dimensional biomarkers from technologies such as imaging and next generation sequencing provide large amounts of patient-level information. And, third, improvements in machine learning methods (especially in the subfield known as deep learning) allow one to create prognostic models that can fully utilize all of these patient data. The intersection of these three key developments — large, analyzable databases containing high-dimensional outcomes, and powerful deep learning models — allows for the generation of more predictive prognostic scores, adjusting for which can substantially reduce variance/confidence intervals, and/or increase power and reduce minimum required sample sizes.

2.3 Objective, Scope and Context-of-use

The objective of this submission is to seek CHMP qualification for the proposed statistical methodology intended to improve the efficiency of Phase 2 and 3 clinical trials by using trial subjects’ predicted control outcomes (prognostic scores) in linear covariate adjustment (PROCOVA™); such prognostic scores can be generated from each subject’s baseline characteristics using a predictive model trained on historical data. Our approach is efficient in the sense that it uses historical data to reduce variance of the treatment response estimates (and thus the minimum sample size required to achieve the desired level of confidence) better than other methods with access to the same baseline covariates.

In this submission, we present mathematical (Section 3.1.2), simulation (Section 3.2), and empirical (Section 3.3) demonstrations that PROCOVA™ is an effective and safe method for leveraging historical data to reduce uncertainty in RCTs. Once the prognostic score has been calculated, the analysis is a standard linear regression. This makes it suitable under current regulatory guidance,[13,14] simple to implement with existing software, and easy to explain and interpret. In comparison to other kinds of historical borrowing methods, PROCOVA™ guarantees unbiased estimates, strict type-I error rate control, and confidence interval coverage, as proven theoretically and demonstrated through simulations in this submission. In anything but the smallest of trials, there is no need for elaborate simulations to demonstrate the trial operating characteristics (as is usually the case for methods that cannot theoretically guarantee control of type-I error). Finally, we provide a simple formula that can be used to calculate power prospectively while benefiting from prognostic score adjustment.

We demonstrate that PROCOVA™ is a robust methodology to optimize both the design and analysis of Phase 2 and 3 RCTs with continuous responses, in prospective context-of-use represented by the following two empirical examples:

Experiment 1. Pre-specified primary analysis of Phase 2 and 3 trials, to deliver higher power/confidence in the results compared to unadjusted analyses.

Experiment 2. Prospective design/sample size estimation for Phase 2 and 3 trials, to attain the desired level of power/level of confidence with a smaller sample size compared to unadjusted trials.

To demonstrate the flexibility of our approach with regard to the prognostic model, we utilize prognostic scores generated by two different models: a random forest and a deep learning model trained on historical data from clinical trials and observational studies.

Our methodology is intended for use in RCTs with continuous responses. When applied to such trials, PROCOVA™ offers two critically important advantages over other approaches. First, it can attain the lowest variance among reasonable analytical approaches with access to the same covariates if the prognostic model is "perfect", i.e., if the computed prognostic score for a subject is equal to his/her actual outcome on control treatment, given his/her baseline covariates. Second, PROCOVA™ is an adjustment for a single covariate derived from a larger set of variables that constitute the input of a prognostic model, providing a substantial dimensionality reduction. Even if the input to the prognostic model is high-dimensional in nature (e.g., a brain image, or a whole transcriptome), PROCOVA™ still represents an adjustment for a single covariate. One only has to measure the Pearson correlation of this single covariate with the actual outcome in a historical population similar to that of the planned trial in order to account for the prognostic score in a prospective sample size estimation.

While our methodology is applicable to in-scope trials in any therapeutic area where historical control data are available, we have chosen Alzheimer's Disease (AD) as our primary initial target because of an exceptionally high, and growing, unmet need; challenging, long and large Phase 2/3 trials; abundant placebo control data from over 150 randomized clinical trials and many observational studies conducted since the 1990's; and largely unchanged SOC and the clinical trial endpoints for symptomatic AD over the last 17 years (ensuring small or no temporal drifts in the data). As such, the predictive models described in the simulations (Section 3.2) and empirical examples/context-of-use (Section 3.3) parts of this submission were constructed on historical data from AD trials contained in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database and the Critical Path for Alzheimer's Disease (CPAD) database (Appendix 5). Our empirical

context-of-use demonstrations involve re-analysis of a Phase 3 trial in patients with AD reported by Quinn et al. [24].

2.4 Out-of-Scope/Future Directions

Several aspects of the proposed methodology are beyond the scope of this submission. For example, it may be possible that prognostic score adjustment retains a statistical advantage relative to direct nonlinear adjustment in trials with other types of response variables including binary variables or time-to-event outcomes, though we have left theoretical investigation of this question to future studies.

Similarly, the estimand targeted by PROCOVA™ as described in this submission, is the difference in the counterfactual population means of a continuous outcome (this is the exact estimand that is targeted by the unadjusted estimator in this setting). Estimands for other types of outcomes are less straightforward and will be considered for further research beyond the scope of this submission.

It should also be possible to combine the advantages of multiple procedures, i.e., to perform adaptive adjustment for a fixed prognostic model trained on historical data.

In addition, the particular choice of prognostic model, and the method used to train it, are beyond the scope of this submission. One of the primary benefits of PROCOVA™ is that it guarantees type-I error rate control for *any* prognostic model, thus separating the concerns of how to build a highly predictive model from how to apply the predictions from a model to maximize power in an RCT. Moreover, the only requirement for prospective powering is the ability to estimate the performance of the prognostic model in the target population.

In the future, PROCOVA™ may be exploited as a component in other kinds of estimators (generalized estimating equation, generalized linear model, survival models etc.). We have limited our theoretical discussion here to the linear model for continuous responses since it is so common, but a prognostic score may be used as a covariate in any analysis that allows for covariate adjustment. In addition, we have limited our discussion to analyses of a single timepoint, but prognostic scores could also be used in analyses with repeated measures. It remains to be seen what optimality properties are satisfied by doing prognostic covariate adjustment in each kind of analysis and under what conditions.

Similarly, one may account for heterogeneous treatment effects by including treatment-by-covariate interactions while estimating the treatment effect. Indeed, some theoretical properties of PROCOVA™ including treatment-by-covariate interactions are presented in Schuler et al. [25]. However, this particular submission describes the use of PROCOVA™ without treatment-by-covariate interactions, in line with the EMA's guidelines on adjustment for baseline covariates in clinical trials [13].

Finally, while this submission is focused exclusively on RCTs with strict type-I error rate control (i.e., in a frequentist framework), we are in the process of developing a Bayesian framework that combines prognostic covariate adjustment with an empirical prior distribution learned from the predictive performances of the prognostic model on past trials [26]. We have shown theoretically that Bayesian PROCOVA™ offers a substantial further increase in statistical power compared to frequentist PROCOVA™, while limiting the type-I error rate under reasonable conditions.

2.5 Preview of the Technical Aspects Detailed in Methods and Results

In the next section, we provide a detailed description of PROCOVA™ and present mathematical proofs of its main statistical properties (Section 3.1). Specifically, we prove that estimates of treatment effects obtained with PROCOVA™ are unbiased and that type-I

error rates of hypothesis tests are controlled at the pre-specified level. These results hold for PROCOVA™ use with any prognostic model. In addition, we prove that PROCOVA™ can attain the maximum power of any estimator with access to the pre-specified baseline covariates if the prognostic model is exact — that is, PROCOVA™ is the optimal estimation procedure if the computed prognostic score for a subject is equal to his/her actual expected outcome under control conditions, given his/her baseline characteristics. In addition, we provide a simple formula to estimate the power/minimum sample size in a prospective trial that will be analyzed with PROCOVA™.

We then describe and quantify the procedure's performance, by demonstrating the efficiency gain associated with the use of PROCOVA™ via several simulations (Section 3.2). These explore how the mean-squared estimation error of the treatment effect varies with and without prognostic covariate adjustment in four scenarios: when the covariate-outcome relationship is linear, when the covariate-outcome relationship is nonlinear, when the treatment effect is heterogeneous, and when the prognostic model is trained on a dataset with different properties from the trial population. We conduct these simulations first using PROCOVA™ alone, and then repeat them for PROCOVA™ combined with standard adjustment for baseline covariates. We show that prognostic covariate adjustment decreases the mean-squared error of the estimated treatment effects in all scenarios, with one exception. There is no change to the mean-squared error when the simulated outcome is a simple linear combination of baseline covariates which are also used individually for standard covariate adjustment.

Next, we present an empirical demonstration of PROCOVA™ through re-analyses of a completed Phase 3 trial in patients with AD, in order to illustrate different benefits of PROCOVA™ (Section 3.3). The first experiment demonstrates that, using the same sample size and randomization ratio as in the original study, adjusting for prognostic scores decreases the magnitude of the estimated standard errors and the width of the confidence intervals. The second experiment demonstrates that accounting for the prognostic scores during sample size estimation results in a trial with fewer subjects but with standard errors of equal magnitude to those in a larger trial designed without PROCOVA™.

We perform these re-analyses using two different types of ML models to generate prognostic scores (Appendix 6), a random forest and a deep learning model (specifically, a Conditional Restricted Boltzmann Machine, or CRBM), in order to emphasize that PROCOVA™ can be applied with different types of prognostic models.

3. Methodology and Results

3.1 The Prognostic Covariate Adjustment (PROCOVA™) Method

Here we describe in detail the steps for using PROCOVA™ to estimate the treatment effect in an RCT and to perform a sample size calculation. We present the mathematical properties of the proposed procedure in a series of theorems, with mathematical proofs and technical details provided in Appendix 1, Appendix 2, and Appendix 3.

3.1.1 Description of PROCOVA™

Our proposed method, Prognostic Covariate Adjustment (PROCOVA™), consists of the following three general steps, described in further detail in Appendix 1:

Step 1: Training and evaluating a prognostic model to predict control outcomes/generate prognostic scores.

We define a prognostic model as a mathematical function of a subject's baseline covariates that predicts the subject's expected outcome if that subject were to receive the control treatment in the planned trial (e.g., placebo). The output of the prognostic model for a given subject is called that subject's prognostic score.

In principle, there are many ways to obtain a prognostic model. The type-I error rate will be controlled for any type of model, whereas the realized increase in trial efficiency will depend on the predictive performance of the model in the target population, defined here and below as subjects meeting the selection criteria in the trial of interest. Machine learning-based methods are especially effective in fitting the model to a collection of historical data and linking subjects' baseline covariates to their outcomes under the control condition. We provide two examples of this type of prognostic model in our empirical analyses.

The minimum sample size required to detect a given effect using PROCOVA™ is a function of the Pearson correlation coefficient between the observed and predicted outcomes in the target population, in addition to the target effect size and the variance of the outcome. The larger the correlation, the smaller the minimum sample size. Therefore, the Pearson correlation coefficient should be estimated using a *separate* set of historical data linking subjects' baseline covariates to their actual outcomes under the control condition, one that was not used to train the prognostic model. The subjects in this historical dataset should have similar baseline characteristics to those in the target population (e.g., they should meet the subject selection criteria of the planned trial). The same dataset can be used to estimate the variance of the outcome.

Step 2: Accounting for the prognostic model while estimating the sample size required for a prospective study.

For a given sample size, an analysis that uses PROCOVA™ will have higher power than an analysis that does not use PROCOVA™. Similarly, a given target effect size can be detected with a smaller sample size in an analysis that uses PROCOVA™ than in an analysis that does not use PROCOVA™. The minimum sample size for a trial can be estimated once the following parameters have been defined: the target effect size, the significance threshold, the desired power level, the proportion of subjects to be randomized to the active treatment arm, and the expected dropout rate. In addition, we need the estimates for the correlation between the prognostic scores and the actual outcomes in the target population as defined in Step 1 above, and the variance of the observed outcomes from Step 1. In many cases, the sponsor of the clinical trial may conservatively choose a correlation that is slightly smaller than estimated, and/or a variance that is slightly larger than estimated, in order to ensure the

planned trial has sufficient power. Typically, these parameters are assumed to be the same for the active treatment and control groups.

With the above parameters now defined, we find the smallest sample size that will achieve the desired power to detect the target effect size. If there are multiple outcomes of interest, such as co-primary endpoints, each with a desired power level and target effect size, then this procedure must be repeated for each outcome, and the largest sample size should be selected. This may require the use of multiple prognostic models (i.e., one to predict each outcome of interest) or a multivariate prognostic model.

Step 3: Estimating the treatment effect from the completed study using a linear model while adjusting for the control outcomes predicted by the prognostic model.

An RCT is performed using its originally estimated minimum sample size, in which each subject is randomized to active treatment or control. Data from subjects who have dropped out of the study should be handled with an appropriate pre-specified method as in any trial analysis [27]. Next, the treatment effect is estimated by fitting a linear model, while adjusting for the estimated prognostic scores. One could also adjust for additional covariates in the regression if desired, so long as the sample size is much greater than the total number of terms in the linear model.

Finally, a null hypothesis (e.g., no treatment effect) can be assessed by computing a two-sided p-value. The null hypothesis is rejected with a two-sided significance test at significance level α if $p < \alpha$.

The PROCOVA™ method described above is a special case of Analysis of Covariance (ANCOVA) with a particular choice of adjustment covariate. As such, PROCOVA™ inherits the statistical properties of ANCOVA; for example, estimated treatment effects will be unbiased and the type-I error rate will be controlled. For these reasons, ANCOVA is widely used in the analysis of clinical trials with continuous responses and is supported by guidance from EMA [13] and draft guidance from FDA [14]. These statistical properties hold for PROCOVA™ using any prognostic model, regardless of the approach to modeling or the data used to inform the model.

It is well known that ANCOVA can improve power in clinical trials if there is a correlation between the outcome and the adjustment covariate. PROCOVA™ is motivated by the fact that the covariate which is most correlated with the outcome is the prediction for the outcome itself. That is, rather than adjusting for a raw baseline covariate, we construct the optimal adjustment covariate. Under certain conditions outlined below, we show that adjusting for the prognostic score in a linear model to estimate the treatment effect achieves the minimum variance among appropriate analytical approaches with access to the same baseline covariates. The mathematical (Section 3.1.2), simulations (Section 3.2), and empirical (Section 3.3) results presented below, demonstrate that, for a given sample size, PROCOVA™ can lead to substantial increases in power without sacrificing control of the type-I error rate. In addition to the traditional assumptions regarding the target effect size, the significance threshold, the desired power level, etc., one only has to measure the Pearson correlation of a single prognostic covariate with the actual outcome in a historical population similar to that of the planned trial in order to account for the prognostic score in a prospective sample size estimation.

3.1.2 Mathematical Results

3.1.2.1 Mathematical Properties of ANCOVA

PROCOVA™ is a special case of an Analysis of Covariance (ANCOVA). As a result, all of the statistical properties of ANCOVA also apply to PROCOVA™. We provide a short review of important properties of ANCOVA, with mathematical details described in Appendix 2, and technical proofs in Appendix 3.

ANCOVA can be used to estimate a treatment effect from an RCT by fitting the linear model while adjusting for a treatment indicator variable, and any other covariates that were measured at or before baseline. The coefficient of the regression on the primary endpoint is an estimate of the treatment effect. The coefficients on the other endpoints or covariates aren't necessarily important, but including those covariates can decrease the uncertainty in the estimate for the treatment effect.

For adjusted estimation based on linear models or generalized linear models, the recently updated draft FDA guidance[14] recommends that sponsors estimate standard errors using the Huber-White robust "sandwich" estimator or the nonparametric bootstrap method, rather than using nominal standard errors. We chose to estimate the standard errors in the regression coefficients using the Huber-White estimator, which is robust to heteroscedasticity.

The following mathematical theorems establish statistical properties of ANCOVA and, as a result, of PROCOVA™. Here, we only present descriptions and implications of the mathematical theorems, leaving rigorous proofs and results to Appendix 2.

Theorem 1:

We consider an ANCOVA analysis in which the adjustment covariates are computed by applying an arbitrary transformation to the raw baseline covariates. We show that the estimate of the treatment effect obtained with ANCOVA is unbiased for any reasonable transformation of the baseline covariates. Moreover, the variance of the estimated treatment effect depends on the covariances between the treatment and control potential outcomes with the transformed baseline covariates. This Theorem has several important corollaries listed below. Both the theorem and the corollaries are described in detail in Appendix 2.

Corollary 1.1 implies that the type-I error rate is controlled using ANCOVA with any reasonable transformation of the baseline covariates.

Corollary 1.2 provides a simple formula to compute the expected power of an ANCOVA analysis, as long as the relevant parameters in the formula for the variance given in Theorem 1 can be estimated.

Corollary 1.3 demonstrates that the formula for the variance of the estimated treatment effect is simplified if the baseline covariates are transformed into a one-dimensional variable. This is useful for prospective power calculations, because it substantially reduces the number of parameters that need to be estimated in order to estimate the minimum sample size required in a future study.

Corollary 1.4 demonstrates that adjusting for a covariate in a trial with equal randomization always decreases the variance of the estimated treatment effect, for any transformation of the baseline covariates into a one-dimensional variable.

Use of ANCOVA is facilitated by the fact that the resulting estimates of treatment effects are unbiased, and type-I error rates of hypothesis tests are controlled. In addition, using ANCOVA always increases power in randomized trials with equal randomization. Therefore,

we propose to choose the transformation that maximizes statistical power, which is PROCOVA™.

3.1.2.2 Mathematical Properties of PROCOVA™

PROCOVA™ is motivated by the theorem presented below, with detailed results provided in Appendix 2 and Appendix 3.

Theorem 2:

If the treatment effect is constant, then the optimal covariate to adjust for in ANCOVA is a prediction of the potential control outcome for a subject, based on that subject's observed baseline covariates. That is, adjusting for a prediction of the potential control outcome minimizes the variance of the estimated treatment effect. These and other related considerations are presented in a more general context elsewhere[25].

An RCT analyzed with PROCOVA™ borrows information from a historical dataset to construct a covariate which, when adjusted for in a regression, minimizes the variance of the estimated treatment effect. As a result, it also maximizes the statistical power of the trial to detect a given effect. If the prognostic model used to predict the control potential outcomes is accurate (i.e., it obtains a high correlation with actual outcomes), then this method obtains the maximum power of any linear analysis using the same baseline covariates that does not include treatment-by-covariate interactions.

A number of recent technological developments have led to substantial improvements in the ability to train highly accurate prognostic models. First, large databases of longitudinal patient data from control arms of historical clinical trials, observational and natural history studies, and real-world sources have become widely available. Second, high dimensional biomarkers from technologies such as imaging and next generation sequencing provide large amounts of information about individual patients. And, third, improvements in machine learning methods (especially in the subfield known as deep learning) allow one to create prognostic models that can fully utilize all of these patient data. The intersection of these three key developments — large, analyzable databases containing high-dimensional outcomes, and powerful deep learning models — allows for the generation of more predictive prognostic scores, adjusting for which can substantially reduce variance/confidence interval, and/or increase power and reduce minimum required sample sizes, as shown in Section 3.2 and Section 3.3.

3.2 Simulation Studies of PROCOVA™

We demonstrate that PROCOVA™ provides more precise estimates of treatment effects than unadjusted estimators in realistic simulated scenarios. By using simulations, we are able to specify the data generating distribution and treatment effect. Since the treatment effect is known, the discrepancy between the estimated and actual treatment effects can be directly measured. Specifically, we used simulation studies to explore how mean-squared estimation error of the treatment effect varies with and without PROCOVA™.

3.2.1 Simulation Study Methods

We simulated four different scenarios that model realistic situations encountered in clinical trials, and that enable us to probe the sensitivity of PROCOVA™ to particular assumptions.

- The Linear simulation describes a scenario in which the outcome-covariate relationship is linear in both the active and control treatment arms with a constant treatment effect.

- The Non-linear simulation describes a scenario in which the outcome-covariate relationship is non-linear in both treatment arms, but the treatment effect is constant.
- The Heterogeneous simulation describes a scenario in which the conditional average effect $\mathrm{E}[Y_1 - Y_0|X] = \mu_1(X) - \mu_0(X)$ is not constant (i.e., $\mathrm{E}[Y_1 - Y_0|X] \neq \mu_1(X) - \mu_0(X)$).
- The Shifted simulation describes a scenario in which the historical population used to train the prognostic model is not representative of the trial population in terms of the baseline covariates (i.e., $P_H(X' = x) \neq P(X = x)$).

Details on the data generating process for each of the simulation scenarios are provided in Appendix 4.

The first two simulation scenarios, covering Linear and Non-linear outcome-covariate relationships, fall under the assumptions in our theoretical results. Therefore, we expect PROCOVA™ to perform well, as long as we use a prognostic model capable of capturing non-linear relationships. In contrast, the Heterogeneous scenario violates the constant treatment effect assumption of Theorem 2, so this scenario probes the sensitivity of PROCOVA™ to that assumption. Although the fourth scenario does not violate any of our assumptions, a prognostic model trained on the simulated historical data in the Shifted scenario may not generalize well to the simulated study population. Therefore, this scenario probes the sensitivity of PROCOVA™ to the predictive performance of the trained prognostic model.

In each simulation scenario, we generated a simulated historical control dataset *and* trained a random forest as a prognostic model. Then, we simulated a randomized trial dataset with 500 subjects randomized 1:1 to the active treatment and control. Finally, we used the prognostic model to generate an estimated prognostic score, and *also* computed the exact prognostic score (i.e., the expected control outcome) using the simulated data generating process. The exact prognostic score represents the performance that could be obtained with a "perfect" prognostic model but, because a random forest is unlikely to learn the *exact* relationship, we expect the estimated prognostic score to perform slightly worse than the exact prognostic score.

We analyzed the data using three estimation procedures: unadjusted, adjusted with the estimated prognostic score obtained with the random forest, and adjusted with the exact prognostic score. The three estimation procedures were repeated for models with and without additional baseline covariates included. Finally, we calculated the squared-error of each estimate relative to the true treatment effect, which is known because it was used to generate the simulated data, repeated this process 10,000 times, and averaged the squared-errors to obtain mean-squared errors for each analysis.

3.2.2 Simulation Study Results

Table 1 and Table 2 present the results obtained in each of the 4 chosen scenarios, including Linear and Non-linear outcome-covariate relationships, both of which can be learned by the random forest prognostic model, and the Heterogeneous and Shifted scenarios, which probe the sensitivity of PROCOVA™ to the violation of the Theorem 2 assumption regarding constant treatment effect, and to the accuracy of the prognostic model, respectively. The two tables differ in that Table 1 does not include any additional covariates besides the prognostic score, while Table 2 includes additional baseline covariates. The Table lists the mean-squared errors of estimated treatment effects obtained in unadjusted analysis; analysis using

adjustment for an estimated prognostic score: and analysis using adjustment for an exact prognostic generated by a "perfect" prognostic model as described above.

Table 1. Mean-squared errors of estimated treatment effects computed from simulations with no additional covariates

Scenario	Unadjusted Analysis	Adjustment for estimated prognostic score	Adjustment for exact prognostic score
Linear	3.49	0.96	0.82
Non-linear	7.73	1.85	0.82
Heterogeneous	5.54	2.32	2.32
Shifted	7.65	6.79	0.82

Table 2. Mean-squared errors of estimated treatment effects computed from simulations with additional baseline covariates

Scenario	Analysis adjusted only for additional covariate	Adjustment for estimated prognostic score and additional covariate	Adjustment for exact prognostic score and additional covariate
Linear	0.84	0.84	0.84
Non-linear	5.11	1.82	0.83
Heterogeneous	2.98	2.19	1.98
Shifted	5.00	4.86	0.83

In agreement with our theoretical results, the mean-squared errors of the analysis with PROCOVA™ were always smaller than or equal to the mean-squared errors without it. In fact, with the exception of the simple linear relationship with additional covariates, the mean-squared errors were substantially smaller with PROCOVA™ and, as expected, using the exact prognostic score always produced a lower mean-squared error than using the estimated prognostic score. The results of the third scenario demonstrate that PROCOVA™ can decrease the mean-squared estimation error even when the assumption of Theorem 2 regarding constant treatment effect is violated. Thus, PROCOVA™ is generally a robust technique for estimating treatment effects from RCTs.

PROCOVA™ provides the largest increases in power when the prognostic model accurately predicts the expected control outcomes in the study population. However, statistical and machine learning-based methods for fitting predictive models may overfit to the population in the training data; leading to a scenario in which the predictive model has a much larger correlation with observed outcomes in the training dataset than in the study population. The shifted scenario illustrates this phenomenon. In this scenario, PROCOVA™ still provides unbiased estimates, type-I error rate control, and decreases the variance of the estimated treatment effect. However, the increase in precision is not as large as could have been obtained with a model that generalized better to the target population. Therefore, while development and validation of the prognostic model to ensure that it achieves good performance in the target population is not necessary to ensure type-I error rate control, it is needed to maximize the efficiencies gained through application of PROCOVA™.

The following simple rules-of-thumb help understand the impact of adjusting for the prognostic score on the trial power:

$$\frac{\text{Variance with PROCOVA}}{\text{Variance without PROCOVA}} \sim 1 - R^2$$

$$\frac{\text{Power with PROCOVA}}{\text{Power without PROCOVA}} \sim 1 + (R^2/2)$$

$$\frac{\text{Minimum sample size with PROCOVA}}{\text{Minimum sample size without PROCOVA}} \sim 1 - R^2$$

Above, R^2 is the squared correlation coefficient between the prognostic scores and actual control outcomes; "with PROCOVA™" means adjusting for the prognostic score; and "without PROCOVA™" means not adjusting for the prognostic score. These rules-of-thumb are not rigorous as the exact ratios depend on various aspects of the trial design. Nevertheless, they provide an idea of the magnitude of the increases in power which can be achieved by applying PROCOVA™ with an advanced prognostic model.

To apply these rules-of-thumb, using a prognostic score with an $R = 0.5$ provides a 25% decrease in variance. Similarly, using a prognostic score with an $R = 0.8$ yields around 64% decrease in variance. Obtaining such correlations is quite realistic with current technologies, driven by the development of large clinical databases and novel machine learning technologies that enable the development of advanced prognostic models.

3.3 Empirical Applications of PROCOVA™

We illustrate the proposed prospective context-of-use for PROCOVA™ through re-analyses of a previously completed clinical trial investigating the effect of docosahexaenoic acid (DHA) on cognitive and functional decline in subjects with mild-to-moderate AD, referred to below as the demonstration trial[24]. First, using two different prognostic models trained on historical data, we illustrate that using PROCOVA™ to add a prognostic covariate to the analyses of this RCT decreases the variance of the treatment effect estimates (*Experiment 1*). Next, using the same prognostic models, we illustrate that PROCOVA™ enables the design of substantially smaller clinical trials with the same statistical power (*Experiment 2*). We use two prognostic models to demonstrate that PROCOVA™ is a general statistical technique that is not tied to a particular type of prognostic model.

3.3.1 Empirical Analyses Methods

We obtained a set of historical controls by combining data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [28] and the Critical Path for Alzheimer's Disease (CPAD) [29,30] The combined dataset was composed of data from 6,919 subjects with early-stage Alzheimer's Disease. Importantly, the historical dataset did not contain data from the demonstration trial. Two different prognostic models were trained to predict control potential outcomes using the ADNI and CPAD datasets: a random forest [31], and a deep learning model [18,32]. For our demonstration, we focused on the 18-month changes in the Alzheimer's Disease Assessment Scale - Cognitive Subscale (ADAS-Cog11) [33] and the Clinical Dementia Rating (CDR) [34]. More details on the training data and the prognostic models are provided in Appendix 5 and Appendix 6.

The demonstration trial was originally performed through the Alzheimer's Disease Cooperative Study (ADCS), a consortium of academic medical centers and private Alzheimer disease clinics funded by the National Institute on Aging to conduct clinical trials on Alzheimer disease. In this trial, 238 subjects were randomized to the active treatment arm,

and 164 subjects were randomized to placebo. The trial measured multiple covariates at baseline including demographics and patient characteristics (e.g., sex, age, region, weight), lab tests (e.g., blood pressure, ApoE4 status [35(p4),36(p4)], and component scores of cognitive tests. More details are provided in Appendix 5.

Experiment 1. Pre-specified primary analysis of Phase 2 and 3 trials, to deliver higher power/confidence in the results compared to unadjusted analyses

After fitting the prognostic models, we analyzed the results from the Quinn et al. trial using three approaches: the unadjusted analysis; PROCOVA™ using the prognostic scores computed from the random forest; and PROCOVA™ using the prognostic scores computed from the deep learning model. This experiment used the same number of subjects and randomization ratio as the original study reported by Quinn et al. Data from subjects who dropped out of the study were not included in any of the analyses. We compared the resulting point estimates and 95% confidence intervals obtained with these three approaches for the effect of treatment on the changes in ADAS-Cog11 and CDR at 18 months.

Experiment 2. Prospective design/sample size estimation for Phase 2 and 3 trials, to attain the desired level of power/level of confidence with a smaller sample size compared to unadjusted trials.

We performed a sample size re-estimation and re-analysis of the Quinn et al. trial in order to demonstrate the clinical utility of accounting for prognostic covariate adjustment during trial design. When training the random forest and deep learning prognostic models, a subset of the ADNI and CPAD datasets were withheld for evaluating the variance and correlation required for the sample size calculation. Of the data that were not used in training the prognostic models, a subset of 345 subjects had (i) baseline Mini-Mental State Exam (MMSE) scores within the same range (14 to 26) as the inclusion criteria of the Quinn et al study, and (ii) had ADAS-Cog11 measurements through 18 months to enable calculation of the necessary standard deviation and correlation coefficients.

The sample size was calculated for a target treatment effect on ADAS-Cog11, though we also include analyses of CDR as a secondary endpoint. The parameters specified in PROCOVA™ Step 2 are given in Table 3.

Table 3. Parameters used in sample size re-estimation for the Quinn et al. study

Parameter	**Value**
Significance level (α)	5%
Desired power (ζ)	80%
Proportion of subjects randomized to treatment arm (π)	3/5
Target treatment effect (β_1^*)	3.1
Expected dropout (d)	0.3
Estimated standard deviation ($\hat{\sigma}_0$)	9.1
Inflation parameter for standard deviation in the control arm (γ_0)	1.0
Inflation parameter for standard deviation in the active treatment arm (γ_1)	1.0
Estimated prognostic correlation, random forest ($\hat{\rho}_0$)	0.36

Estimated prognostic correlation, deep learning model ($\hat{\rho}_0$)	0.43
Deflation parameter for prognostic correlation in the control arm (λ_0)	0.9
Deflation parameter for prognostic correlation in the active treatment arm (λ_1)	0.9

The sample size calculation was carried out using a binary search in a custom software library. We compared the original trial design and results to those obtained with PROCOVA™ based on the number of subjects as well as the resulting point estimates and 95% confidence intervals for the treatment effect on ADAS-Cog11 and CDR at 18 months. Additional details are provided in Appendix 7.

Of note, the only difference between *Experiment 1* and *Experiment 2* is the choice of the deflation parameters for prognostic correlation in the control and active treatment arms, λ_0 and λ_1, respectively. In *Experiment 1,* $\lambda_0 = \lambda_1 = 0$, which discounts the correlation to zero. That is, the estimated minimum sample size is the same as originally prespecified (before accounting for the prognostic score). *Experiment 2*, by contrast, uses $\lambda_0 = \lambda_1 = 0.9$, which assumes that the correlation of the prognostic model to observed outcomes in the study population will be slightly smaller than the one estimated from historical data.

3.3.2 Empirical Analyses Results

Experiment 1. Pre-specified primary analysis of Phase 2 and 3 trials, to deliver higher power/confidence in the results compared to unadjusted analyses.

Table 4 shows the results of three different approaches to estimating the treatment effect of DHA on the change in ADAS-Cog11 and CDR at 18 months: the unadjusted, difference-in-means analysis; PROCOVA™ while adjusting for prognostic score computed from the random forest; and PROCOVA™ while adjusting for prognostic score computed from the deep learning model. The data presented are point estimates and 95% confidence intervals for the estimated treatment effects.

Table 4. Reanalysis of the Quinn et al. trial at 18 months using two different prognostic scores

	Unadjusted analysis	Analysis adjusting for random forest prognostic score	Analysis adjusting for deep learning prognostic score
ADAS-Cog11	-0.10 ± 2.03	-0.11 ± 1.96	0.28 ± 1.88
CDR-SB	-0.02 ± 0.66	-0.02 ± 0.66	-0.11 ± 0.64

Concordant with the simulation studies, the standard errors for the effects obtained using prognostic covariate adjustment were smaller than or equal to those obtained using the unadjusted analysis. This led to narrower confidence intervals, which are still mathematically guaranteed to have the correct frequentist coverage.

While the point estimates for the treatment effects were modified to some extent when prognostic score adjustment was applied, the changes were minimal relative to the size of the estimated standard errors. Adjusting for baseline covariates or a prognostic score does not add bias [12,37,38], even though the point estimates for individual endpoints may change. That is, differences in point estimates between adjusted and unadjusted analyses are random, and do not persist in expectation. The original analysis of this particular trial[24] did not demonstrate

statistically significant improvements on any of the endpoints of interest, and nor did any of our re-analyses.

Experiment 2. Prospective design/sample size estimation for Phase 2 and 3 trials, to attain the desired level of power/level of confidence with a smaller sample size compared to unadjusted trials.

In designing a trial, one can set a desired statistical power for detecting a target treatment effect and then estimate the minimum number of subjects required to achieve that power. Using PROCOVA™ enables one to achieve a desired statistical power in a trial with fewer subjects. To demonstrate the efficiency gains associated with the use of PROCOVA™ during trial design, we performed a sample size re-estimation and re-analysis of the demonstration trial[24] introduced earlier.

Table 5 shows the minimum number of subjects required to achieve the desired power, estimated using an unadjusted analysis; using PROCOVA™ with a prognostic score computed from a random forest, and using PROCOVA™ with a prognostic score computed from a deep learning model. The Table also presents the point estimates and 95% confidence intervals for the estimated treatment effects on the two endpoints of interest.

Table 5. Re-analysis of the Quinn et al. study using different sample sizes that account for the impact of the prognostic score

	Unadjusted analysis	**Analysis using adjustment for random forest prognostic score**	**Analysis using adjustment for deep learning prognostic score**
Actively-treated Subjects	238	217	206
Placebo Subjects	164	144	137
Total Subjects	402	361	343
ADAS-Cog11	-0.10 ± 2.03	-0.14 ± 2.05	0.23 ± 2.04
CDR-SB	-0.02 ± 0.66	-0.02 ± 0.69	-0.11 ± 0.70

Using the random forest prognostic score resulted in a 10% reduction in the total number of required subjects compared to the unadjusted analysis, while using the deep learning prognostic score resulted in a 15% reduction in the total number of required subjects compared to the unadjusted analysis. Despite the reduced sample sizes, the widths of the confidence intervals for the effect on ADAS-Cog11 in the trial designs using PROCOVA™ are effectively the same.

Both hypothetical trial designs using PROCOVA™ have confidence intervals for the treatment effect on CDR that are 6% larger than in the unadjusted analysis. That is because the sample sizes were estimated from the performance of the respective prognostic models on ADAS-Cog11, with the goal of detecting a given effect on ADAS-Cog11. If one desires to achieve a given level of statistical power on multiple endpoints, then the sample size estimation procedure should be repeated for each of these endpoints and the largest sample size should be used. In addition, such applications will require either multiple prognostic models (i.e., one for each endpoint, as in our random forest example) or a multivariate prognostic model (i.e., one model that predicts all endpoints, as in our deep learning model).

4. Conclusions

In summary, our mathematical, simulation, and empirical results demonstrate that PROCOVA™ is a robust and efficient statistical methodology to leverage historical control arm data and predictive modeling (of any type). Its application significantly decreases the uncertainty in treatment effect estimates without compromising strict type-I error rate control in the large sample setting in Phase 2 and 3 trials. We have shown that our methodology increases the efficiency of both the design and analysis of RCTs measuring continuous responses in prospective applications.

Specifically, our mathematical results (Section 3.1.2) prove that PROCOVA™ improves over traditional ANCOVA methods that adjust for raw baseline covariates by constructing the optimal adjustment covariate – a prediction of a potential outcome under control conditions for all trial participants, conditioned on their observed baseline covariates. Specifically, Theorem 1 proves that estimates of treatment effects with PROCOVA™ are unbiased, and that Type-1 error rates of hypothesis tests are controlled at pre-specified levels, while Theorem 2 proves that such prediction of the potential outcome is the optimal covariate to adjust for in the analysis.

Our simulations (Section 3.2) show marked decreases in the mean-squared error of the estimated treatment effects associated with the use of PROCOVA™ alone or in combination with standard adjustment for baseline covariates, under four sets of conditions that model realistic situations encountered in clinical trials. Our results also indicate that prognostic covariate adjustment is a robust method that performs well even if the treatment effect is not constant, and when the prognostic model only approximates the expected control potential outcome of a subject conditioned on his/her baseline covariates.

And finally, our empirical results (Section 3.3) demonstrate that the prospective application of PROCOVA™ to Phase 2 and 3 RCTs (our stated context-of-use) significantly decreases variance in treatment effect estimates while maintaining type-I error rate control. In pre-specified primary analysis (*Experiment 1*), the use of PROCOVA™ delivers higher power and confidence in the results compared to unadjusted analyses; specifically, the width of the confidence intervals is decreased by up to 8%. In prospective design/sample size estimation (*Experiment 2*), its application attains desired level of power/level of confidence with a smaller sample size compared to unadjusted trials; specifically, the minimum total sample size is decreased by up to 15%. These benefits are realized using different types of prognostic models, illustrating that PROCOVA™ is a robust statistical methodology that can be applied with any prognostic model.

A number of recent technological developments, such as the development of large clinical databases, high dimensional biomarkers, and novel machine learning technologies, have led to substantial improvements in the ability to train highly accurate prognostic models. Using a simple rule of thumb, a prognostic model that obtains a correlation of R with observed outcomes can be used with PROCOVA™ to decrease the variance of the estimated treatment effect by a factor of $1 - R^2$, approximately. For example, using a prognostic score with $R = 0.5$ provides up to 25% decrease in variance, whereas using a prognostic score with $R = 0.8$ provides up to 64% decrease in variance. Due to the recent technological developments, it is now feasible to train prognostic models that obtain correlations of this magnitude for a variety of continuous responses in multiple therapeutic areas. Therefore, using PROCOVA™ to adjust for these more predictive prognostic scores can substantially reduce variance and widths of confidence intervals, and/or increase power and reduce minimum required sample sizes.

While the current application focuses on sample size and treatment effect estimation for RCTs with continuous variables under the requirement of strict type-I error rate control, ongoing and future work will develop PROCOVA™ applications to/in other areas including, but not limited to, RCTs with repeated measurements, binary or count outcomes, and time-to-event outcomes, as well Bayesian analogues that provide more statistical power while limiting the type-I error rate under reasonable conditions.

5. Questions for the EMA

Questions on Statistical Properties of PROCOVA™

Question 1. Does the EMA agree that PROCOVA™ produces unbiased treatment effect estimates and controls the type-I error rate, given that:

a. **PROCOVA™ is a special case of ANCOVA in which the covariate used for adjustment is a prognostic score, computed from data collected at or before baseline using a pre-specified prognostic model;**

b. **ANCOVA can decrease the variance of the estimated treatment effect if the adjustment covariate is correlated with the response;**

c. **Using ANCOVA to adjust for a covariate produces unbiased treatment effect estimates and controls the type-I error rate, as long as the covariate is computed from data collected at or before baseline.**

Supporting Evidence:

ANCOVA is known to possess several desirable statistical properties: with its use, estimated treatment effects will be unbiased, the type-I error rate will be controlled, and trial power will be increased if there is a correlation between the outcome and the adjustment covariate. Because of these statistical properties, ANCOVA is widely used in the analysis of clinical trials with continuous responses and is supported by guidance from EMA [13] and draft guidance from FDA [14].

Our mathematical results (Section 3.1.2) demonstrate that PROCOVA™ is a special case of ANCOVA with a particular choice of adjustment covariate. As such, PROCOVA™ inherits the statistical properties of ANCOVA described above, and these statistical properties hold for PROCOVA™ when used in conjunction with any prognostic model, regardless of the approach to modeling or the data used to inform the model.

Moreover, PROCOVA™ improves over traditional ANCOVA methods that adjust for raw baseline covariates by constructing the optimal adjustment covariate – a prediction of a potential outcome under control conditions for all trial participants, conditioned on their observed baseline covariates collected at or prior to the randomization. Theorem 1 proves that estimates of treatment effects with ANCOVA, and therefore PROCOVA™, are unbiased, and that type-1 error rates of hypothesis tests are controlled at pre-specified levels, while Theorem 2 proves that such prediction of the potential outcome is the optimal covariate to adjust for in the analysis. Detailed mathematical results are provided in Appendix 2 and Appendix 3.

The type-1-error rate control is further illustrated by the results of our simulations described in Section 3.2.2 and Appendix 4.

Questions 2. Does the EMA agree that PROCOVA™ can decrease the variance of the estimated treatment effect, and that it achieves lower variance when the prognostic score is more highly correlated with the response?

Supporting Evidence:

Theorem 2 proves that a prognostic score, i.e., the prediction of a potential outcome under control conditions for all trial participants conditioned on their observed baseline covariates, is the optimal covariate to adjust for in ANCOVA. Theorem 2 is presented

and further discussed in Section 3.1.2.2, Appendix 2 and Appendix 3.

Our simulation results described in Section 3.2 and specifically in Table 1 and Table 2, as well as in Appendix 4, demonstrate that the higher the correlation between the prognostic score and the observed control outcomes, the greater the reduction in the variance of treatment effect estimates. This finding held when PROCOVA™ was applied alone (Table 1) or combined with adjustment for baseline covariates (Table 2).

Additional evidence is provided by our empirical demonstration presented in Section 3.3, with further technical details included in Appendix 5, Appendix 6, and Appendix 7. Specifically, the results in Table 4 and Table 5 show that greater reductions in variance can be achieved when the prognostic score is more highly correlated with the observed outcome.

Question 3. Does the EMA agree that applying adjustment for the prognostic score during sample size estimation can result in a smaller minimum sample size required to achieve the desired level of power?

Supporting Evidence:

We describe the relationship between variance and power in our mathematical results (Section 3.1.2, Appendix 2 and Appendix 3), as well as in our simulations (Section 3.2 and Appendix 4). Our empirical application of PROCOVA™ (Section 3.3) shows that the use of PROCOVA™ allows to maintain power at lower sample sizes, as outlined in Section 3.3.2 and specifically in Table 5, as well as in Appendix 7.

Questions on the Context-of-Use

Question 4. Does the EMA agree that PROCOVA™ is an acceptable statistical method to estimate treatment effects in phase 2 and 3 clinical trials with continuous responses, given that:

a. **PROCOVA™ is a special case of ANCOVA;**
b. **ANCOVA is an acceptable statistical method to estimate treatment effects in phase 2 and 3 clinical trials with continuous responses under current regulatory guidance.**

Supporting evidence:

ANCOVA is known to possess several desirable statistical properties: with its use, estimated treatment effects will be unbiased, the type-I error rate will be controlled, and trial power will be increased if there is a correlation between the outcome and the adjustment covariate. Because of these statistical properties, ANCOVA is widely used in the analysis of clinical studies with continuous responses, including registration trials, and is supported by guidance from EMA [13] and draft guidance from FDA [14]. This information is summarized in Section 3.1.1 (in particular, Step 3), Appendix 2 and Appendix 3.

Our overview of PROCOVA™ (Section 3.1.1) and our mathematical results (Section 3.1.2) establish that PROCOVA™ is a special case of ANCOVA with a particular choice of adjustment covariate. As such, PROCOVA™ inherits the statistical properties of ANCOVA described above, and these statistical properties hold for PROCOVA™ when used in conjunction with any prognostic model, regardless of the approach to modeling or the data used to inform the model.

Therefore, PROCOVA™ is also acceptable and should be recommended for use to estimate treatment effects in pre-specified analyses of pivotal/registration trials.

Question 5. Does the EMA agree that it is acceptable to account for the adjustment of the prognostic score using PROCOVA™ during sample size estimation for a phase 2 and 3 clinical trials with continuous responses?

Supporting evidence:

We have provided three lines of evidence demonstrating that the use of PROCOVA™ can reduce variance of the treatment effect estimates: mathematical results (Section 3.1.2), simulations (Section 3.2 and specifically Table 1 and Table 2) and empirical examples (Section 3.3 – Experiment 2 and Table 4).

In addition, we have shown that the same power can be delivered with a smaller sample size and lower variance (reduced via application of PROCOVA™), as with a larger sample size and higher variance. This was established in our simulations described in Section 3.2 and in empirical demonstration presented in Section 3.3 (see Experiment 2) and Table 5.

The technical details for our mathematical results are provided in Appendix 2 and Appendix 3; for our simulations – in Appendix 4, for empirical demonstrations – in Appendix 5 and Appendix 6, and for sample size estimation – in Appendix 7.

Question 6. Does the EMA agree that PROCOVA™, combined with a predictive prognostic model and if implemented as described, could enable increases in power and/or decreases in minimum sample sizes in phase 2 or 3 clinical trials with continuous responses?

Supporting evidence:

Our approach is designed to prospectively decrease the uncertainty, or variance, in treatment effect estimates from RCTs without compromising strict type-1 error rate control in the large-sample setting. We achieve this by combining curated historical control arm data, highly predictive modeling, and covariate adjustment for the prognostic score generated through modeling.

Our mathematical results (Section 3.1.2, Appendix 2, and Appendix 3), simulations (Section 3.2 and specifically Table 1 and Table 2, as well as Appendix 4) and empirical examples (Section 3.3, Appendix 5, Appendix 6, and Appendix 7) demonstrate that PROCOVA™ can reduce variance of the treatment effect estimates in trials with continuous responses.

This reduction in variance can be leveraged either by increasing analytical power without increasing the sample size (Section 3.3, Experiment 1), or by reducing the minimum required sample size while maintaining the power (Section 3.3, Experiment 2). The Sponsor can make that choice depending on the circumstances of a particular trial but must prospectively pre-specify the application of PROCOVA™ prior to unblinding, to avoid bias.

In summary, our method is scientifically sound since it only adjusts for a single covariate (or single additional covariate) derived from information collected at baseline/prior to randomization; produces unbiased estimates for treatment effects;

controls the type-I error rate; and leads to correct confidence interval coverage. It is also consistent with current FDA and EMA regulatory guidance. As such, PROCOVA™ can be used to prospectively increase the power or reduce the minimum required sample size in studies that support drug approvals, i.e., pivotal/confirmatory Phase 3, and occasionally Phase 2, clinical trials.

6. References and Appendices

6.1 References

1. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis ICh, Shi W. Federated learning of predictive models from federated Electronic Health Records. *Int J Med Inf*. 2018;112:59-67. doi:10.1016/j.ijmedinf.2018.01.007

2. Dankar FK, El Emam K. The application of differential privacy to health data. In: *Proceedings of the 2012 Joint EDBT/ICDT Workshops*. EDBT-ICDT '12. Association for Computing Machinery; 2012:158-166. doi:10.1145/2320765.2320816

3. Baker SG, Lindeman KS. Rethinking historical controls. *Biostatistics*. 2001;2(4):383-396. doi:10.1093/biostatistics/2.4.383

4. Ibrahim JG, Chen M-H, Gwon Y, Chen F. The Power Prior: Theory and Applications. *Stat Med*. 2015;34(28):3724-3749. doi:10.1002/sim.6728

5. Kopp-Schneider A, Calderazzo S, Wiesenfarth M. Power gains by using external information in clinical trials are typically not possible when requiring strict type I error control. *Biom J*. 2020;62(2):361-374. doi:https://doi.org/10.1002/bimj.201800395

6. Lim J, Walley R, Yuan J, et al. Minimizing Patient Burden Through the Use of Historical Subject-Level Data in Innovative Confirmatory Clinical Trials: Review of Methods and Opportunities. *Ther Innov Regul Sci*. 2018;52(5):546-559. doi:10.1177/2168479018778282

7. Ghadessi M, Tang R, Zhou J, et al. A roadmap to using historical controls in clinical trials – by Drug Information Association Adaptive Design Scientific Working Group (DIA-ADSWG). *Orphanet J Rare Dis*. 2020;15(1):69. doi:10.1186/s13023-020-1332-x

8. Anonymous. ICH E9 statistical principles for clinical trials. European Medicines Agency. Published September 17, 2018. Accessed March 26, 2021. https://www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials

9. Kahan BC, Jairath V, Doré CJ, Morris TP. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials*. 2014;15(1):139. doi:10.1186/1745-6215-15-139

10. Lin W. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Ann Appl Stat*. 2013;7(1):295-318. doi:10.1214/12-AOAS583

11. Raab GM, Day S, Sales J. How to Select Covariates to Include in the Analysis of a Clinical Trial. *Control Clin Trials*. 2000;21(4):330-342. doi:10.1016/S0197-2456(00)00061-1

12. Yang L, Tsiatis AA. Efficiency Study of Estimators for a Treatment Effect in a Pretest-Posttest Trial. *Am Stat*. 2001;55(4):314-321.

13. Guideline on adjustment for baseline covariates in clinical trials, 26 February 2015, EMA/CHMP/295050/2013; Committee for Medicinal Products for Human Use (CHMP).

14. Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products, Guidance for Industry, US FDA/CDER/CBER, May 2021.

15. Ambrosius WT, Sink KM, Foy CG, et al. The design and rationale of a multicenter clinical trial comparing two strategies for control of systolic blood pressure: The Systolic Blood Pressure Intervention Trial (SPRINT). *Clin Trials J Soc Clin Trials*. 2014;11(5):532-546. doi:10.1177/1740774514537404

16. Austin SR, Wong Y-N, Uzzo RG, Beck JR, Egleston BL. Why summary comorbidity measures such as the Charlson Comorbidity Index and Elixhauser score work. *Med Care*. 2015;53(9):e65-e72. doi:10.1097/MLR.0b013e318297429c

17. Cooney MT, Dudina AL, Graham IM. Value and Limitations of Existing Scores for the Assessment of Cardiovascular Risk. *J Am Coll Cardiol*. 2009;54(14):1209-1227. doi:10.1016/j.jacc.2009.07.020

18. Fisher CK, Smith AM, Walsh JR. Machine learning for comprehensive forecasting of Alzheimer's Disease progression. *Sci Rep*. 2019;9(1):13622. doi:10.1038/s41598-019-49656-2

19. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539

20. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform*. 2018;19(6):1236-1246. doi:10.1093/bib/bbx044

21. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *Npj Digit Med*. 2018;1(1):1-10. doi:10.1038/s41746-018-0029-1

22. Dubois S, Romano N, Jung K, Shah N, Kale DC. THE EFFECTIVENESS OF TRANSFER LEARNING IN ELECTRONIC HEALTH RECORDS DATA. Published online 2017:4.

23. General Data Protection Regulation (GDPR) – Official Legal Text. General Data Protection Regulation (GDPR). Accessed March 26, 2021. https://gdpr-info.eu/

24. Quinn JF, Raman R, Thomas RG, Yurko-mauro K, Nelson PEB, Md A. *Trial Registration Clinicaltrials.Gov Identifier: NCT00440050*.

25. Schuler A, Walsh D, Hall D, Walsh J, Fisher C. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. *ArXiv201209935 Cs Stat*. Published online December 17, 2020. Accessed March 6, 2021. http://arxiv.org/abs/2012.09935

26. Walsh D, Schuler A, Hall D, Walsh J, Fisher C. Bayesian prognostic covariate adjustment. *ArXiv201213112 Stat*. Published online December 24, 2020. Accessed March 6, 2021. http://arxiv.org/abs/2012.13112

27. Anonymous. Missing data in confirmatory clinical trials. European Medicines Agency. Published September 17, 2018. Accessed March 9, 2021. https://www.ema.europa.eu/en/missing-data-confirmatory-clinical-trials

28. Petersen RC, Aisen PS, Beckett LA, et al. Alzheimer's Disease Neuroimaging Initiative (ADNI). *Neurology*. 2010;74(3):201-209. doi:10.1212/WNL.0b013e3181cb3e25

29. Neville J, Kopko S, Broadbent S, et al. Development of a unified clinical trial database for Alzheimer's disease. *Alzheimers Dement*. 2015;11(10):1212-1221. doi:10.1016/j.jalz.2014.11.005

30. Romero K, de Mars M, Frank D, et al. The Coalition Against Major Diseases: Developing Tools for an Integrated Drug Development Process for Alzheimer's and Parkinson's Diseases. *Clin Pharmacol Ther*. 2009;86(4):365-367. doi:10.1038/clpt.2009.165

31. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

32. Bertolini D, Loukianov AD, Smith AM, et al. Modeling Disease Progression in Mild Cognitive Impairment and Alzheimer's Disease with Digital Twins. *ArXiv201213455 Cs Q-Bio*. Published online December 24, 2020. Accessed March 6, 2021. http://arxiv.org/abs/2012.13455

33. Rosen WG, Mohs RC, Davis KL.A new rating scale for Alzheimer's disease. *Am J Psychiatry*. 1984;141(11):1356-1364. doi:10.1176/ajp.141.11.1356

34. Morris JC. The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology*. 1993;43(11):2412-2412. doi:10.1212/WNL.43.11.2412-a

35. Coon KD, Myers AJ, Craig DW, et al. A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *J Clin Psychiatry*. 2007;68(4):613-618. doi:10.4088/jcp.v68n0419

36. Safieh M, Korczyn AD, Michaelson DM. ApoE4: an emerging therapeutic target for Alzheimer's disease. *BMC Med*. 2019;17(1):64. doi:10.1186/s12916-019-1299-4

37. Leon S, Tsiatis AA, Davidian M. Semiparametric Estimation of Treatment Effect in a Pretest-Posttest Study. *Biometrics*. 2003;59(4):1046-1055. doi:10.1111/j.0006-341X.2003.00120.x

38. Wang B, Ogburn EL, Rosenblum M. Analysis of covariance in randomized trials: More precision and valid confidence intervals, without model assumptions. *Biometrics*. 2019;75(4):1391-1400. doi:10.1111/biom.13062

39. Rubin DB. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. :10.

40. Hausman J, Palmer C. Heteroskedasticity-robust inference in finite samples. *Econ Lett*. 2012;116(2):232-235. doi:10.1016/j.econlet.2012.02.007

41. Robins JM, Rotnitzky A, Zhao LP. Estimation of Regression Coefficients When Some Regressors are not Always Observed. *J Am Stat Assoc*. 1994;89(427):846-866. doi:10.1080/01621459.1994.10476818

42. Rosenblum M, van der Laan MJ. Simple, Efficient Estimators of Treatment Effects in Randomized Trials Using Generalized Linear Models to Leverage Baseline Variables. *Int J Biostat*. 2010;6(1). doi:10.2202/1557-4679.1138

43. Tsiatis A. *Semiparametric Theory and Missing Data*. Springer-Verlag; 2006. doi:10.1007/0-387-37345-4

44. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *Mach Learn PYTHON*.:6.

45. Wickham H. Tidy Data. *J Stat Softw*. 2014;059(i10). Accessed March 9, 2021. https://ideas.repec.org/a/jss/jstsof/v059i10.html

46. Ackley DH, Hinton GE, Sejnowski TJ. A Learning Algorithm for Boltzmann Machines*. *Cogn Sci*. 1985;9(1):147-169. doi:https://doi.org/10.1207/s15516709cog0901_7

47. Generating Digital Twins with Multiple Sclerosis Using Probabilistic Neural Networks | bioRxiv. Accessed March 6, 2021. https://www.biorxiv.org/content/10.1101/2020.02.04.934679v2

48. Le Roux N, Bengio Y. Representational Power of Restricted Boltzmann Machines and Deep Belief Networks. *Neural Comput*. 2008;20(6):1631-1649. doi:10.1162/neco.2008.04-07-510

49. Mnih V, Larochelle H, Hinton GE. Conditional Restricted Boltzmann Machines for Structured Output Prediction. :9.

50. Taylor GW, Hinton GE. Factored conditional restricted Boltzmann Machines for modeling motion style. In: *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*. ACM Press; 2009:1-8. doi:10.1145/1553374.1553505

51. Tieleman T. Training restricted Boltzmann machines using approximations to the likelihood gradient. In: *Proceedings of the 25th International Conference on Machine Learning - ICML '08*. ACM Press; 2008:1064-1071. doi:10.1145/1390156.1390290

6.2 Appendices

Appendix 1. Detailed Description of PROCOVA™

This section elaborates on the description of PROCOVA™. Specifically, PROCOVA™ consists of the following three steps:

Step 1: Training and evaluating a prognostic model to predict control outcomes.

Let $E[Y_0|X]$ denote the expected outcome for a subject with baseline covariates X if that subject were to receive the control treatment in the planned trial (e.g., placebo). We define this as the prognostic score for this subject.

We define a prognostic model m as a function of the baseline covariates that approximates the prognostic score, i.e., $m(X) \approx E[Y_0|X]$. The prognostic model can be any non-constant function of the baseline covariates; the type-I error rate is controlled for any type of model (e.g., a mechanistic model, machine learning-model, or deep neural network) whereas the realized power only depends on the predictive performance of the model.

To obtain a prognostic model, statistical or machine learning-based methods can be used to fit the model to a collection of historical data, $D_H = \{(X', Y')\}$, linking subjects' baseline covariates to their outcomes under the control condition. We will provide two examples of this type of prognostic model in our empirical analyses, but we reiterate that PROCOVA™ is independent from the method used to produce the prognostic model.

In order to use the prognostic model to estimate the minimum sample size in Step 2 below, we need to estimate its performance in the patient population that will be enrolled in the planned study (target population). To do so, we collect a set of historical data $D_{H''} = \{(X'', Y'')\}$ linking subjects' baseline covariates to their outcomes under the control condition. We recommend choosing this dataset such that it is not part of the dataset used to fit the prognostic model. Subjects from $D_{H''}$ who have similar baseline characteristics to those in the target population (e.g., those who meet the inclusion criteria of the planned trial) are selected and the variance $\hat{\sigma}_0^2 = \text{Var}[Y'']$ and correlation $\hat{\rho}_0 = \text{Cov}[Y'', m(X)]/\sqrt{\text{Var}[Y'']\text{Var}[m(X)]}$ are computed.

Step 2: Accounting for the prognostic model while estimating the sample size required for a prospective study.

Define the target effect size, β_1^*, the significance threshold α, the desired power level, ζ, fraction of subjects to be randomized to the active arm, π, and dropout rate, d. In addition, define $\gamma_w \geq 1$ and $\lambda_w \in [0,1]$ for $w = 0,1$. Here, the variance of the potential outcome under active treatment w in the planned trial will be assumed to be $\gamma_w^2\hat{\sigma}_0^2$, such that choosing large γ_w inflates the estimated variance. Similarly, the correlation between the potential outcome and the prognostic model under active treatment w will be assumed to be $\lambda_w\hat{\rho}_0$, such that choosing small λ_w deflates the estimated correlation.

With the above parameters now defined, we use a numerical optimization algorithm (such as a binary search to minimize n such that

$$\zeta \geq \Phi\left(\Phi^{-1}\left(\frac{\alpha}{2}\right)+\frac{\beta_1^*}{\nu}\right)+\Phi\left(\Phi^{-1}\left(\frac{\alpha}{2}\right)-\frac{\beta_1^*}{\nu}\right)$$

with $\nu^2 = \frac{1}{n}\left(\frac{\gamma_0^2\hat{\sigma}_0^2}{1-\pi}+\frac{\gamma_1^2\hat{\sigma}_0^2}{\pi}+\frac{\hat{\theta}^2-2\hat{\theta}_*\hat{\theta}}{\pi(1-\pi)}\right)$, $\hat{\theta} = \hat{\rho}_0\hat{\sigma}_0\big((1-\pi)\lambda_0\gamma_0+\pi\lambda_1\gamma_1\big)$, and $\hat{\theta}_* = \hat{\rho}_0\hat{\sigma}_0(\pi\lambda_0\gamma_0+(1-\pi)\lambda_1\gamma_1)$. The minimum sample size is estimated to be $n_d = \frac{n}{1-d}$.

If there are multiple outcomes of interest, each with a desired power level and target effect size, then this procedure must be repeated for each outcome, and the largest sample size should be selected. Note that this may require the use of multiple prognostic models (i.e., one to predict each outcome of interest) or a multivariate prognostic model.

Step 3: Estimating the treatment effect from the completed study using a linear model while adjusting for the control outcomes predicted by the prognostic model.

An RCT with n_d subjects is performed in which each subject is randomized to active treatment ($W_i = 1$) or control ($W_i = 0$). The result is a dataset (X_i, W_i, Y_i) for $i = 1, \ldots, n_d$. Data from subjects who have dropped out of the study should be handled with an appropriate, pre-specified, method as in any trial analysis [27]. Next, the treatment effect is estimated by fitting a linear model,

$$Y_i = \beta_0 + \beta_1 W_i + \beta_2 m(X_i) + \epsilon_i$$

while adjusting for the estimated prognostic score. One could also adjust for additional covariates in the regression if desired, so long as the sample size, n, is much greater than the total number of terms in the linear model, $k + 2$.

Finally, the null hypothesis $H0: \beta_1 = \beta_1^*$ can be assessed by computing a two-sided p-value,

$$p = 2\left(1-\Phi\left(\frac{|\hat{\beta}_1-\beta_1^*|}{\sigma_{\hat{\beta}_1}}\right)\right)$$

in which $\sigma_{\hat{\beta}_1}$ is the standard error on the estimated treatment effect $\hat{\beta}_1$, and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. The null hypothesis is rejected with a two-sided significance test at level α if $p < \alpha$.

The PROCOVA™ method described above is a special case of Analysis of Covariance (ANCOVA) with a particular choice of adjustment covariate. As such, PROCOVA™ inherits the statistical properties of ANCOVA; for example, estimated treatment effects will be unbiased and the type-I error rate will be controlled. For these reasons, ANCOVA is widely used in the analysis of clinical trials with continuous responses and is supported by guidance from EMA [13] and draft guidance from FDA [14]. These properties hold for PROCOVA™ using any prognostic model, regardless of the approach to modeling or the data used to inform the model.

PROCOVA™ improves over traditional ANCOVA methods that adjust for raw baseline covariates by constructing the optimal adjustment covariate, $m(X) \approx E[Y_0|X]$. Under certain

conditions outlined below, we show that adjusting for this covariate in a linear model to estimate the treatment effect achieves the minimum variance among all asymptotically unbiased estimators with access to the same baseline covariates. Thus, for a given sample size, PROCOVA™ can lead to substantial increases in power without sacrificing control of the type I error rate. Moreover, only a small number of parameters need to be estimated from historical data when accounting for the effect of the prognostic score on the analysis while estimating the required minimum sample size with PROCOVA™.

Appendix 2. General Mathematical Results

A.2.1 Notation

We consider an RCT with n subjects who are randomized to the active treatment arm (denoted $W_i = 1$) or to the control arm (denoted $W_i = 0$). There is a vector X_i of covariates measured for each subject i at baseline (i.e., before the first treatment is given), and a continuous variable $Y_i \in R$ measured at the end of the study. Thus, the trial dataset is a set of n tuples (X_i, W_i, Y_i), which we denote $(\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{Y})$.

Let $Y_{0,i}$ and $Y_{1,i}$ be the potential control and active treatment outcomes for subject i, respectively, and let $Y_W = WY_1 + (1 - W)Y_0$ denote the observed outcomes, which are a function of the random treatment assignment [39]. Our structural assumption about the trial is,

$$P(\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{Y}, \boldsymbol{Y}_0, \boldsymbol{Y}_1) = \mathbf{1}(\boldsymbol{Y} = \boldsymbol{Y}_{\boldsymbol{W}}) P(\boldsymbol{W}) \prod_{\mathrm{i}} \mathrm{P}\left(\mathrm{X_i}, \mathrm{Y_{0,i}}, \mathrm{Y_{1,i}}\right) \qquad \textbf{Eq. 1}$$

In other words, a) the observed outcomes are the potential outcomes corresponding to the assigned treatment, b) the treatment is assigned randomly and independently of all other variables, and c) the trial subjects are independent and identically distributed. Denote the conditional average outcomes under each treatment condition as $\mu_w(X) = \mathrm{E}[Y_w|X]$, and the population average outcomes under each treatment condition w as $\mu_w = \mathrm{E}[Y_w]$, such that the average treatment effect is the difference in means, $\tau \coloneqq \mu_1 - \mu_0$.

We also presume access to a historical dataset (X', Y') containing n' identically and independently distributed observations of baseline covariates, X', and outcomes, Y'. Note that we assume that these are the same baseline covariates to be measured in the RCT, but in a potentially different population. These have some joint distribution $P_H(X', Y')$ in the historical population. We will see that our best-case scenario is $P_H(X', Y') = P(X, Y_0)$ so that the historical population is maximally informative about the trial control arm.

In addition, let $\pi = P(W_i = 1)$ and $1 - \pi = P(W_i = 0)$ be the probability that a subject is assigned to the active treatment or control arm in the trial, respectively. Finally, let $\mathrm{Var}[A]$ denote the variance of A and $\mathrm{Cov}[A, B]$ denote the covariance between A and B.

In what follows, we abbreviate the usual empirical (sample) average of identically and independently distributed variables $A_1, \ldots, A_n \sim A$ with the notation $\bar{A} \coloneqq \widehat{\mathrm{E}}[A] = \frac{1}{n}\sum_i A_i$. In addition, we assume that the covariates we are adjusting for in the ANCOVA estimator (defined below) are computed as function of the raw covariates, i.e., $F = f(X)$. The function f is essentially arbitrary except for the minor restriction that the covariance matrix of the transformed covariates, $\mathrm{Cov}[F]$, is invertible.

A.2.2 Analysis of Covariance (ANCOVA)

Analysis of Covariance (ANCOVA) can be used to estimate a treatment effect from an RCT by fitting the linear model,

$$Y_i = \beta_0 + \beta_1 W_i + \beta_2^T f(X_i) + \epsilon_i \qquad \text{Eq. 2}$$

using the method of least squares. Here, $f(\mathrm{X})$ is an arbitrary, potentially multidimensional, function of the baseline covariates, that defines a set of adjustment covariates in the regression. Commonly, $f(X) = [X_1, \ldots, X_k]^T$ is simply the identify function, but it is also common to include squared terms (e.g., X_1^2) or interactions (e.g., $X_1 X_2$) chosen *a priori*. However, the statistical framework presented below holds for all these, as well more general types of functions.

If we define the matrix of regressors as $Z = [1, W, f(X)^T]^T$, then we can write the point estimate for $\beta = [\beta_0, \beta_1, \beta_2^T]^T$ in matrix notation as,

$$\hat{\beta} = (Z^T Z)^{-1} Z^T Y \qquad \text{Eq. 3}$$

The standard errors in the regression coefficients, which we denote as $\Sigma_{\hat{\beta}}$, could be estimated with the standard maximum likelihood estimator under the assumption of normally distributed homoscedastic errors, or with an estimator that is robust to heteroscedasticity such as:

$$\Sigma_{\hat{\beta}} = \frac{n}{n-k-2} (ZZ^T)^{-1} Z \mathrm{diag}(e_i^2) Z^T (ZZ^T)^{-1} \qquad \textbf{Eq. 4}$$

in which $e_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 W_i + \hat{\beta}_2^T f(X_i)\right)$ are the residual errors, and k is the dimension of $f(X)$. Here, we use the heteroscedasticity robust standard errors. A discussion of alternative estimators for the standard errors can be found in [40].

The null hypothesis $H0: \beta_1 = \beta_1^*$ can be assessed by computing a p-value,

$$p = 2\left(1 - \Phi\left(\frac{|\hat{\beta}_1 - \beta_1^*|}{\sqrt{\Sigma_{\hat{\beta}_1}}}\right)\right) \qquad \text{Eq. 5}$$

in which $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. Here, and throughout, we rely on asymptotic theory taken in the limit that $n \to \infty$ while the number of terms in the linear model (i.e., $k + 2$) is fixed. The null hypothesis is rejected with a two-sided significance test at level α if $p < \alpha$. Similarly, a confidence interval for the treatment effect can be constructed as $\hat{\beta}_1 \pm \Phi^{-1}(\alpha/2)\sqrt{\Sigma_{\hat{\beta}_1}}$.

A.2.3 Mathematical Properties of ANCOVA

The following mathematical theorems establish statistical properties of ANCOVA estimators. As PROCOVA™ is a special case of ANCOVA, each of these properties also applies to PROCOVA™ estimators. All proofs are provided in Appendix 3.

Theorem 1: The ANCOVA estimator $\hat{\beta}_1$ is asymptotically distributed as $\hat{\beta}_1 \sim \mathcal{N}(\tau, \nu^2)$ with mean $\tau = \mu_1 - \mu_0$ and variance,

$$\nu^2 \coloneqq \mathrm{Var}[\hat{\beta}_1] = \frac{1}{n}\left(\frac{\sigma_0^2}{1-\pi} + \frac{\sigma_1^2}{\pi} + \frac{(\xi^T - 2\xi_*^T)\mathrm{V}^{-1}\xi}{\pi(1-\pi)}\right)$$

in which $\xi = (1-\pi)\text{Cov}[Y_0, f(X)] + \pi\text{Cov}[Y_1, f(X)]$, $\xi_* = \pi\text{Cov}[Y_0, f(X)] + (1-\pi)\text{Cov}[Y_1, f(X)]$, and $V = \text{Cov}[f(X)]$ is the covariance matrix of the adjustment covariates. Note that V must be invertible, which places mild restrictions on $f(X)$.

Corollary 1.1 The two-sided significance test for the null-hypothesis, $H_0: \tau = \tau_*$, conducted by computing the p-value, $p = 2\left(1 - \Phi\left(\frac{|\hat{\beta}_1 - \tau_*|}{\nu}\right)\right)$, and rejecting the null-hypothesis if $p < \alpha$ has asymptotic type-I error rate α.

Corollary 1.2: The power of this two-sided significance test to detect a treatment effect $\Delta\tau = \tau - \tau_*$ is asymptotically,

$$\text{Power}(\Delta\tau) = \Phi\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) + \frac{\Delta\tau}{\nu}\right) + \Phi\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) - \frac{\Delta\tau}{\nu}\right). \quad \text{Eq. 6}$$

Corollary 1.3: If $F = f(X)$ is one dimensional, then the asymptotic variance simplifies to,

$$\nu^2 = \frac{1}{n}\left(\frac{\sigma_0^2}{1-\pi} + \frac{\sigma_1^2}{\pi} + \frac{\theta^2 - 2\theta_*\theta}{\pi(1-\pi)}\right) \quad \text{Eq. 7}$$

in which write $\theta = (1-\pi)\rho_0\sigma_0 + \pi\rho_1\sigma_1$, $\theta_* = \pi\rho_0\sigma_0 + (1-\pi)\rho_1\sigma_1$, with $\rho_0 = \text{Cov}[Y_0, f(X)]/\sqrt{\text{Var}[Y_0]\text{Var}[f(X)]}$ and $\rho_1 = \text{Cov}[Y_1, f(X)]/\sqrt{\text{Var}[Y_1]\text{Var}[f(X)]}$.

This corollary demonstrates that the formula for the asymptotic variance of the estimated treatment effect simplifies if the baseline covariates are transformed into a one-dimensional variable by application of the function. This is useful for prospective power calculations, because it substantially reduces the number of parameters that need to be estimated in order to estimate the minimum sample size required in a future study.

Corollary 1.4: If $F = f(X)$ is one dimensional, and equal numbers of subjects are randomized to active treatment and control, then the asymptotic variance further simplifies to,

$$\nu^2 = \frac{2}{n}\left(\sigma_0^2 + \sigma_1^2 - \frac{(\rho_0\sigma_0 + \rho_1\sigma_1)^2}{2}\right)$$

Moreover, this variance is always less than that of the unadjusted estimator.

A.2.4 Mathematical Properties of PROCOVA™

We propose Prognostic Covariate Adjustment (PROCOVA™) as a procedure for estimating a treatment effect from an RCT measuring a continuous response. The method is motivated by the following theorem.

Theorem 2: Let $\mathcal{F}$ be the set of all functions $f: X \to \mathcal{R}$ that map the baseline covariates to a real number. If the treatment effect is constant and the variances are equal, that is $\mu_1(X) = \tau + \mu_0(X)$ and $\sigma_0 = \sigma_1$, then the asymptotic variance of the treatment effect estimated by ANCOVA is minimized by the choice $f(X) = \text{E}[Y_0|X]$.

Appendix 3. Proofs of Mathematical Results

A.3.1 Definitions

Difference-in-Means Estimator

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

ANCOVA Estimator

The "ANCOVA" estimator of $\tau = \mu_1 - \mu_0$ (denoted $\hat{\beta}_1$) is the effect estimated using a linear regression with predictors $Z = [1, W, F]^T$ and outcome Y.

A.3.2 Mathematical Theorems

Lemma 1:

The influence function for the linear regression treatment effect estimator is $\psi = \psi_1 - \psi_0$ in which

$$\psi_w = \frac{W_w}{\pi_w}\left(Y - \hat{\mu}_w^*(F)\right) + \left(\hat{\mu}_w^*(F) - \hat{\mu}_w^*\right)$$

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

Proof: This follows from results in [41]. An accessible presentation for the case of generalized linear models is given in [42]. ■

Lemma 2:

The difference-in-means estimator has asymptotic variance given by $\mathrm{Var}[\hat{\tau}_\Delta] = \frac{1}{n}\left(\frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1}\right)$ in which $\sigma_w^2 = \mathrm{Var}[Y_w]$, $\pi_0 = 1 - \pi$, and $\pi_1 = \pi$.

Proof: Follow the outline of Theorem 1 below taking $Z^T = [1, W]$. ■

Theorem 1: The ANCOVA estimator $\hat{\beta}_1$ is asymptotically distributed as $\hat{\beta}_1 \sim \mathcal{N}(\tau, v^2)$ with mean $\tau = \mu_1 - \mu_0$ and variance,

$$v^2 \coloneqq \mathrm{Var}\left[\hat{\beta}_1\right] = \frac{1}{n}\left(\frac{\sigma_0^2}{1-\pi} + \frac{\sigma_1^2}{\pi} + \frac{(\xi^T - 2\xi_*^T)\mathrm{V}^{-1}\xi}{\pi(1-\pi)}\right)$$

in which $\xi = (1-\pi)\mathrm{Cov}[Y_0, F] + \pi\mathrm{Cov}[Y_1, F]$, $\xi_* = \pi\mathrm{Cov}[Y_0, F] + (1-\pi)\mathrm{Cov}[Y_1, F]$, and $V = \mathrm{Cov}[F]$ is the covariance matrix of the adjustment covariates. Note that V must be invertible, which places mild restrictions on $f(X)$.

Proof: We begin by applying Lemma 1. Minimization of the expected log-likelihood shows that $\hat{\beta}^* = \mathrm{E}[ZZ^T]^{-1}\mathrm{E}[ZY]$.

The identity $\mathrm{E}[AB] = \mathrm{Cov}[A, B] + \mathrm{E}[A]\mathrm{E}[B]$, and the fact that our structural assumptions imply $W_w \perp Y_w, X$ and $W_w Y = W_w Y_w$, may be used to show that

$$\mathrm{E}[ZZ^T]^{-1} = \begin{bmatrix} \pi_0^{-1} + \eta^T V^{-1}\eta & -\pi_0^{-1} & -V^{-1}\eta \\ -\pi_0^{-1} & \pi_0^{-1}\pi_1^{-1} & 0 \\ -V^{-1}\eta & 0 & V^{-1} \end{bmatrix}$$

and

$$\mathrm{E}[ZY] = \begin{bmatrix} \mu \\ \pi_1\mu_1 \\ \mu\eta + \xi \end{bmatrix}$$

in which $\eta = \mathrm{E}[F]$, $\mu = \mathrm{E}[Y]$, $\xi = \mathrm{Cov}[F, Y]$, and $V = \mathrm{Cov}[F]$.

Applying the above demonstrates that $\hat{\beta}^* = [\mu_0, \tau, V^{-1}\xi^T]^T$. Thus, $\hat{\mu}_w^*(F) = \mu_0 + w\tau + \tilde{F}^T V^{-1}$ and , $\hat{\mu}_w^* = \mu_w$. In this equation and from here on, let $\tilde{F} = F - \mathrm{E}[F]$. Then, from Lemma 1,

$$\psi_w = \frac{W_w}{\pi_w}(Y - \mu_w) - \frac{\widetilde{W}_w}{\pi_w}\tilde{F}^T V^{-1}\xi$$

in which $\widetilde{W_w} = W_w - \pi_w$. It will be helpful to define $h_w(F) \coloneqq -\tilde{F}^T V^{-1}\xi$. An application of Lemma 1 and some algebra gives

$$\psi = \frac{W_1}{\pi_1}(Y - \mu_1) - \frac{W_0}{\pi_0}(Y - \mu_0) - (W_1 - \pi)\frac{\tilde{F}^T V^{-1}\xi}{\pi_0\pi_1}$$

It's helpful to define $\psi_{1,\Delta} \coloneqq \frac{W_1}{\pi_1}(Y - \mu_1)$, $\psi_{0,\Delta} \coloneqq \frac{W_0}{\pi_0}(Y - \mu_0)$, $\psi_\Delta = \psi_{1,\Delta} - \psi_{0,\Delta}$, $h(F) \coloneqq -\frac{\tilde{F}^T V^{-1}\xi}{\pi_0\pi_1}$, and $\phi = (W_1 - \pi)\frac{\tilde{F}^T V^{-1}\xi}{\pi_0\pi_1}$. Using this notation, we can write $\psi = \psi_\Delta - \phi$.

It is known that all regular and asymptotically linear estimators of the treatment effect have an influence function of this form with $h(F)$ dependent on the choice of estimator [37,43].

By the theory of influence functions, our estimator has a limiting distribution [43]

$$\sqrt{n}\left(\hat{\beta}_1 - \tau\right) \to \mathcal{N}(0, \mathrm{E}[\psi^2])$$

The asymptotic variance of $\hat{\beta}_1$ is thus $\mathrm{E}[\psi^2] = \mathrm{E}[(\psi_\Delta - \phi)^2] = \mathrm{E}[\psi_\Delta^2] - 2\mathrm{E}[\psi_\Delta\phi] + \mathrm{E}[\phi^2]$. The first term is the variance of the influence function for the difference-in-means (also called "unadjusted") estimator. It may be verified that this evaluates to $\mathrm{E}[\psi_\Delta^2] = \frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1}$ in which $\sigma_w^2 = \mathrm{Var}[Y_w]$. The variance of ϕ is

$$\mathrm{E}[\phi^2] = \mathrm{E}\left[\left(\frac{W_1 - \pi_1}{\pi_1\pi_0}\tilde{F}^T V^{-1}\xi\right)^2\right] = \frac{\mathrm{E}[(W_1 - \pi_1)^2]}{\pi_1^2\pi_0^2}\xi^T V^{-1}\mathrm{E}[\tilde{F}\tilde{F}^T]V^{-1} = \frac{1}{\pi_1\pi_0}\xi^T V^{-1}\xi$$

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

$$\mathrm{E}[\psi_\Delta\phi] = \mathrm{E}\left[\psi_{1,\Delta}\phi\right] - \mathrm{E}\left[\psi_{0,\Delta}\phi\right] = \left(\frac{1}{\pi_1}\xi_1^T + \frac{1}{\pi_0}\xi_0^T\right)V^{-1}\xi = \frac{1}{\pi_1\pi_0}\xi_*^T V^{-1}\xi$$

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

Corollary 1.1 The two-sided significance test for the null-hypothesis, $H_0: \tau = \tau_*$, conducted by computing the p-value, $p = 2\left(1 - \Phi\left(\frac{|\widehat{\beta}_1 - \tau_*|}{\nu}\right)\right)$, and rejecting the null-hypothesis if $p < \alpha$ has asymptotic type-I error rate α.

Proof: Theorem 1 implies that the test statistic $T = \frac{\widehat{\beta}_1 - \tau_*}{\nu} \sim \mathcal{N}(0,1)$ follows a standard normal distribution under the null hypothesis in the asymptotic regime. Therefore, the null-hypothesis will be rejected with probability $Pr\left(\Phi(T) < \frac{\alpha}{2}\right) + Pr\left(1 - \Phi(T) < \frac{\alpha}{2}\right) = Pr\left(\Phi(T) < \frac{\alpha}{2}\right) + Pr\left(\Phi(T) > 1 - \frac{\alpha}{2}\right) = Pr\left(\mathrm{T} < \Phi^{-1}\left(\frac{\alpha}{2}\right)\right) + Pr\left(T > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) = \Phi\left(\Phi^{-1}\left(\frac{\alpha}{2}\right)\right) + 1 - \Phi\left(\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) = \alpha$. ■

Corollary 1.2: The power of this two-sided significance test to detect a treatment effect $\Delta\tau = \tau - \tau_*$ is asymptotically,

$$\mathrm{Power}(\Delta\tau) = \Phi\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) + \frac{\Delta\tau}{\nu}\right) + \Phi\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) - \frac{\Delta\tau}{\nu}\right).$$

Proof: Theorem 1 implies that the test statistic $T = \frac{\widehat{\beta}_1 - \tau_*}{\nu}$ will follow the distribution $\mathcal{N}\left(\frac{\Delta\tau}{\nu}, 1\right)$ in the asymptotic regime. The null-hypothesis will not be rejected if $Pr\left(\Phi(T) > \frac{\alpha}{2}\right) + Pr\left(1 - \Phi(T) > \frac{\alpha}{2}\right) = Pr\left(\Phi(T) > \frac{\alpha}{2}\right) + Pr\left(\Phi(T) < 1 - \frac{\alpha}{2}\right) = Pr\left(\mathrm{T} > \Phi^{-1}\left(\frac{\alpha}{2}\right)\right) + Pr\left(T < \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) = \Phi\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) + \frac{\Delta\tau}{\nu}\right) + \Phi\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) - \frac{\Delta\tau}{\nu}\right)$. ■

Corollary 1.3: If $F = f(X)$ is one dimensional, then the asymptotic variance simplifies to,

$$\nu^2 = \frac{1}{n}\left(\frac{\sigma_0^2}{1-\pi} + \frac{\sigma_1^2}{\pi} + \frac{\theta^2 - 2\theta_*\theta}{\pi(1-\pi)}\right)$$

in which write $\theta = (1-\pi)\rho_0\sigma_0 + \pi\rho_1\sigma_1$, $\theta_* = \pi\rho_0\sigma_0 + (1-\pi)\rho_1\sigma_1$, with $\rho_0 = \mathrm{Cov}[Y_0, F]/\sqrt{\mathrm{Var}[Y_0]\mathrm{Var}[F]}$ and $\rho_1 = \mathrm{Cov}[Y_1, F]/\sqrt{\mathrm{Var}[Y_1]\mathrm{Var}[F]}$.

Proof: If $F = f(X)$ is one dimensional then we can write $\xi = (1-\pi)\rho_0\sigma_0\sigma_F + \pi\rho_1\sigma_1\sigma_F$, $\xi_* = \pi\rho_0\sigma_0\sigma_F + (1-\pi)\rho_1\sigma_1\sigma_F$, and $V = \mathrm{Cov}[F] = \sigma_F^2$. Plugging these into the formula for the asymptotic variance gives $\nu^2 = \frac{1}{n}\left(\frac{\sigma_0^2}{1-\pi} + \frac{\sigma_1^2}{\pi} + \frac{\xi^2 - 2\xi_*\xi}{\pi(1-\pi)\sigma_F^2}\right)$. Plugging in $\theta = \xi/\sigma_F$ and $\theta_* = \xi_*/\sigma_F$ gives the desired result. ■

Corollary 1.4: If $F = f(X)$ is one dimensional, and equal numbers of subjects are randomized to active treatment and control, then the asymptotic variance further simplifies to,

$$\nu^2 = \frac{2}{n}\left(\sigma_0^2 + \sigma_1^2 - \frac{(\rho_0\sigma_0 + \rho_1\sigma_1)^2}{2}\right)$$

Moreover, this variance is always less than that of the unadjusted estimator.

Proof: Follows directly from Corollary 1.3 by simple algebra. ■

Theorem 2: Let $\mathcal{F}$ be the set of all functions $f: X \to \mathcal{R}$ that map the baseline covariates to a real number. If the treatment effect is constant and the variances are equal, that is $\mu_1(X) = \tau + \mu_0(X)$ and $\sigma_0 = \sigma_1$, then the asymptotic variance of the treatment effect estimated by ANCOVA is minimized by the choice $f(X) = \mathrm{E}[Y_0|X]$.

Proof: If the treatment effect is constant and $\sigma_0 = \sigma_1 = \sigma$, then $\rho_0 = \rho_1 = \rho$ and $\theta = \theta_* = \rho\sigma$. Therefore, the asymptotic variance of the ANCOVA estimator simplifies to

$$v^2 = \frac{\sigma^2}{n}\left(\frac{1}{1-\pi} + \frac{1}{\pi} - \frac{\rho^2}{\pi(1-\pi)}\right)$$

The variance is minimized when ρ^2 is maximized, which is equivalent to maximizing the correlation between $\mu_0(X)$ and $f(X)$. The correlation of a random variable with another is always maximized by a linear transformation of that random variable itself, which implies that $f(X) = a + b\mu_0(X) = a + b\mathrm{E}[Y_0|X]$. Technically, this holds for any $b \neq 0$, but we will generally set $a = 0$ and $b = 1$. ■

Appendix 4. Details of Simulation Studies

Each of our simulation scenarios is defined by particular choices for the pair of distributions $P_H(X', Y')$ and $P(X, Y_0, Y_1)$. In all cases, the distribution of covariates in the simulated historical and trial data were 10-dimensional uniform random variables in the prism $[l, h]^{10}$. Distributional shift was modeled by choosing different values of l and h for $P_H(X')$ and $P(X)$. The distributions $P_H(Y'|X')$, $P(Y_0|X)$, and $P(Y_1|X)$ were of a Gaussian quadratic-mean form $\mathcal{N}(aX^T\mathbf{1}X + bX^T\mathbf{1} + c, 1)$ in all scenarios. The parameter a controls the degree of non-linearity, with $a = 0$ representing the linear case. In this context, treatment effect heterogeneity refers to the situation in which a or b is different for $P(Y_0|X)$ and $P(Y_1|X)$. Large constant effects are encoded with different values for c in $P(Y_0|X)$, and $P(Y_1|X)$ while keeping a and b the same. The specific values of l, h for each covariate distribution and of a, b, and c are shown in Table 6.

Table 6. Distributional shifts and degrees of non-linearity for each scenario

Scenario	$P_H(X')$		$P(X)$		$P_H(Y'\|X')$			$P(Y_0\|X)$			$P(Y_1\|X)$		
	l'	h'	l	h	a'	b'	c'	a_0	b_0	c_0	a_1	b_1	c_1
Linear	-1	1	-1	1	0	1	0	0	1	0	0	1	0
Non-linear	-1	1	-1	1	0.5	1	0	0.5	1	0	0.5	1	5
Heterogeneous	-1	1	-1	1	0.5	1	0	0.5	1	0	0	1	0
Shifted	-2	0	-1	1	0.5	1	0	0.5	1	0	0.5	1	0

In each simulation scenario, we generated a historical control dataset $(\boldsymbol{X}', \boldsymbol{Y}')$ by drawing 10,000 identically and independently distributed samples from a specified distribution, $P_H(X', Y') = P_H(Y'|X')P_H(X')$. These simulated historical data were used to train to a random forest (1000 trees, with other parameters set to defaults in the python package sklearn [44]) as a prognostic model, $m: \mathcal{X} \to \mathcal{Y}$, Then, we simulated a randomized trial dataset $(\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{Y})$ with 500 subjects, equally randomized to the active and control arms. The data-generating process for these data involved drawing 500 IID samples from a counterfactual distribution $P(Y_1, Y_0, X) = P(Y_1|X)P(Y_0|X)P(X)$, evenly splitting the sample into active treatment and control arms, and then setting $Y = Y_1$ for the actively-treated subjects and $Y = Y_0$ for the controls. Finally, we used the prognostic model to generate the prognosis, $\boldsymbol{M} = m(\boldsymbol{X})$, and analyzed the data varying the estimation procedures.

Three estimation procedures were used: unadjusted, adjusted with the estimated prognostic score obtained with the random forest, and adjusted with the exact prognostic score. The three estimation procedures were repeated for models with and without additional baseline covariates included. Where baseline covariates were included, the full set of baseline covariates leveraged for the random forest models was used such that the standard covariate adjustment had access to the same amount of data as the models used to produce the estimated prognostic scores.

The result was a set of 24 effect estimates (4 data generation models times 3 estimation procedures without additional covariates, plus 4x3 with additional covariates). We calculated the squared-error of each estimate relative to the true treatment effect, which is known from

the data-generating counterfactual distribution, repeated this process 10,000 times, and averaged the squared-errors to obtain mean-squared errors for each estimator.

Appendix 5. Historical Data Sources

A.5.1 Quinn et al. Dataset

Our empirical demonstrations use a trial reported by Quinn et al. [24] that was conducted to determine if docosahexaenoic acid (DHA) supplementation slows cognitive and functional decline for individuals with mild to moderate Alzheimer's disease. The trial was performed through the Alzheimer's Disease Cooperative Study (ADCS), a consortium of academic medical centers and private Alzheimer disease clinics funded by the National Institute on Aging to conduct clinical trials on Alzheimer disease.

Quinn et al. randomized 238 subjects to the active treatment arm administered DHA, and 164 subjects to the control arm administered placebo. The primary outcome of interest for our reanalysis was the increase in the Alzheimer's Disease Assessment Scale - Cognitive Subscale (ADAS-Cog 11, a quantitative measure of cognitive ability) [33] over 18 months. Increase in the Clinical Dementia Rating (CDR) score [34] was a secondary endpoint. The trial measured a number of variables at baseline including demographics and patient characteristics (e.g., sex, age, region, weight), lab tests (e.g. blood pressure, ApoE4 status [35,36]), and component scores of cognitive tests. The baseline variables from the Quinn et al. study that were used in our reanalysis are shown in Table 7, which also contains the variables included in the two datasets of historical data used to train the prognostic models.

Table 7. Variables included in the Quinn et al., 2010 trial and in the two historical datasets used to train the prognostic models.

Variable	CODR-AD	ADNI	Quinn et al., 2010
Age	Yes	Yes	Yes
Education Level	No	Yes	Yes
Region	Yes	Yes	Yes
Sex	Yes	Yes	Yes
Placebo Controlled	Yes	Yes	Yes
Baseline Visit	Yes	Yes	Yes
Height	Yes	Yes	Yes
Weight	Yes	Yes	Yes
Heart Rate	Yes	Yes	Yes
Diastolic Blood Pressure	Yes	Yes	Yes
Systolic Blood Pressure	Yes	Yes	Yes
History of Arterial Hypertension	Yes	Yes	Yes
History of Diabetes Mellitus Type II	Yes	Yes	Yes
Acetylcholinesterase Inhibitors and Memantine Concomitant Medication Usage	Yes	Yes	Yes
ApoE ε4 Allele Count	Yes	Yes	Yes
CSF Total Tau	No	Yes	No
CSF Phosphorylated Tau 181	No	Yes	No
Amyloid Status	No	Yes	No
ADAS Cancellation	Yes	Yes	No

Variable	CODR-AD	ADNI	Quinn et al., 2010
ADAS Commands	Yes	Yes	Yes
ADAS Comprehension	Yes	Yes	Yes
ADAS Construction	Yes	Yes	Yes
ADAS Delayed Word Recall	Yes	Yes	No
ADAS Ideational	Yes	Yes	Yes
ADAS Naming	Yes	Yes	Yes
ADAS Orientation	Yes	Yes	Yes
ADAS Remember Instructions	Yes	Yes	Yes
ADAS Spoken Language	Yes	Yes	Yes
ADAS Word Finding	Yes	Yes	Yes
ADAS Word Recall	Yes	Yes	Yes
ADAS Word Recognition	Yes	Yes	Yes
CDR Community	Yes	Yes	Yes
CDR Home and Hobbies	Yes	Yes	Yes
CDR Judgement	Yes	Yes	Yes
CDR Memory	Yes	Yes	Yes
CDR Orientation	Yes	Yes	Yes
CDR Personal Care	Yes	Yes	Yes
MMSE Attention and Calculation	Yes	Yes	Yes
MMSE Language	Yes	Yes	Yes
MMSE Orientation	Yes	Yes	Yes
MMSE Recall	Yes	Yes	Yes
MMSE Registration	Yes	Yes	Yes
Alanine Aminotransferase	Yes	No	Yes
Alkaline Phosphatase	Yes	No	Yes
Aspartate Aminotransferase	Yes	No	Yes
Total Cholesterol	Yes	No	Yes
Creatine Kinase	Yes	No	Yes
Creatinine	Yes	No	Yes
Eosinophils	Yes	No	Yes
Gamma Glutamyl Transferase	Yes	No	Yes
Glucose	Yes	No	Yes
Hematocrit	Yes	No	Yes
Hemoglobin	Yes	No	Yes
Hemoglobin A1c	Yes	No	No
Indirect Bilirubin	Yes	No	No
Lymphocytes	Yes	No	Yes
Monocytes	Yes	No	Yes

Variable	CODR-AD	ADNI	Quinn et al., 2010
Platelets	Yes	No	Yes
Potassium	Yes	No	Yes
Sodium	Yes	No	Yes
Triglycerides	Yes	No	Yes
Vitamin B12	Yes	No	Yes
Serious Adverse Events	Yes	Yes	Yes

A.5.2 CPAD and ADNI Datasets

The random forest and probabilistic neural network prognostic models were trained using a common dataset that was created by combining data from two databases of longitudinal clinical data on Alzheimer's Disease.

One source is the C-Path Online Data Repository for Alzheimer' s Disease (CODR-AD), a database provided by the Critical Path for Alzheimer's Disease (CPAD) consortium [29,30] that consists of the control arms of 29 mostly mild to moderate AD clinical trials with more than 7000 subjects. The other source is from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [28], a collection of four long-running observational studies enrolling subjects across the AD disease spectrum since 2004, focusing primarily on MCI and cognitively normal subjects.

Some studies in the CODR-AD and ADNI databases did not have sufficient data for inclusion in the training set. First, we excluded all cognitively normal subjects from the ADNI study and excluded the ADNI3 study because it did not have sufficiently frequent visits to model the longitudinal behavior of the outcomes of interest. In addition, studies from CODR-AD that did not record ADAS-Cog were excluded. The resulting dataset had 6,919 subjects and 34,224 subject-visits across 21 studies, with approximately 25% of subjects with MCI and the remainder with AD.

Even after excluding some of the studies with data that did not conform to our requirements, the remaining studies still had varying duration, visit intervals, inclusion criteria, and measured variables. For example, the ADNI studies typically have a 6-month cadence and an extremely broad set of observations that include imaging, biomarker data, and important disease severity measures. We reprocessed both databases to extract measured variables and encoded them into a consistent wide-form ("tidy") tabular format [45]. The detailed steps used for processing the data from ADNI and CPAD to create the common training dataset are described in Bertolini et al. [32] and Fisher et al [18].

Sixty-three variables were selected for inclusion in the prognostic models based on clinical significance determined by recommendations from subject-matter experts and on the availability of data. These variables are shown in Table 7.

Note that the set of variables included in the training set for the prognostic model was determined before examining the variables present in the Quinn et al. study. This resulted in some missing baseline observations when computing the prognostic scores, but this also

models a realistic scenario that would likely be encountered in a clinical trial with a pre-specified prognostic model.

Only 70% of the combined ADNI/CPAD dataset was used for training the prognostic models. We refer to this portion of the dataset as the "training dataset". The remaining 30% was held-out and used to estimate the standard deviations and correlations required for the sample size calculations for PROCOVA™. We refer to this portion of the dataset as the "test dataset".

A.5.3 Data Availability

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

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

Certain data used in the preparation of this submission were obtained from the University of California, San Diego Alzheimer's Disease Cooperative Study Legacy database; that data are from the Quinn et al. study [24].

Appendix 6. Prognostic Models

A.6.1 Variable Selection for Empirical Analyses

A 2019 study conducted by Fisher et al [18], which focused on the use of probabilistic neural networks for forecasting disease progression in patients with AD, demonstrated that component-wise longitudinal models of clinical data could be useful for predicting composite score outcomes, including ADAS-Cog11. An updated version of that model, trained on a larger dataset described above, corresponds to the probabilistic neural network prognostic model described below. Therefore, similar criteria were used for selecting the variables; that is, clinical significance for AD assessed by subject-matter experts, and the fraction of time a variable was observed in the training set. Notably, the variables were selected without reference to the Quinn et al. study, which mirrors pre-specification of a prognostic model in a prospective trial. Variables that were selected for use in the prognostic model, but weren't measured in the Quinn et al study, were treated as missing observations and were imputed as described in the relevant modeling sections.

A.6.2 Random Forest Model

A random forest is a standard supervised learning algorithm capable of capturing nonlinear input-output relationships [31]. More precisely, a random forest is trained to learn a function $f: X \to Y_0$ that minimizes the mean squared error of the predictions. This amounts to learning a function $f(X) \simeq \mathrm{E}[Y_0|X]$. In theory, a consistent estimator will learn this function exactly in the limit that the number of samples in the training set goes to infinity, but the trained predictor will always be an approximation for any finite training dataset.

For our empirical analyses, we need prognostic models for two outcomes, the change in the ADAS-Cog11 and CDR-SB scores at 18 months. Therefore, we used the training dataset to fit two different random forests, one to predict the change in ADAS-Cog11 over 18 months and another to predict the change in CDR-SB over 18 months. Any missing baseline observations were mean imputed during training. The random forests were fit using the python software library Scikit Learn [44] using 1000 trees.

As mentioned previously, some of the baseline variables from the training dataset were missing when using the trained random forests to predict prognostic outcomes for the subjects in the Quinn et al. study. Therefore, any missing baseline variables were imputed with the mean of that variable measured in the training dataset.

For each outcome (ADAS-Cog11 and CDR-SB) and visit, a random forest model is trained to predict that endpoint. For each subject, the prediction of the random forest model is used as the prognostic covariate in the estimator.

A.6.3 Deep Learning Prognostic Model

We referred to the random forest as supervised learning model because it was specifically trained to minimize the mean squared prediction error on the change in ADAS-Cog11 and CDR-SB over 18 months. Unsupervised learning with probabilistic neural networks offers a different approach, in which one aims to train a model on the probability distribution of all of the variables over time. This probability distribution can be used to generate multivariate

clinical trajectories, and predictions of any desired quantity can be computed from these sampled clinical trajectories.

Our aim here is not to advocate for one or the other approach to prognostic modeling, only to provide one example of each type of algorithm as demonstrations for how they can be used with PROCOVA™. The properties of PROCOVA™ do not depend on the particular type of prognostic model.

Let $X(t)$ denote the vector of the variables in Table 7 measured at time $t = 0, 3, 6, ...$, with t in units of months and $t = 0$ denoting the baseline visit. The probabilistic neural network is trained to model the probability distribution $P\big(X(3), X(6), ... \mid X(0)\big)$. With this distribution, we can compute the expected value of any function of the variables such as the change in ADAS-Cog11 or CDR-SB over 18 months.

The particular type of probabilistic neural network used here is called a Conditional Restricted Boltzmann Machine (CRBM) [18,32,46–51]. A CRBM models the temporal dynamics as a Markov process,

$$p\big(X(t = 0, \dots, T)\big) = p\big(X(t = 0, \dots, L)\big) \prod_{t=L}^{T} p\big(X(\tau) \big| X(t = \tau - 1, \dots, \tau - L)\big)$$

in which the transition operator has a form

$$p\big(X(\tau) \big| X(t = \tau - 1, \dots, \tau - L)\big) = Z^{-1} \int d\, h\, e^{-U\left(X(\tau), h \mid X(t = \tau - 1, \dots, \tau - L)\right)}.$$

Here, L is the lag of the Markov process, and the interactions between variables are mediated through a vector of hidden variables, denoted h. The energy function, U, depends on parameters that are optimized by stochastic gradient descent to maximize the likelihood of the data and/or to minimize the ability for another model to distinguish between samples from the data distribution and samples from the model distribution. Details on the training of this model are provided in Bertolini et al [32].

The probabilistic neural network (i.e., CRBM) and the random forest were both trained on the training dataset, with the same input variables. The CRBM used all of the timepoints and modeled the distribution of all variables across all times, from which the predicted changes in ADAS-Cog11 and CDR-SB at 18 months could be computed. In contrast, the random forest only predicted the 18-month changes in ADAS-Cog11 and CDR-SB. Even though the two prognostic models are very different, they can both be used to generate the prognostic scores for PROCOVA™, providing increases in power while guaranteeing type-I error rate control.

Appendix 7. Details of Sample Size Estimation

Estimating σ_w^2 and ρ_w for power calculations

One method for obtaining estimates for the marginal potential outcome variances ($\boldsymbol{\sigma_2^2}$) and potential outcome-prognostic score correlations ($\boldsymbol{\rho_w}$) is to use prior data, for example data from the placebo control arm of a previous trial performed on a similar population. In this case, we presume we have access to a vector $Y'' = \left[Y_1'' \ldots, Y_{n''}''\right]$ of outcomes for these subjects and their corresponding prognostic scores $M'' = \left[M_1'' \ldots M_{n''}''\right]$, calculated by applying the prognostic model, m, to each subject's vector of baseline covariates X, i.e., $M_i'' = m(X_i'')$.

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

$$\hat{\sigma}_0^2 = \frac{1}{n'' - 1}\sum_i (Y_i'' - \bar{Y}'')^2 \qquad \text{Eq. 8}$$

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

$$\hat{\rho}_0 = \frac{\sum_i (Y_i'' - \bar{Y}'')(M_i'' - \bar{M}'')}{\sqrt{\sum_i (Y_i'' - \bar{Y}'')^2}\sqrt{\sum_i (M_i'' - \bar{M}'')^2}} \qquad \text{Eq. 9}$$

which is the usual sample correlation coefficient.

It is common that a prognostic model may perform differently in new population. If the variance in the trial population is larger than estimated, or if the magnitude of the correlation between the predicted and observed outcomes is smaller than estimated, then using these estimates could lead to a trial with lower than anticipated power. Therefore, it can be useful to inflate the estimated variance by a factor $\gamma_0 \geq 1$, or to deflate the correlation by a factor $\lambda_0 \in [0,1]$, in order to buffer against this effect.

It is usually difficult to directly estimate the corresponding values for the active treatment arm due to scarcity of data from subjects receiving the investigational therapeutic. As a result, we propose using the estimates obtained from the database of historical controls, $\hat{\sigma}_0^2$ and $\hat{\rho}_0$. Moreover, this approach is in-line with the constant treatment effect scenario outlined in Theorem 2. As before, it can be useful to inflate the estimated variance by a factor $\gamma_1 \geq 1$, or to deflate the correlation by a factor $\lambda_1 \in [0,1]$.