# Response to EMA Qualification Advice List of Issues Received 10 September 2021

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

20 September 2021

# SUMMARY

We appreciate a comprehensive review of our submission by the EMA, and an opportunity to engage in a scientific discussion scheduled for 27 September 2021. Here we provide an overview of the four main areas that comprise the majority of issues outlined in Qualification Advice List of Issues dated 10 September 2021, followed by our responses to individual issues.

1. Under what conditions does PROCOVA™ have the potential for an attainable advantage over ANCOVA with adjustment for a single covariate (or a limited number of covariates)? Conversely, when is PROCOVA™ not advantageous/should not be used?

PROCOVA™ has potential to attain an advantage over ANCOVA under two main scenarios. In the first scenario, multiple baseline covariates are known or suspected to be prognostically important, exceeding the number of covariates to be safely used with ANCOVA. The combination of machine learning and PROCOVA™ methodology overcomes this hurdle by integrating the multiple baseline covariates into a single prognostic covariate. In the second scenario, the prognostic value resides in the non-linear relationships between/among the baseline covariates. Again, machine learning methods used to generate the prognostic score to be used in PROCOVA™ provide an advantage over an ANCOVA analysis. These and other potential scenarios are discussed in greater detail in our response to Issue 3b.

Conversely, PROCOVA™ is not likely to attain an advantage over ANCOVA if all or most of the prognostic value is known to be associated with a small number of baseline covariates; if the relationship between these covariates and the outcome is known to be linear; and if the non-linear combinations of baseline covariates are known to have no predictive value. In addition, PROCOVA™ is not warranted if no prognostically useful information exists in the baseline covariates. Again, additional details can be found in our response to Issue 3b.

Finally, if the prognostic score cannot be robustly validated for the population of interest (i.e., the population matching the future trial enrollment criteria), then PROCOVA™ should not be used to reduce the sample size. The lack of robust validation of the prognostic score may result from a significant mismatch between the population used to train and validate the prognostic model (Step 1 of PROCOVA™ described in Section 3.1.1 and Appendix 1 of our submission), and the target population of the future trial, and when no other suitable population exists on which to validate the prognostic score. With a less robust validation, the use of PROCOVA™ to attain a higher level of power may still be warranted, even if the amount of power increase is uncertain (see below).

2. What is the procedure for the validation of the prognostic score, and what are the uncertainties?

Robust validation of the prognostic score is an integral part of Step 1 of PROCOVA™. Robust validation of the prognostic score is essential prior to using PROCOVA™ to prospectively reduce the sample size of a planned trial while maintaining the power. For applications where PROCOVA™ is used to add power, robust validation of the prognostic score reduces uncertainty about the amount of power gain. If the trial is well-powered without PROCOVA™, the size of the gain with PROCOVA™ may be less of a concern.

Robust validation means obtaining evidence showing the correlation between the prognostic score and the actual outcomes in a population/populations of interest, similar to that of the future trial (i.e., meeting the main patient selection criteria of that future trial). Once this correlation coefficient is estimated as a result of the prognostic score validation, the value can be incorporated into the sample size calculation using the methods described in Appendix 1 (Step 2) in our submission. The analytic properties of PROCOVA™ can be further reinforced by simulation studies for relevant parameters, like those presented and discussed in Section 3.3 of our submission. We discuss the validation of the prognostic score in greater detail in our responses to Issues 4 and 9.

As discussed above, robust validation of the prognostic score may not be possible if there is a significant mismatch between the population used to train and validate the prognostic model (Step 1 of PROCOVA™) and the target population of the future trial, and when no other suitable population exists on which to validate the prognostic score.

3. What is the procedure to avoid the use of optimistic assumptions in prospective sample size calculations?

As with traditional approaches to sample size calculations, even in the presence of a robustly validated prognostic score, a Sponsor may choose to be more conservative than simply assuming that the correlation coefficient between the prognostic score and the planned trial outcome will be identical to the correlation between the prognostic score and the outcomes in the validation dataset. To this end, we propose that inflation/deflation factors lambda and gamma can be used to provide a more conservative estimate of the value of the prognostic score, effectively lowering the correlation coefficient by some amount. For example, these parameters can be used to account for the expected sampling variability in populations, which is usually under 10% of the correlation coefficient. The choice of these parameters is further discussed in our responses to Issues 3 and 9. The relative benefit of PROCOVA™ versus a simpler covariate adjustment via traditional ANCOVA should also be considered, as discussed in our response to Issue 2.

4. How does PROCOVA™ influence/interact with a number of traditional elements of trial design and analysis, such as stratification, standard error calculation, subgroup analyses, imputation of missing covariates, multiple endpoint scenarios, and non-inferiority and equivalence designs?

We discuss these elements in our responses to Issues 3c and 4-8.

## EMA QUALIFICATION ADVICE ISSUE 1

**Please discuss available experience with prediction models in different therapeutic areas with regard to the ability to predict outcomes of future trials and attainable performance and correlations and their associated uncertainties.**

### SPONSOR RESPONSE

As described in our submission (Section 2.3 and Appendix 5), our methodology is applicable to any therapeutic area where historical data on the patient population in question are available, such that one can build a prognostic model to predict control outcomes (generate prognostic score) with sufficient accuracy given the subjects' measured baseline covariates.

In addition to the prognostic model in Alzheimer's Disease (AD) described in our submission, we have used our deep learning methodology to generate a prognostic model for Multiple Sclerosis[1], a complex neurodegenerative disorder with a highly varied and unpredictable progression course. We also have several prognostic models in development in other disease areas, including Amyotrophic Lateral Sclerosis, Huntington's Disease, Parkinson's Disease, Inflammatory Bowel Disease, and Immunology (Lupus).

Regardless of the disease area, the performance of a prognostic model and the ultimate gain in a particular application of PROCOVA™ may be directly affected by the nature of the historical data (e.g., frequency of assessments or evolution in the standard of care) and data quality (e.g., proportion of missing values). Most importantly, the model performance will depend on sufficient availability and access to fit-for-purpose historical data, which are needed to (1) train and validate the model, and (2) estimate the correlation between the prognostic score and the observed outcomes in the target population (the last step should use a dataset not used to train the predictive model and should possess characteristics similar to the target population). The implications and uncertainties associated with the evaluation of this correlation and the decision to apply PROCOVA™ are further discussed in the Summary above and in our response to Issues 3b and 4 below.

When considering the availability of adequate historical data, it is important to realize that over the last two decades, many placebo-controlled clinical trials conducted in a particular disease area have been likely collecting the same variables and utilizing many of the same approaches to evaluating efficacy and safety, based on the medical and regulatory importance of these variables and approaches for that disease area. Furthermore, while standards of care evolve, they rarely undergo changes so drastic that the prognostic model would become obsolete. Thus, if a sufficiently large database of placebo arm data exists in a particular disease area, there is a good chance it will be adequate for model training.

Finally, we anticipate that the ability to train highly accurate prognostic models will continue to grow due to ongoing improvements in deep learning methods and expansion of large databases of longitudinal patient data, including new, high dimensional biomarkers from technologies that provide large amounts of patient-level information.

## EMA QUALIFICATION ADVICE ISSUE 2

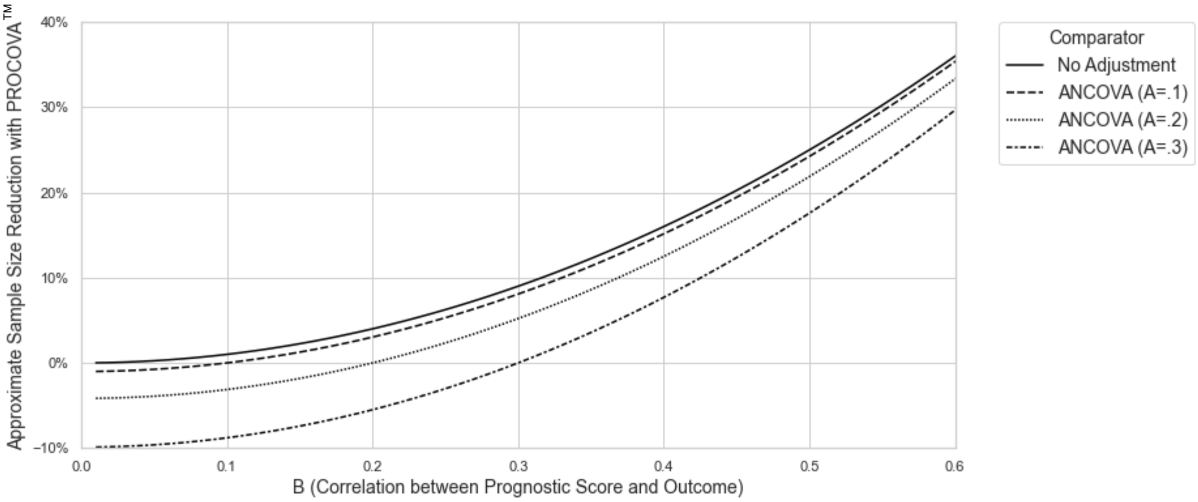
**Please discuss the option to implement a step in the PROCOVA™ procedure justifying an advantage over ANCOVA adjustment with single covariates.**

### SPONSOR RESPONSE

Our procedure, as described in Experiment 1 and Experiment 2 in Section 3.3.1, compares PROCOVA™ to a primary analysis with no covariate adjustment. This comparison was selected because sample size estimates routinely do not account for covariate adjustment even when traditional covariates are pre-specified in the primary analysis.

In order to assess if there is an attainable advantage of the PROCOVA™ procedure over ANCOVA with single covariates, we propose that Sponsors use the rules of thumb in Section 3.2.2. The approximate formula treats the single baseline covariate (which is expected to be correlated with the outcome) as a rudimentary prognostic score. This formula could be applied equally to a single baseline covariate and a prognostic score. For example, suppose that the correlation between the single baseline covariate and the outcome is A and the correlation between the prognostic score and the outcome is B. The attainable sample size reductions are approximately $1-A^2$ and $1-B^2$ respectively.

If a trial was planned to prospectively account for the power gains from using a single covariate, and one wished to consider using a prognostic score instead, the approximate sample size reduction would be (1-B$^2$)/(1-A$^2$). Consider a scenario where A=.2 and B=.4. In this scenario, the reduced sample size using PROCOVA™ versus a single covariate would be approximately .84/.96 = 87.5% (i.e., a 12.5% reduction in the sample size rather than the 16% reduction attainable compared to no adjustment). This is illustrated in the figure below, which includes a range of possible values of A and B.



It is important to note that PROCOVA™ results in a disadvantage compared to ANCOVA if A is greater than B, and therefore PROCOVA™ is not recommended under these conditions. We do not propose a specific lower bound in the sample size reduction, which must be achievable before implementing PROCOVA™; however, a given trial Sponsor may choose to only use PROCOVA™ when there is reason to believe that the sample size reduction can be, for example, at least 10%. Furthermore, a trial Sponsor applying that rule might only choose PROCOVA™ over no adjustment if the correlation between the prognostic score and the outcome was greater than 0.3 or choose PROCOVA™ over a strong single variable (e.g., one where the correlation was 0.2) if the correlation between the prognostic score and the outcome was greater than 0.35.

Step 2 of our procedure, as described in Section 3.1.1 and Appendix 1 of our submission, is "Accounting for the prognostic model while estimating the sample size required for a prospective study". To address the important considerations described above, Step 2 may be expanded into Steps 2a, 2b, and 2c as follows:

Step 2a. Estimate the effect of PROCOVA™ versus no adjustment, generally following the same approach as in the original Step 2 but paying greater attention to setting lambda (discussed further in response to Issue 3a and Issue 9).

Step 2b. Determine the effect on the power or sample size attainable by adjusting for a single covariate (or a limited number of covariates) to determine if ANCOVA is a viable alternative to no adjustment for the primary analysis.

Step 2c. Assess the relative pros and cons of using PROCOVA™ or ANCOVA, and make a final determination to choose one of the three paths: no adjustment, ANCOVA with one or more pre-specified covariates, or PROCOVA™.

## EMA QUALIFICATION ADVICE ISSUE 3

**The major difference of PROCOVA™ to the more conventional approach to address prognostic covariates is i) the method to evaluate the robustness of the sample size estimate, and ii) the inclusion of a single covariate using fixed weights to combine important baseline covariates. The Applicant is thus invited to:**

a) **clarify the choice of the de/inflation factors λ and γ, and compare the PROCOVA™ approach to the usual evaluation of the robustness of the sample size estimate with respect to deviations from assumptions made. Discuss options to avoid that sponsors would understand inclusion of the deflation and inflation parameters as exhaustive for covering uncertainties in assumptions**

### SPONSOR RESPONSE

This question pertains to the parameters lambda and gamma, which are described in our submission (Table 3 of Section 3.3.1 and also Step 2 in Appendix 1). The intent of lambda and gamma is to deflate or inflate estimates of the estimated correlation and standard deviation, respectively.

When developing sample size estimates and power calculations, statisticians will often use a conservative estimate for one or more of the key inputs to the calculation, either due to a lack of confidence in a particular estimate or a simple desire to be conservative. Our goal in including these parameters in our method was to be explicit and transparent for addressing uncertainties in model assumptions.

With respect to lambda in our case study, we selected a value that was close to 1 because our AD model was built from a very large dataset and it was validated both through in-sample cross validation and external datasets which gave us a high degree of confidence in our correlation estimates. However, there was a possibility of having an overly optimistic estimate of the correlation since the external datasets used in validation were limited. Therefore, we selected a lambda of 0.9 to allow for the possibility that the true correlation might be 10% less than the estimate from our data (to account for the expected sampling variability in populations). With respect to gamma, there is often a wealth of historical data about the standard deviation of a particular endpoint, so using 1.0 for gamma (i.e., no inflation) will often be a reasonable choice.

It must be noted that the selections of lambda and gamma were made in consultation with the Unlearn data scientists who developed and validated the AD disease progression model, from which the prognostic score was computed. The simplicity of the final step of the PROCOVA™ implementation may unintentionally obscure the complexity of the first step (model development). Similarly, we wish to underscore the importance of discussion in the second step, which we proposed breaking out into Steps 2a, 2b, and 2c (see response to Issue 2).

b) **discuss the impact of using a single covariate with fixed weights in case the relationship between outcome and baseline characteristic differs in the study setting as compared to the historical trials. It may be of interest to study the impact of such distributional inhomogeneities on PROCOVA™ in some further simulation experiments**

**SPONSOR RESPONSE**

An advantage of PROCOVA™ is that all of the prognostic value of the baseline data is captured in a single covariate. Consider a simple case where the prognostic score is a simple linear combination of 4 baseline values (e.g., prognostic score = 1*CovA + 2*CovB + 3*CovC + 4*CovD). Only one beta coefficient for the prognostic score is needed in PROCOVA™ versus 4 beta coefficients (one each for CovA, CovB, CovC, and CovD) in standard ANCOVA.

In this example, if the covariates are measured on a similar scale, the prognostic score equation suggests that CovD is substantially more predictive than CovA. If, in a future dataset, it is the opposite (i.e., CovA is substantially more predictive than CovD), a model, which fit new individual coefficients to CovA and CovD, would likely have lower variance than the model based on the prognostic score. That is, one can construct scenarios where an ANCOVA model, with multiple covariates, would outperform a PROCOVA™ model.

PROCOVA™ is likely to outperform ANCOVA in situations where the traditional ANCOVA approach cannot provide optimal predictions. Such conditions include: (1) multiple baseline covariates are known or suspected to be prognostically important; (2) prognostic variables have non-linear effects; (3) prognostic variables interact with one another; and (4) prognostic variables have the potential to be missing in the trial. In case (1) above, one has to pre-specify the limited set of covariates to use in ANCOVA. In case (2), one has to pre-specify which non-linear terms (if any) to include in ANCOVA. In case (3), one has to pre-specify which interactions (if any) to include, and with respect to case (4), one has to pre-specify an imputation plan for each individual covariate. While datasets may exist with none of these problems, most datasets have at least one of these issues, if not all of them. An advantage of PROCOVA™ is that these issues do not impair its performance.

A general difficulty with directly comparing PROCOVA™ to ANCOVA is that many choices are made in determining exactly how to apply ANCOVA, so one cannot directly compare PROCOVA™ to a generic ANCOVA model. One may compare PROCOVA™ to no adjustment and one may compare some specific implementation of ANCOVA to no adjustment, and then a rational decision can be made about the pros and cons of each method. A case study involving a direct comparison of PROCOVA™ to ANCOVA is provided in our response to part d) of this question.

c) **clarify whether the prognostic score is planned to be taken into account in the study design (e.g., using stratification) using PROCOVA™. Discuss options to implement stratification and specifically stratified randomisation (only) for well-understood prognostic factors.**

**SPONSOR RESPONSE**

Two common uses of individual baseline covariates are stratified randomization and covariate adjustment. Thus, it is natural to consider whether or not a prognostic score could have both uses.

We propose that PROCOVA™ is strictly a covariate adjustment technique and not a stratification technique. Because the prognostic score is explicitly a prediction for the expected outcomes in the control arm, there is no reason to believe it would be a predictive biomarker. Additionally, the prognostic score is derived from a potentially large set of variables that constitute the input to a prognostic score. For the prognostic score to be used at the time of randomization, it would have to be computed in real-time via a randomization hotline, which we do not believe would be practical.

d) **consider the Alzheimer study, discuss the major difference to the standard approach (where e.g. the baseline ADAS-Cog as well as other prognostic factors like the APOE4 status would be considered in design and analysis), and quantify in how far precision of estimates or power would be improved**

**SPONSOR RESPONSE**

As shown in our submission (Table 3 in Section 3.3.1), the 95% confidence interval for the ADAS-Cog11 treatment effect is reduced from ±2.03 to ±1.88 comparing the unadjusted result to the PROCOVA™ analysis using the deep learning model. These numbers correspond to heteroskedasticity-corrected variances of 1.07 and 0.92 respectively.

To address the Issue posed by the EMA, we performed an additional analysis. The variance of the treatment effect under a wider range of models is shown below.

| Covariate Adjustment | Variance of Treatment Effect Estimate for Change in ADAS-Cog11 |
|---|---:|
| None | 1.0715 |
| Baseline ADAS-Cog11 score | 1.0209 |
| APOE4 +/- | 1.0721 |
| Baseline ADAS-Cog11 and APOE4 +/- | 1.0219 |
| Prognostic Score (PROCOVA™) | 0.9192 |

The variance adjusting for baseline ADAS-Cog11 alone is 1.02. The variance adjusting for APOE status is almost the same as no adjustment (1.07), and the variance adjusting for both APOE status and baseline ADAS-Cog11 is 1.02. The lowest variance is attained with PROCOVA™ (0.92).

## EMA QUALIFICATION ADVICE ISSUE 4

**Please discuss implications in clinical trial settings with small sample sizes for standard error calculation and potential limitations of small historical data sets with regard to the number of covariates in a prediction model and uncertainties for correlation estimation. Consider the use of the t-distribution for statistical inference.**

**SPONSOR RESPONSE**

The issues related to small sample sizes can arise from two distinct sources, the size of the historical dataset used to build the model, and the size of the trial dataset to which the prognostic scores are applied. We address them in that order below.

Small sample sizes in the historical datasets may affect the construction and validation of the predictive model used to generate the prognostic score. A detailed discussion of potential methods for constructing predictive models which can generate prognostic scores for use with

PROCOVA™ is beyond the scope of the current application. Nevertheless, the size (and population coverage) of the historical datasets used for training the predictive model should be sufficient to enable the model to achieve out-of-sample predictive performance, i.e., in the same range as the in-sample predictive performance. For example, Riley et al[2] propose that the shrinkage factor should be >=.9.

The size of the required dataset will vary according to the number of covariates included in the model, its ability to capture nonlinear relationships, as well as the training procedure used (e.g., use of prior knowledge and/or regularization techniques).

Sample size estimation with PROCOVA™ requires an estimate of the correlation between the prognostic score and the observed outcomes in the target population. We propose to estimate this correlation from a held-out dataset, which was not used to train the predictive model, that has similar characteristics to the target population. Therefore, a key consideration is the availability of sufficient data for constructing such a validation dataset.

It is possible that one may not be able to achieve sufficiently large out-of-sample correlations (e.g., stronger correlations than individual known baseline covariates) if the training dataset is too small. In such a case, using PROCOVA™ would not provide efficiency gains beyond the basic analysis, but it would still preserve the type-I error rate control.

The size of the trial dataset is a separate issue. The theoretical sections of the application related to the power of PROCOVA™ analyses apply in the large sample setting in which the test statistic is approximately normally distributed. Therefore, we used the normal distribution for our p-value and power calculations as it is theoretically justified in this regime. We considered the use of Student’s t-distribution, but were unable to find a theoretical justification for it when using the Huber-White “sandwich” estimator for the standard errors (in contrast to the nominal estimators of the standard errors). However, as the p-value obtained from the Student’s t-distribution is always larger than the p-value obtained from the normal distribution, substituting the t-distribution with the appropriate degrees of freedom would be more conservative at smaller samples sizes. In cases in which control of the type-I error rate is particularly important (e.g., phase 3 clinical trials), typical sample sizes are large enough that differences between the normal and t-distributed statistics are minimal. Alternatively, one could estimate standard errors and p-values using a nonparametric bootstrap approach[3].

## EMA QUALIFICATION ADVICE ISSUE 5

**Please discuss options to implement subgroup analysis and a scenario with differential treatment effects (based on covariates) and possible approaches to characterise the treatment effect.**

### SPONSOR RESPONSE

Our submission focused on leveraging PROCOVA™ to attain either a reduction in the necessary sample size without affecting power, or an increase in power without increasing sample size. Because sample sizes are generally focused on the primary analysis of the entire study cohort, we did not discuss implications for subgroups.

In the case of improved power, without a sample size increase, one would expect the precision gains for the entire study cohort to exist for subgroups as well. Thus, all estimates would be more precise with PROCOVA™ than without PROCOVA™. Although studies are not typically powered to detect treatment effects in specific subgroups, trends would be observable with greater precision.

In the case of a reduced sample size with the same power, the precision of the treatment effect estimates in subgroups should generally be very similar to a larger study conducted without PROCOVA™. A notable exception is the case where a subgroup indicator has a very high correlation with the outcome. Here, the variance of the outcome within the subgroup will be lower than the overall variance and the corresponding variance reduction with PROCOVA™ might be slightly lessened. Generally, single variables, particularly categorical variables, do not have strong enough correlations with the outcome to have this effect. Nonetheless, if subgroup analysis is critically important (e.g., if there is a concern about differential treatment effects based on the biology of the disease or the trends in the data), this is a legitimate issue to be considered when deciding whether or not to use PROCOVA™ or any other covariate adjustment method.

## EMA QUALIFICATION ADVICE ISSUE 6

**Please discuss options for recommending imputation approaches for missing covariates.**

### SPONSOR RESPONSE

In our submission (Executive Summary), we note “a missing data imputation scheme should be pre-specified”, though we were not prescriptive about the best approach. The deep learning models used by Unlearn in general, and used in our case study, are built to produce a prognostic score regardless of how many of the model inputs are missing.

Regardless of the modeling approach, the method for calculating a prognostic score should be fully pre-specified for every patient. There should be no risk of excluding a patient from the analysis due to missing data nor should there be any risk of a post-hoc imputation scheme being necessary to avoid case-wise deletion.

It would also be permissible to have a fully pre-specified imputation scheme for any missing model inputs rather than having a model that is robust to missing inputs. The two obligatory requirements are that (1) the final prognostic score should not be missing for any patient in the ITT population, and (2) all imputation methods utilized should be fully pre-specified. Pre-specification can occur either through the model having a clear method for handling missing data, as in the case of Unlearn’s AD model, or, if a model requires non-missing inputs, each input must have a fully pre-specified imputation plan that is based on baseline data alone.

## EMA QUALIFICATION ADVICE ISSUE 7

**Please further discuss sample size estimation in a multiple endpoint scenario, covering co-primary endpoints and scenarios for which a precision of estimates would be targeted.**

### SPONSOR RESPONSE

The benefits of PROCOVA™ rely on the correlation between the prognostic score and the primary endpoint. In the case of prognostic scores built from disease progression models, there will be many endpoints in the same trial with corresponding prognostic scores. As shown in the case study included in our submission, the strength of those predictions will vary from one endpoint to another, so the size of the sample size decrease will depend on the specific endpoint selected.

In the case of co-primary endpoints, the endpoint with the weakest correlation to the corresponding prognostic score must be considered if PROCOVA™ is being used to reduce the sample size. If, on the other hand, PROCOVA™ is being used purely for increased power and precision, it is permissible for the benefits to vary between the two co-primary endpoints. The extent of the power gain will depend on the particular approach for handling multiple comparisons.

## EMA QUALIFICATION ADVICE ISSUE 8

**Please clarify the applicability of PROCOVA™ in non-inferiority and equivalence studies.**

### SPONSOR RESPONSE

All of the benefits of PROCOVA™ stem from reducing the variance of the treatment effect. Because an estimate of the variance of the treatment effect is essential to properly sizing and powering non-inferiority and equivalence studies, it could theoretically be used in that setting in addition to superiority studies.

Nonetheless, our simulations, case studies, and mathematical proofs were entirely focused on superiority studies. Therefore, we do not believe we have provided sufficient evidence to claim that PROCOVA™ is equally beneficial for non-inferiority studies or equivalence studies.

## EMA QUALIFICATION ADVICE ISSUE 9

**Please discuss prognostic model validation, and the problem of optimistic measures of predictive accuracy due to overfitting and how this will be addressed in practice (e.g. if extensive external validation is not feasible).**

### SPONSOR RESPONSE

The correlation between the observed trial outcome and the prognostic score is the single most important parameter for determining the expected benefits of PROCOVA™. This is illustrated by the rules of thumb provided in our submission (Section 3.2.2). Our approach, as illustrated by the parameters in Table 3 corresponding to our case study, involves a parameter for the estimated correlation as well as a deflation parameter, recognizing that the correlation may be prone to some level of optimism.

The question of optimism is directly related to Issue 3a, i.e., how to set lambda. Our technique is intended to be sufficiently general that it can be applied in cases where the prognostic score is highly predictive and extensively validated and also in cases where the prognostic score is only weakly predictive and less well validated.

First, consider a case where a model was developed and validated on the same data. We strongly recommend against this approach, but there certainly exist predictive models where model fitting occurred without separating model development and validation data. Some type of post-hoc cross-validation might be used in these cases to estimate the degree of optimism in the original estimate, but the best approach is likely to use a very conservative lambda (e.g., 0.5) or not to use PROCOVA™ at all.

Next, consider a case where development data and validation data have been carefully partitioned. In this case, a model can be locked prior to validation, and estimates of the correlation in the validation data should be reasonably accurate estimates for concurrent studies that are in the same population as the studies used to validate the model. In practice, the studies for which PROCOVA™ is intended will occur in the future, and therefore, it should be expected that the correlation may be somewhat weaker. Some level of optimism is likely because the future observations are effectively outside of the data space of the original estimates, even if calendar time is a relatively unimportant prognostic variable. Here, we might choose a lambda of 0.9, as we did in the case study included in our submission.

We may also consider a case where extensive validation has occurred in external datasets. In this case, the validation will have needed to occur prior to the design and finalization of the protocol in which PROCOVA™ will be used, so, even extensive external validation cannot guarantee accurate estimation for a trial which involves future patients. A lambda of 0.95 may be appropriate in this case.

Ultimately, there must be judgment calls about every estimate that feeds into the sample size calculation. This has always been true of any sample size calculation, though there are admittedly more parameters for which assumptions are needed using PROCOVA™. Some of these additional parameters, such as lambda, are intended to make the judgment calls more explicit, which we hope will engender more thoughtful discussion among stakeholders.

## REFERENCES

1. Walsh J, Smith, A, Pouliot Y, et al. Generating Digital Twins with Multiple Sclerosis Using Probabilistic Neural Networks. *bioRxiv preprint* doi: https://doi.org/10.1101/2020.02.04.934679, posted April 19, 2020.

2. Riley R, Snell K, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: Part I – Continuous outcomes. *Statistics in Medicine.* 2018:1-14.

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