20 September 2022
EMADOC-1700519818-907465
Committee for Medicinal Products for Human Use (CHMP)

Qualification opinion for Prognostic Covariate Adjustment (PROCOVA™)

Draft agreed by Scientific Advice Working Party (SAWP)	10 February 2022
Adopted by CHMP for release for consultation	24 February 2022[1]
Start of public consultation	22 March 2022 [2]
End of consultation (deadline for comments)	03 May 2022[3]
Adopted by CHMP	15 September 2022

Keywords	Qualification of Novel Methodology, Statistical methodology, Prognostic Covariate Adjustment, Sample size estimation

[1] Last day of relevant Committee meeting.
[2] Date of publication on the EMA public website.
[3] Last day of the month concerned.

CHMP Qualification Opinion overview

The qualification opinion document uses the following structure:

Separate to this document, but relevant to the qualification opinion, are the Applicant's PROCOVA handbook, and the responses to the responses to the first and second EMA list of issues.

1. CHMP Qualification Opinion statement and conclusion

Prognostic Covariate Adjustment (PROCOVA) is a 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.

The Applicant proposes the PROCOVA method for estimation and statistical inference on the treatment effect in randomized controlled clinical trials measuring continuous outcomes. The procedure involves developing a prognostic score for the outcome under control based on a historical data set that is independent from the study data and then applying the prognostic score as covariate in an ANCOVA model for the actual data analysis of a clinical trial.

The methodology comprises three steps:

> Step 1: Training and evaluating a prognostic model to predict outcomes under the control condition (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.

CHMP qualifies PROCOVA as prognostic score adjustment and the proposed procedures, as described in a handbook for trial statisticians, could enable increases in power or precision of treatment effect estimates in controlled randomised clinical trials with continuous outcomes. The presented mathematical properties, simulation exercises and empirical application support this use. The assumed reduction in residual variance due to a prognostic score may in principle be taken into account to reduce sample size, if it can be ensured that the calculation is considering uncertainties in the assumptions made, and if the resulting sample size is large enough to meet other relevant purposes of the clinical trial apart from the primary hypothesis test and treatment effect estimation.

Regarding the mathematical properties of PROCOVA, as implemented, the method can be regarded a special case of ANCOVA sharing the properties of type I error control and asymptotically unbiased estimates of the treatment effect with sufficiently large sample sizes. The method uses a number of assumptions that are similar to those required by a large variety of parametric frequentist methods that are regularly applied and accepted from a regulatory perspective. Therefore, the proposed prognostic covariate procedure is an acceptable statistical approach for primary analysis of clinical trials.

CHMP does not intend to single out any specific method for statistical modelling using adjustment for covariates as 'the' method to be used, but qualifies use of methods as acceptable ones. There are other options than as proposed by PROCOVA to implement covariate adjustment with or without use of historical data. Depending on the specific trial setting and the expected sample size, Applicants may

compare the properties of candidate methods that apply adjustment for covariates, and they may not only compare these to an ANCOVA with individual covariate adjustment. Within the approaches that are deemed acceptable and sufficient according to regulatory requirements, the Applicant may select the one that they consider to serve best the objectives of their clinical trial.

An attainable advantage over using ANCOVA with single covariate adjustment should be justified to support application of the PROCOVA method. The Applicant demonstrates that this should be the case if the prognostic score is able to capture a nonlinear relationship between covariates and outcomes of interest. The potential sample size reductions using PROCOVA depend on the ratio of the correlation between a single baseline covariate (or a linear combination of the covariates that would typically be considered in the analysis) and the outcome and the correlation between the single prognostic (PROCOVA) score and the outcome. The gain in sample size (or likewise in power or precision of the treatment effect estimates) should be evaluated at the stage of planning a trial, also taking the optimism due to prognostic model building into account. The relative pros and cons of using PROCOVA or ANCOVA should be compared to make a final determination to choose one of the three paths: no adjustment, ANCOVA with one or more pre-specified covariates, or PROCOVA. The PROCOVA handbook provides step-by-step instructions for the trial statistician to make an informed choice among these three paths.

The potential advantages of the PROCOVA procedure and prognostic score adjustment in general depend on the availability of appropriate historical data and the derivation of a prognostic model that would allow outcome prediction in a future clinical trial. The number of covariates that can be included in the modelling approach is determined by the size and quality of the historical dataset(s). Establishing external validity of historical data is of paramount importance when applying a prognostic model in a future clinical trial. Type I error control, unbiased effect estimation and confidence interval coverage are not dependent on the choice or performance of the prognostic model. PROCOVA can be used together with adjustment using additional covariates and stratified randomisation, but the consequence of using the prognostic score together with additional prognostic covariates (one or more) needs to be carefully considered. Where such additional covariates and/or stratification factors are already included in the prognostic score, the impact of the potential multicollinearity on the precision of the estimated coefficients may outweigh the proposed advantage. Recommendations given in the handbook for trial statisticians on subgroup analysis should be followed. CHMP notes that the impact of multicollinearity when applying PROCOVA is not fully understood and additional research is desirable. In addition to the recommendations in the handbook for use of PROCOVA for a case that involves adjusting for additional covariates, including stratification factors for trials with stratified randomization, an alternative option to exclude these additional covariates from the prognostic score model may be explored before application.

CHMP cannot qualify a formalised procedure for prognostic model development in Step 1 as part of the PROCOVA method. Only specific settings were explored and it cannot be foreseen if successful outcome prediction from each model development will be possible for the proposed very general context of use. However, it is emphasised that the three steps are not independent, and Step 1 is a necessary part of the procedure. It is acknowledged that this puts the burden on Applicants who want to use the procedure in specific settings. There is existing and increasing literature on derivation of prognostic models, and existing knowledge should be taken into account. There may be disease conditions for which prediction of endpoints selected for clinical trials is not possible with a desired precision or was not successful in previous settings. Outcomes from historical data may not allow prediction of control arm outcomes of future trials in case of changes in the therapeutic landscape. In addition, CHMP cannot issue a statement about the precision of prognostic models in general and over therapeutic areas and if these models would allow meaningful improvement in power or reductions in sample size. However, it is noted that prognostic models could help understanding disease characteristics or even mechanistic properties.

The chosen approach to prognostic model development is according to the Applicant explicitly out of scope of this qualification procedure. There are advances in statistical 'learning' methods, the ability to

handle high-dimensional data and progress with e.g., machine learning or deep learning methods. However, derivation of a prognostic model would require careful work by Applicants or independent groups with access to appropriate data sets. Applicants should be aware of the risk of overfitting when using more complex prognostic modelling approaches, including machine learning and artificial intelligence methodology. Therefore, assessment of correlation between observations and outcomes with data independent of training data is of importance to avoid too optimistic estimates of this correlation. The updated handbook provides guidance for the choice of a deflation factor λ, and for the conduct of sensitivity analyses taking into account a potential over-optimism of the prognostic model and the fact that the correlation of the prognostic score with the outcome may be smaller in a future trial that investigates the experimental treatment.

In simulations performed by the Applicant, potential differences between the historical population used to derive the prognostic model and the trial population were addressed with simulations using a 'shifted population'. While this is acknowledged, the robustness of the planned PROCOVA approach with regard to availability of covariates in historical and future data, data quality with regard to misspecification or measurement error and missing or incomplete covariates need to be carefully assessed.

Approaches with non-linear models for analysis and direct comparisons to such models, as well as models with treatment-by-covariate interactions are out of scope of this qualification procedure.

2. Executive Summary as submitted by the Applicant

The 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.

This methodology (outlined in the **Novel Methodology** section below) 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.

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.

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

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. Although such methods do generally increase power, they cannot strictly control the type-I error rate reducing the relevance of such methods, particularly for pivotal/ confirmatory/ Phase 3 RCTs. 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.

The Novel Methodology

We propose a novel approach that leverages historical control arm data and predictive modelling 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). Our approach is entirely pre-specifiable, is generic enough to be integrated into many analysis plans and is supported by regulatory guidance.

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. 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 meagre. 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.

Adjustment for composite or computed covariates such as body mass index, Charlson comorbidity index, or Framingham risk score, is not new. 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, analysable 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.

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 simulation and empirical 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, 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 and empirical examples/context-of-use 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). Our empirical context-of-use demonstrations involve re-analysis of a Phase 3 trial in patients with AD reported by Quinn et al.

3. Questions on Statistical Properties of PROCOVA from the Applicant and CHMP Answers

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.

Applicant's position

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.

CHMP answer

The Applicant proposes a method, PROCOVA, to perform estimation and statistical inference on the treatment effect in randomized controlled clinical trials. The methodology comprises three steps:

> Step 1: Training and evaluating a prognostic model to predict outcomes under the control condition (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 key idea is to first develop a prognostic score for the outcome based on a historical data set that is independent from the study data and then apply the prognostic score as covariate in an ANCOVA model for the actual data analysis.

Following the Applicant's arguments, modern methods of statistical learning, such as random forests or neural networks could allow for modeling the functional relationship between prognostic variables and the outcome with higher accuracy than e.g. a simple linear combination would provide. Hence, the approach would improve the efficiency of the analysis over other methods of adjustment by providing a prognostic score that is more strongly correlated with the outcome.

The Applicant's position that PROCOVA is a special case of ANCOVA and hence is an appropriate method for the analysis of randomized trials is agreed to with minor comments and proposals, which will be addressed below and in the answers to the specific questions.

The following table summarises the differences between the conventional approach addressing prognostic factors and PROCOVA.

	Standard approach	**PROCOVA**
Design Stage	Most important prognostic factors are identified and considered in the study design (stratification)	A prognostic model is developed and preferably validated (using "external" data set) It is unclear whether the prognostic score will be used for stratification
Sample Size considerations	Sample size is estimated based on α β, difference to be detected and variability based on historical studies	Sample size is estimated based on α β, difference to be detected and variability, as well as ρ (correlation coefficient between prognostic

	The gain in efficiency including covariates may be incorporated (which is not often done in practice) Sensitivity of sample size estimates with respect to assumptions taken is evaluated	index and outcome) based on historical studies Uncertainty in variability and prognostic ability is accounted for (using parameters λ and γ)
Analysis	Stratification factors (and possibly other variables) are included as covariates in the regression model	A single prognostic index (and possibly other variables) are included as covariate(s) in the regression model

Overall, there are two major differences between the conventional approach and PROCOVA:

- the method to evaluate the robustness of the sample size estimate, which will be addressed in the answer to Questions 3 and 5

- the inclusion of a single covariate using fixed weights to combine important baseline covariates, which will be addressed in the answer to Questions 2 and 4.

With regard to the answer to Question 1, CHMP would like to refer to the proposed context of use. The Applicant suggests that the approach represents a special case of analysis of covariance (ANCOVA) that can be performed in a large-sample setting using standard linear regression. It is claimed that it can use historical data to reduce the variance of the treatment response estimates better than other available approaches, potentially reducing the minimum sample size required to achieve the same level of confidence. The methodology is recommended for use in trials with continuous variables for which historical data in a similar patient population is available that allows building a prognostic model to predict control outcomes with sufficient accuracy using the measured baseline covariates for the subjects. The variables used by the prognostic model must be measured at baseline for subjects in the historical data set and the new clinical trial.

Theorem 1 and corollaries 1.1 to 1.4 of the Mathematical Results section in the briefing document are acknowledged. These demonstrate analytically important properties of the PROCOVA method in a controlled parallel group clinical trial setting with equal randomisation to the groups.

CHMP agrees that the proposed method is an application of an ANCOVA model in which a predefined prognostic score is used as covariate. Properties regarding bias and control of type I error rate will be those of usual ANCOVA models. I.e., in a randomized trial, treatment effect estimates will be asymptotically unbiased and finite sample bias will typically be negligible. The type I error rate is controlled asymptotically under the assumption of equal variances in both groups or equal group sizes. Indeed, in this setting the asymptotic variance of a covariate-adjusted treatment effect estimate is lower than the variance of an unadjusted estimate, if there is a non-zero correlation between the covariate and the outcome, hence adjusting for prognostic covariates is generally beneficial in terms of power.

An important prerequisite for PROCOVA to inherit the properties of ANCOVA is that the definition of the prognostic score is independent of the study data, and this point is obviously acknowledged by the Applicant.

For further considerations on the conditions defined in the question by the Applicant and the consequences for the proposed context of use (Questions 4 to 6), please see the answers of the following questions.

Question 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?

Applicant's position

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.

CHMP answer

The Applicant shows, under the assumption of a constant treatment effect across all covariate values and the assumption of equal variances of the outcome variable under treatment and control, that an ANCOVA model that is adjusted for the true functional relationship between covariates and outcome results in minimal variance of the treatment effect estimate among all models that are adjusted for a function of the same covariates. This is an intuitive, albeit relevant result. The sample size of the clinical trial must be large enough to ensure that the asymptotic variance is a reasonable estimate for the variance. Some additional, weaker assumptions commonly applied for statistical modelling are also needed (Schuler et al., arXiv:2012.09935v2 2021). Under these conditions, it can generally be agreed that the proposed prognostic covariate procedure can achieve a lower variance of the treatment effect estimate if the correlation of the prognostic score with the outcome of interest is higher.

Extensive modelling (and model validation) to attain a prognostic index (linear or non-linear predictor of baseline variables) is a valuable exercise in general in order to predict the natural disease course (or the disease course under some standard therapy). The reduction of variance of treatment effect estimates due to adjustment for prognostic covariates is well established and will be achieved with the proposed method if the applied score is correlated with the outcome.

The relevant difference between usual ANCOVA models and the proposed PROCOVA method is that the latter aims to use a prognostic score that is close to the true functional relationship between the included covariates and the outcome under the control condition. In contrast, ANCOVA usually is used with (a limited number of) linear predictors without interactions such that a linear approximation to the true functional relationship is applied. It is agreed that a model that resembles the true functional form more closely will likely produce a treatment effect estimate with lower variance.

A drawback of PROCOVA, however, is that the prognostic score must be prespecified including a scale factor, and weights used within the score cannot be adjusted to possible differences between the training setting and the actual trial setting. In contrast, in a usual ANCOVA model the functional relationship is a linear approximation, but it is chosen optimal to the observed data among all linear approximations. There may be situations in which the optimal linear approximation may outperform the approximation by a function that is correct in principle, but has misspecified coefficient values.

A particular situation where coefficient values may differ between training and trial data sets may arise if the distribution of an included variable is different in the training and the trial population and the prognostic score does not perfectly resemble the true relationship but is still an approximation. For illustration, consider the case of a true quadratic relationship and a linear approximation: The slope of the best linear approximation depends on the distribution of the covariate values across patients and even if the slope was completely known for a training population, it would not be the optimal choice in an analysis model for a different population with another distribution of the covariate where a model that estimates the required coefficient from the data may be more efficient. The impact of such distributional inhomogeneities that may occur in the practical application of PROCOVA should be investigated in advance (using simulation experiments).

The simulation studies performed to support the statement of Theorem 2 in four different scenarios with variations of the strict assumptions (outcome-covariate relationship linear, outcome-covariate relationship linear non-linear, conditional average treatment effect not constant, shifted trial population) are appreciated. They show that even if these assumptions are not strictly fulfilled, the mean squared errors with prognostic covariate adjustment were lower than without. This is acknowledged.

The empirical application to existing data sets shows that the postulated decrease in variance can be attained in a realistic scenario with real data and is considered supportive for application of the proposed procedures.

Of note, the prognostic score may be used together with further covariates as the Applicant explored in one of their simulation experiments. SAWP issued a second list of issues that addressed multicollinearity when implementing stratified randomisation in trials using individual baseline covariates and PROCOVA at the same time. The Applicant provided a written response to this second list of issues and an updated handbook to be used by trial statisticians when applying PROCOVA. When applying PROCOVA together with stratified randomisation, a linear model for primary analysis adjusting for the prognostic score and any additional pre-specified baseline covariate(s), provides an unbiased point estimate of the treatment effect in the overall trial population. However, this primary analysis model does not produce an unbiased estimate of a subgroup effect. The instructions for trial statisticians state that subgroup effects or treatment-by-subgroup interactions should not be evaluated using the same linear model that is used for primary analysis of the treatment effect, since applying this model may introduce multicollinearity and could impact the accuracy of subgroup-specific treatment effect estimates. It is emphasised by the Applicant that the prognostic score is not intended as a stratification factor.

In addition, it is acknowledged that a prognostic score in PROCOVA may utilise a large number of covariates, if the training data set is sufficiently large, whereas with usual ANCOVA the number of covariates is limited to be much less than the number of included subjects.

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?

Applicant's position

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.

CHMP answer

It can be agreed that applying adjustment for the PROCOVA prognostic score or a set of covariates for ANCOVA in general could lead to a smaller minimum sample size to achieve a desired level of power. As outlined by the Applicant, the minimum sample size is a function of the Pearson correlation coefficient between observations and predictions of the prognostic model. During sample size planning an investigator may take into account explained variation due to covariates, such as the prognostic score in PROCOVA, which will result in smaller sample size than assuming an unadjusted analysis or zero correlation between covariates and outcome. However, overly optimistic assumptions on the effect of covariates may result in too low sample sizes and inconclusive studies. It is noted that the Applicant recommends using a separate data set independent from the training data to estimate the correlation coefficient and thus avoid overestimation of the correlation; this is supported. Please see the answer to Question 5 for further considerations and more detailed comments regarding sample size planning.

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.

Applicant's position

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.

CHMP answer

As outlined in the answer to question 1, CHMP agrees that the proposed method is a special case of ANCOVA. Therefore, similar to other ANCOVA models adjusted for a prognostic score, the proposed method will be acceptable to estimate the treatment effect and perform statistical inference on it in randomized trials. The proposed PROCOVA procedure can be considered an acceptable formal

presentation of approaches that were used in clinical trial settings before when prognostic covariates were included in analysis models, e.g. by imaging based risk scores in oncology or covariate based risk scores in cardiovascular diseases.

Regarding use of linear models for estimation, it is noted that from a regulatory perspective for a primary estimand and analysis, application of a linear ANCOVA model with covariate adjustment would be acceptable even if the linear model does not model the relationship between treatment, covariates and outcomes correctly if an average treatment effect for a population-level summary is targeted. It is though acknowledged that an improved modelling of the true relationship between treatment, (a larger set of) covariates and outcome can be beneficial and can improve the precision of the estimator and could potentially also allow better understanding of conditional treatment effects if relevant in a particular disease setting.

The Applicant proposes to perform statistical inference on the treatment effect using large sample normal approximations to the respective test statistic. While this approach is asymptotically valid, it neglects the variability of the estimate for the residual variance nuisance parameter. It is therefore recommended to use t-distributions (which take into account this variability under the assumption of normally distributed residuals) to avoid too liberal test decisions. This is particularly emphasized as the sample a size may be small in phase II, and even phase III studies. The Applicant agreed during the discussion meeting that using the t-distribution is a reasonable, conservative approach for trials with smaller sample sizes.

The Applicant further proposes to use robust "sandwich" variance estimation in inferential procedures. This is acceptable, however certain properties of the robust variance estimator need to be taken into account: Using a bias-adjusted estimator is required as the small sample bias of the unadjusted robust variance estimator may be considerable. The bias adjustment proposed by the Applicant is acceptable. The robust estimator has larger variability than the model-based estimator. Hence it may not be suitable with small sample sizes. In any case, hypothesis tests and confidence intervals should be based on t-distributions as discussed above. In the discussion meeting, the Applicant pointed out that there is no definite way for choosing the degrees of freedom for a reference t-distribution when using robust variance estimation. This is acknowledged, however using an approximate number of degrees of freedom is considered acceptable. E.g., the work by Lipsitz, Ebrahim and Parzen 1999 on a respective Satterthwaite approximation may be considered (Lipsitz, S. R., Ibrahim, J. G., & Parzen, M. (1999). A degrees-of-freedom approximation for a t-statistic with heterogeneous variance. Journal of the Royal Statistical Society: Series D (The Statistician), 48(4), 495-506).

The following further specific concerns may need to be addressed in an actual application:

1) Since the prognostic score is trained under control conditions, it is possible that its correlation to the outcome is larger under control than under treatment. This could result in unequal residual variance in the two groups, which may lead to inflation of the type I error rate in trials with unequal group sizes. The robust variance estimation as proposed by the Applicant is an acceptable remedy of this issue.

2) A score that includes complex transformations of the considered variables may be prone to result in skewed distribution with some outliers, even if the included variables have unsuspicious distributions at their original scale. Outliers in the prognostic score may turn out to be influential points in fitting the analysis model, which may raise concerns regarding the robustness of results. It is recommended that the PROCOVA analysis should be supported by appropriate model diagnostics to assess the robustness of the analysis results with respect to deviations in single observations.

3) The Applicant claims that with recent methodological developments a prognostic score with considerable correlation can be obtained for a variety of continuous responses in multiple therapeutic areas. Correlation values around 0.4 are considered in the empirical examples and values up to 0.8 are

considered in the theoretical sections. Considering the conventional approach, a strong prognostic index with a correlation of such a magnitude would usually be accounted for in the study planning, e.g. using stratified randomisation. The Applicant clarified during the discussion meeting that the prognostic score to be used in the PROCOVA analysis is not intended to be used for stratification. As the prognostic score is derived from a potentially large set of variables, it is not considered practical to be implemented in the randomization procedure. This aspect was further addressed in a second list of issues, and the updated handbook developed by the Applicant instructs trial statisticians to consider (a limited number of) the strongest prognostic factors for stratified randomization taking into account that (some of) these candidate stratification factors could already be included in the prognostic score.

4) It is expected that data on all variables included in the prognostic score will be collected in the randomised trial. Concerning incomplete data on covariates for prognostic score adjustment, there are be several options and a missing data imputation scheme should be pre-specified. Missing data was further addressed in the second list of issues. Additional instructions were provided for situations where significant differences in data completeness exist between the new trial and the validation dataset. The correlation coefficient R may be lower in a new trial if one or more important variables are expected to be missing frequently (or with a different pattern of missingness). While the prognostic model would be able to generate prognostic scores for all subjects, regardless of missing data, the advantage of PROCOVA may be decreased. Generally, if the proportion of missing data is low and imputation is considered, multiple imputation could be preferable and imputations should not depend on data of post-baseline measurements in the target trial. It is acknowledged that baseline covariates cannot be impacted by intercurrent events.

5) While it is understood that the prognostic score adjustment targets an average treatment effect for a trial population, subgroup analysis based on covariates could be relevant for characterisation of the treatment effect. This would be of particular relevance in case of (expected) differential treatment effects. The Applicant provided further instructions on how such situations should be addressed at the design and analysis stage when using PROCOVA. Please refer to the answer to Question 2.

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?

Applicant's position

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.

CHMP answer

As stated in the answer to Question 3, it is agreed that taking into account explained variation due to covariates, such as the prognostic score in PROCOVA, results in reduced residual variance and hence will result in smaller sample size than assuming an unadjusted analysis.

Nonetheless, for such a planning approach to be acceptable potential uncertainties in the assumption on the variance explained by the prognostic score need to be taken into account. Overly optimistic assumptions on the effect of covariates may result in too low sample sizes and inconclusive studies. Most trials are planned conservatively without taking into account possible gains in power due to adjusting for covariates and the actual power may then be larger than the planning assumption of, e.g. 80% or 90%. Also in usual sample size planning, different assumptions regarding the variance and other relevant parameters are explored to assess the impact of deviations from the made assumptions on the resulting power.

As a first step, an attainable advantage over using ANCOVA with individual covariate adjustment should be justified. The Applicant demonstrates that this should be the case if the prognostic score is able to capture a nonlinear relationship between covariates and outcomes of interest. This is discussed in Schuler et al. (Schuler et al., arXiv:2012.09935v2 2021), and there would be no gain in efficiency when adjusting with a prognostic score assuming a linear relationship between covariates and outcome. During the discussion meeting, the Applicant further elaborated on the attainable advantage of the PROCOVA procedure over ANCOVA with individual covariates. The potential sample size reductions using PROCOVA depend on the ratio of the correlation between a single baseline covariate (or a linear combination of the covariates that would typically be considered in the analysis) and the outcome and the correlation between the single prognostic (PROCOVA) score and the outcome. The gain in sample size (or likewise in power or precision of the estimates) can then be evaluated (graphically) and should also take the optimism due to prognostic model building into account. The relative pros and cons of using PROCOVA or ANCOVA are compared to make a final determination to choose one of the three paths: no adjustment, ANCOVA with one or more pre-specified covariates, or PROCOVA. This issue was raised in a second list of issues and was addressed by the Applicant in a handbook for trial statisticians guiding the application of PROCOVA. The handbook provides guidance to help the trial statistician make an informed choice among the three paths with step-by-step instructions.

In the original procedure described by the Applicant, an inflation parameter (γ) for standard deviation in the control arm, as well as a deflation parameter (λ) for prognostic correlation in both arms need to be selected. The latter has been set to $\lambda=0.9$ in the analysis of the Alzheimer data set. A clear rationale for that choice was not provided. In an actual application, it needs to be carefully considered how λ and γ are chosen. Evaluation of the robustness of the sample size or power estimate with respect to deviations from assumptions, as outlined above, seems generally more informative than relying on the two modifying parameters. At the discussion meeting and in the written responses to CHMP's first list of issues, the Applicant outlined rules of thumb for the choice of the deflation factor λ for the correlation coefficient. The choice is proposed to depend on the extent of model validation. The value may be close to 1 if there was extensive validation using external data sets, it may be chosen conservatively (e.g. $\lambda=0.5$) if the model was developed and validated on the same data set, or it may be decided to not use PROCOVA at all. It was considered important by SAWP to provide the practitioner with such rules of thumb but also to advise conduct of sensitivity analyses to prevent under-powered trials. The updated handbook provides guidance for the choice of the deflation factor λ, and for the conduct of sensitivity analyses taking into account a potential over-optimism of the prognostic model and the fact that the correlation of the prognostic score with the outcome may be smaller under experimental treatment. It should still be kept in mind that the approach using λ and γ may not cover the range of all parameters relevant for assessing the robustness of the sample size and should not be understood as prescriptive by Applicants to account for all uncertainties.

Establishing external validity of historical data was raised as an issue in the second list of issues and the Applicant addressed this with the updated guidance documents. The handbook provides definitions and instructions to validate the prognostic model. Instructions include recommendations to collaborate with model developers to establish the external validity of historical validation data sets. Specific comments are provided on how to match the validation dataset to the trial population, on how to account for the potential changes in the SOC, and how to address different extent of missing data between the validation dataset and the trial data. These instructions are acknowledged. Prognostic model validation using a data set that is independent from the historical training data and from the study data, as proposed by the Applicant, is certainly endorsed to avoid too optimistic estimates of the correlation coefficient. However, the feasibility of this step may be limited by the availability of additional validation data that have similar properties as the planned study data.

Moreover, it should be kept in mind that the sample size of a clinical trial should in most cases be sufficient not only for the primary hypothesis test but also for providing a sufficiently large safety database or, in some cases, to address more than one endpoint or the precision in important subgroups (see Q4).

With regard to the scenarios addressed with the empirical application of PROCOVA provided with the briefing document, these are considered to be of relevance and the results of Experiment 1 and 2 support the application of the proposed procedures. It is noted that data from patients who dropped out of the study were not included in the analysis (p. 21, briefing document). This would not be acceptable for regulatory purposes. It is also noted that the empirical applications mention two outcomes of interest (ADAS-Cog11 and CDR at 18 months). While the sample size in the example cases was calculated for ADAS-Cog11, analyses for CDR are also reported. With respect to co-primary endpoints, the Applicant states in Section 3.1.1 "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 approach is not in general appropriate as it may result in insufficient power to reject all co-primary endpoints simultaneously. Instead, the conjunctive power should be the basis for sample size calculations with co-primary endpoints. However, it is agreed that in case of multiple endpoints of interest using multiple prognostic models or a multivariate prognostic model may be necessary.

The Applicant uses two-sided tests in the sample size and power calculations. Rejections due to observed effects in both directions are counted as rejection of the null hypothesis. It is noted that from a regulatory perspective, only one part of the comparisons may be relevant for study success. This should usually be reflected in the hypothesis testing. With respect to considering the expected dropout rate d, accounting for dropouts in sample size considerations as proposed by the Applicant using $n_d=n/(1-d)$ is generally reasonable. However, typically all randomised subjects should be included in the primary analysis and a strategy to address post-randomisation events affecting the outcome as well as missing data handling should be taken into account.

In summary, the assumed reduction in residual variance due to a prognostic score may in principle be taken into account to reduce sample size, if it can be ensured that the calculation is conservative with respect to uncertainties in the assumptions made, and if the resulting sample size is large enough to meet other relevant purposes apart from the primary hypothesis test.

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?

Applicant's position

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 Applicant 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.

CHMP answer

In principle, CHMP agrees that implementing PROCOVA as prognostic score adjustment using a prognostic model derived from independent data and the proposed procedures could enable increases in power and/or decreases in sample size in phase 2 and 3 clinical trials with continuous outcomes. The presented mathematical properties, simulation exercises and empirical application support this use. Regarding choice of sample size, the answers to Questions 3 and 5 should be considered to safeguard that the selected sample size is suitable for the trial objectives.

Regarding the mathematical properties of PROCOVA, as implemented the method can be regarded a special case of ANCOVA sharing the properties of type I error control and asymptotically unbiased estimates of the treatment effect with sufficiently large sample sizes. For the weaker assumptions the Applicant uses the term 'technical' assumptions (Schuler et al., arXiv:2012.09935v2 2021), which may be debated. However, it can be agreed that similar assumptions are required for a large variety of parametric frequentist methods regularly applied and accepted from a regulatory perspective. Therefore, the proposed prognostic covariate procedure is an acceptable statistical approach.

The potential advantages of the PROCOVA procedure and prognostic score adjustment more broadly, depend on the availability of appropriate historical data and the derivation of a non-linear prognostic model that would allow outcome prediction in a future clinical trial. The number of covariates that can be included in the modelling approach is determined by the size and quality of the historical dataset. However, it is clear that type I error control, unbiased effect estimation and confidence interval coverage are not dependent on the choice or performance of the prognostic model. It is noted that prognostic score adjustment can be used together with adjustment using single covariates. The consequence of using the prognostic score together with additional prognostic covariates (one or more) needs to be carefully considered. The impact of the potential multicollinearity on the precision of the estimated coefficients may outweigh the proposed advantage of using PROCOVA and should thus be investigated in advance in order to inform the parameterisation to be used in the final primary analysis model (as

well as subgroup analyses). Using PROCOVA together with individual covariates for stratified randomisation was addressed in a second list of issues. Subgroup analyses based on covariates included in the prognostic score are addressed in an updated handbook for application of the PROCOVA method (see also the answer to Question 2). This includes subgroup analysis for covariates that could be predictive of treatment effect. If the treatment effect is expected to differ between subgroups due to predictive biomarkers as covariate (in contrast to a prognostic covariate) and precision of the treatment effect is especially important, additional power calculations are recommended to ensure sufficient power for subgroup analysis. Additionally, the need for pre-specification of the prognostic model may be a disadvantage in case of only a low number of covariates relevant for outcome prediction that could instead be included in an ANCOVA as single covariates with potential advantages in interpretation of results.

4. Background information as submitted by the Applicant

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

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.

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.

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].

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.

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.

Methodology and Results

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.

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 Applicant 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.

Mathematical Results

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 Applicants 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™.

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.

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™.

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.

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.

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.

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.

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).

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.