



16 March 2022  
Case No.: EMA/SA/0000059571  
Committee for Medicinal Products for Human Use (CHMP)

## DRAFT 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 <sup>1</sup>
Start of public consultation	22 March 2022 <sup>2</sup>
End of consultation (deadline for comments)	03 May 2022 <sup>3</sup>

Comments should be provided using this [template](#). The completed comments form should be sent to [ScientificAdvice@ema.europa.eu](mailto:ScientificAdvice@ema.europa.eu)

<b>Keywords</b>	Qualification of Novel Methodology, Statistical methodology, Prognostic Covariate Adjustment, Sample size estimation
-----------------	--

---

<sup>1</sup> Last day of relevant Committee meeting.

<sup>2</sup> Date of publication on the EMA public website.

<sup>3</sup> Last day of the month concerned.



1 **Executive Summary**

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

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

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

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

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

31 We show that PROCOVA™ is optimal if the prognostic model attains the maximal possible correlation  
32 with the actual outcomes of subjects under control conditions. However, one can realize gains in  
33 efficiency even with imperfect prognostic models. The other important advantage of PROCOVA™ is that  
34 it involves an adjustment for a single covariate derived from a larger set of variables that constitute  
35 the input of a prognostic model, providing a substantial dimensionality reduction. Even if the input to  
36 the prognostic model is high-dimensional in nature (e.g., a brain image, or a whole transcriptome),  
37 PROCOVA™ still represents an adjustment for a single covariate. One only has to measure the Pearson  
38 correlation of this single covariate with the actual outcome in a similar historical population in order to  
39 account for the prognostic score in a prospective sample size estimation for a planned trial. We present  
40 mathematical proof and an actual demonstration of a prospective application of PROCOVA™ to power a  
41 trial without estimating or assuming a large number of population parameters.

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

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

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

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

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

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

## 61 **Statement of the Need for and Impact of the Proposed Novel Methodologies in Clinical Drug** 62 **Development**

### 63 ***Background***

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

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

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

82 Various approaches to historical borrowing have been proposed and their properties extensively  
83 evaluated, ranging from directly inserting subjects from previous studies into the current sample, to  
84 using previous studies to derive prior distributions for Bayesian analyses. Although such methods do  
85 generally increase power, they cannot strictly control the type-I error rate reducing the relevance of

86 such methods, particularly for pivotal/ confirmatory/ Phase 3 RCTs. A common approach to addressing  
87 the risk of type-I error rate inflation when information is borrowed is to carry out multiple simulation  
88 studies to quantify this effect.

### 89 ***The Novel Methodology***

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

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

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

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

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

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

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

124 A number of recent technological developments have led to substantial improvements in the ability to  
125 train highly accurate prognostic models. First, large databases of longitudinal patient data from control  
126 arms of historical clinical trials, observational and natural history studies, and real-world sources have  
127 become widely available. Second, high dimensional biomarkers from technologies such as imaging and  
128 next generation sequencing provide large amounts of patient-level information. And, third,  
129 improvements in machine learning methods (especially in the subfield known as deep learning) allow  
130 one to create prognostic models that can fully utilize all of these patient data. The intersection of these

131 three key developments — large, analysable databases containing high-dimensional outcomes, and  
132 powerful deep learning models — allows for the generation of more predictive prognostic scores,  
133 adjusting for which can substantially reduce variance/confidence intervals, and/or increase power and  
134 reduce minimum required sample sizes.

### 135 **Objective, Scope and Context-of-use**

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

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

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

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

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

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

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

175 While our methodology is applicable to in-scope trials in any therapeutic area where historical control  
176 data are available, we have chosen Alzheimer’s Disease (AD) as our primary initial target because of  
177 an exceptionally high, and growing, unmet need; challenging, long and large Phase 2/3 trials;  
178 abundant placebo control data from over 150 randomized clinical trials and many observational studies  
179 conducted since the 1990’s; and largely unchanged SOC and the clinical trial endpoints for  
180 symptomatic AD over the last 17 years (ensuring small or no temporal drifts in the data). As such, the  
181 predictive models described in the simulations and empirical examples/context-of-use parts of this  
182 submission were constructed on historical data from AD trials contained in the Alzheimer’s Disease  
183 Neuroimaging Initiative (ADNI) database and the Critical Path for Alzheimer’s Disease (CPAD)  
184 database). Our empirical context-of-use demonstrations involve re-analysis of a Phase 3 trial in  
185 patients with AD reported by Quinn et al.

186 **Background information as submitted by the Applicant**<sup>i</sup>

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

189 **Background**

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

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

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

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

215 **The Novel Methodology**

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



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

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

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

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

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

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

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

## 262 **Objective, Scope and Context-of-use**

263 The objective of this submission is to seek CHMP qualification for the proposed statistical methodology  
264 intended to improve the efficiency of Phase 2 and 3 clinical trials by using trial subjects’ predicted

265 control outcomes (prognostic scores) in linear covariate adjustment (PROCOVA™); such prognostic  
266 scores can be generated from each subject's baseline characteristics using a predictive model trained  
267 on historical data. Our approach is efficient in the sense that it uses historical data to reduce variance  
268 of the treatment response estimates (and thus the minimum sample size required to achieve the  
269 desired level of confidence) better than other methods with access to the same baseline covariates.

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

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

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

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

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

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

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



311 **Alzheimer's** Disease (CPAD) database ([Appendix 5](#)). Our empirical context-of-use demonstrations  
312 involve re-analysis of a Phase 3 trial in patients with AD reported by Quinn et al. <sup>24</sup>.

### 313 **Out-of-Scope/Future Directions**

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

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

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

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

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

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

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

### 348 **Preview of the Technical Aspects Detailed in Methods and Results**

349 In the next section, we provide a detailed description of PROCOVA™ and present mathematical proofs  
350 of its main statistical properties (Section 3.1). Specifically, we prove that estimates of treatment  
351 effects obtained with PROCOVA™ are unbiased and that type-I error rates of hypothesis tests are  
352 controlled at the pre-specified level. These results hold for PROCOVA™ use with any prognostic model.  
353 In addition, we prove that PROCOVA™ can attain the maximum power of any estimator with access to  
354 the pre-specified baseline covariates if the prognostic model is exact — that is, PROCOVA™ is the  
355 optimal estimation procedure if the computed prognostic score for a subject is equal to his/her actual

356 expected outcome under control conditions, given his/her baseline characteristics. In addition, we  
357 provide a simple formula to estimate the power/minimum sample size in a prospective trial that will be  
358 analyzed with PROCOVA™.

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

370 Next, we present an empirical demonstration of PROCOVA™ through re-analyses of a completed Phase  
371 3 trial in patients with AD, in order to illustrate different benefits of PROCOVA™ (Section 3.3). The first  
372 experiment demonstrates that, using the same sample size and randomization ratio as in the original  
373 study, adjusting for prognostic scores decreases the magnitude of the estimated standard errors and  
374 the width of the confidence intervals. The second experiment demonstrates that accounting for the  
375 prognostic scores during sample size estimation results in a trial with fewer subjects but with standard  
376 errors of equal magnitude to those in a larger trial designed without PROCOVA™.

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

## 381 **Methodology and Results**

### 382 **The Prognostic Covariate Adjustment (PROCOVA™) Method**

383 Here we describe in detail the steps for using PROCOVA™ to estimate the treatment effect in an RCT  
384 and to perform a sample size calculation. We present the mathematical properties of the proposed  
385 procedure in a series of theorems, with mathematical proofs and technical details provided in [Appendix](#)  
386 [1](#), [Appendix 2](#), and [Appendix 3](#).

#### 387 **Description of PROCOVA™**

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

#### 390 **Step 1: Training and evaluating a prognostic model to predict control outcomes/generate** 391 **prognostic scores.**

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

396 In principle, there are many ways to obtain a prognostic model. The type-I error rate will be controlled  
397 for any type of model, whereas the realized increase in trial efficiency will depend on the predictive  
398 performance of the model in the target population, defined here and below as subjects meeting the  
399 selection criteria in the trial of interest. Machine learning-based methods are especially effective in

400 fitting the model to a collection of historical data and linking subjects' baseline covariates to their  
401 outcomes under the control condition. We provide two examples of this type of prognostic model in our  
402 empirical analyses.

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

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

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

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

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

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

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

442 The PROCOVA™ method described above is a special case of Analysis of Covariance (ANCOVA) with a  
443 particular choice of adjustment covariate. As such, PROCOVA™ inherits the statistical properties of  
444 ANCOVA; for example, estimated treatment effects will be unbiased and the type-I error rate will be  
445 controlled. For these reasons, ANCOVA is widely used in the analysis of clinical trials with continuous

446 responses and is supported by guidance from EMA <sup>13</sup> and draft guidance from FDA <sup>14</sup>. These statistical  
447 properties hold for PROCOVA™ using any prognostic model, regardless of the approach to modeling or  
448 the data used to inform the model.

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

## 462 **Mathematical Results**

### 463 **Mathematical Properties of ANCOVA**

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

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

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

477 The following mathematical theorems establish statistical properties of ANCOVA and, as a result, of  
478 PROCOVA™. Here, we only present descriptions and implications of the mathematical theorems,  
479 leaving rigorous proofs and results to [Appendix 2](#).

#### 480 **Theorem 1:**

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

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

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

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

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

499 Use of ANCOVA is facilitated by the fact that the resulting estimates of treatment effects are unbiased,  
500 and type-I error rates of hypothesis tests are controlled. In addition, using ANCOVA always increases  
501 power in randomized trials with equal randomization. Therefore, we propose to choose the  
502 transformation that maximizes statistical power, which is PROCOVA™.

### 503 **Mathematical Properties of PROCOVA™**

504 PROCOVA™ is motivated by the theorem presented below, with detailed results provided in [Appendix 2](#)  
505 and [Appendix 3](#).

#### 506 **Theorem 2:**

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

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

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

### 529 **Simulation Studies of PROCOVA™**

530 We demonstrate that PROCOVA™ provides more precise estimates of treatment effects than  
531 unadjusted estimators in realistic simulated scenarios. By using simulations, we are able to specify the  
532 data generating distribution and treatment effect. Since the treatment effect is known, the discrepancy  
533 between the estimated and actual treatment effects can be directly measured. Specifically, we used

534 simulation studies to explore how mean-squared estimation error of the treatment effect varies with  
535 and without PROCOVA™.

## 536 **Simulation Study Methods**

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

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

541 The Non-linear simulation describes a scenario in which the outcome-covariate relationship is non-  
542 linear in both treatment arms, but the treatment effect is constant.

543 The Heterogeneous simulation describes a scenario in which the conditional average effect  
544  $E[Y_1 - Y_0|X] = \mu_1(X) - \mu_0(X)$  is not constant (i.e.,  $E[Y_1 - Y_0|X] \neq \mu_1(X) - \mu_0(X)$ ).

545 The Shifted simulation describes a scenario in which the historical population used to train the  
546 prognostic model is not representative of the trial population in terms of the baseline  
547 covariates (i.e.,  $P_H(X' = x) \neq P(X = x)$ ).

548 Details on the data generating process for each of the simulation scenarios are provided in [Appendix 4](#).

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

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

565 We analyzed the data using three estimation procedures: unadjusted, adjusted with the estimated  
566 prognostic score obtained with the random forest, and adjusted with the exact prognostic score. The  
567 three estimation procedures were repeated for models with and without additional baseline covariates  
568 included. Finally, we calculated the squared-error of each estimate relative to the true treatment  
569 effect, which is known because it was used to generate the simulated data, repeated this process  
570 10,000 times, and averaged the squared-errors to obtain mean-squared errors for each analysis.

## 571 **Simulation Study Results**

572 Table 1 and Table 2 present the results obtained in each of the 4 chosen scenarios, including Linear  
573 and Non-linear outcome-covariate relationships, both of which can be learned by the random forest  
574 prognostic model, and the Heterogeneous and Shifted scenarios, which probe the sensitivity of  
575 PROCOVA™ to the violation of the Theorem 2 assumption regarding constant treatment effect, and to  
576 the accuracy of the prognostic model, respectively. The two tables differ in that Table 1 does not  
577 include any additional covariates besides the prognostic score, while Table 2 includes additional



578 baseline covariates. The Table lists the mean-squared errors of estimated treatment effects obtained  
 579 in unadjusted analysis; analysis using adjustment for an estimated prognostic score: and analysis  
 580 using adjustment for an exact prognostic generated by a “perfect” prognostic model as described  
 581 above.

582 **Table 1. Mean-squared errors of estimated treatment effects computed from**  
 583 **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

584 **Table 2. Mean-squared errors of estimated treatment effects computed from**  
 585 **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

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

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

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

$$\frac{\text{Variance with PROCOVA}}{\text{Variance without PROCOVA}} \sim 1 - R^2$$
$$\frac{\text{Power with PROCOVA}}{\text{Power without PROCOVA}} \sim 1 + (R^2/2)$$
$$\frac{\text{Minimum sample size with PROCOVA}}{\text{Minimum sample size without PROCOVA}} \sim 1 - R^2$$

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

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

## 618 **Empirical Applications of PROCOVA™**

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

## 628 **Empirical Analyses Methods**

629 We obtained a set of historical controls by combining data from the Alzheimer's Disease Neuroimaging  
630 Initiative (ADNI)<sup>28</sup> and the Critical Path for Alzheimer's Disease (CPAD)<sup>29,30</sup>. The combined dataset was  
631 composed of data from 6,919 subjects with early-stage Alzheimer's Disease. Importantly, the historical  
632 dataset did not contain data from the demonstration trial. Two different prognostic models were  
633 trained to predict control potential outcomes using the ADNI and CPAD datasets: a random forest<sup>31</sup>,  
634 and a deep learning model<sup>18,32</sup>. For our demonstration, we focused on the 18-month changes in the  
635 Alzheimer's Disease Assessment Scale - Cognitive Subscale (ADAS-Cog11)<sup>33</sup> and the Clinical Dementia  
636 Rating (CDR)<sup>34</sup>. More details on the training data and the prognostic models are provided in [Appendix](#)  
637 [5](#) and [Appendix 6](#).

638 The demonstration trial was originally performed through the Alzheimer's Disease Cooperative Study  
639 (ADCS), a consortium of academic medical centers and private Alzheimer disease clinics funded by the  
640 National Institute on Aging to conduct clinical trials on Alzheimer disease. In this trial, 238 subjects  
641 were randomized to the active treatment arm, and 164 subjects were randomized to placebo. The trial  
642 measured multiple covariates at baseline including demographics and patient characteristics (e.g., sex,  
643 age, region, weight), lab tests (e.g., blood pressure, ApoE4 status<sup>35(p4),36(p4)</sup>), and component scores of  
644 cognitive tests. More details are provided in [Appendix 5](#).

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

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

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

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

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

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

Parameter	Value
Significance level ( $\alpha$ )	5%
Desired power ( $\zeta$ )	80%
Proportion of subjects randomized to treatment arm ( $\pi$ )	3/5
Target treatment effect ( $\beta_1^*$ )	3.1
Expected dropout ( $d$ )	0.3
Estimated standard deviation ( $\hat{\sigma}_0$ )	9.1
Inflation parameter for standard deviation in the control arm ( $\gamma_0$ )	1.0
Inflation parameter for standard deviation in the active treatment arm ( $\gamma_1$ )	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 ( $\lambda_0$ )	0.9
Deflation parameter for prognostic correlation in the active treatment arm ( $\lambda_1$ )	0.9

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

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

## 680 Empirical Analyses Results

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

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

688 **Table 4. Reanalysis of the Quinn et al. trial at 18 months using two different**  
689 **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

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

694 While the point estimates for the treatment effects were modified to some extent when prognostic  
695 score adjustment was applied, the changes were minimal relative to the size of the estimated standard  
696 errors. Adjusting for baseline covariates or a prognostic score does not add bias [12,37,38](#), even though  
697 the point estimates for individual endpoints may change. That is, differences in point estimates  
698 between adjusted and unadjusted analyses are random, and do not persist in expectation. The original  
699 analysis of this particular trial<sup>24</sup> did not demonstrate statistically significant improvements on any of  
700 the endpoints of interest, and nor did any of our re-analyses.

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

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

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

713 **Table 5. Re-analysis of the Quinn et al. study using different sample sizes that account**  
714 **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

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

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

## 729 **Conclusions**

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

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

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

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

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

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



776 **Questions on Statistical Properties of PROCOVA from the Applicant**

777 **Question 1**

778 **Does the EMA agree that PROCOVA™ produces unbiased treatment effect estimates and**  
779 **controls the type-I error rate, given that:**

- 780 **a. PROCOVA™ is a special case of ANCOVA in which the covariate used for adjustment is a**  
781 **prognostic score, computed from data collected at or before baseline using a pre-**  
782 **specified prognostic model;**
- 783 **b. ANCOVA can decrease the variance of the estimated treatment effect if the adjustment**  
784 **covariate is correlated with the response;**
- 785 **c. Using ANCOVA to adjust for a covariate produces unbiased treatment effect estimates**  
786 **and controls the type-I error rate, as long as the covariate is computed from data**  
787 **collected at or before baseline.**

788 ***Applicant's position***

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

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

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

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

810 **CHMP answer**

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

813 Step 1: Training and evaluating a prognostic model to predict outcomes under the control  
814 condition (generate prognostic score).

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

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

819 The key idea is to first develop a prognostic score for the outcome based on a historical data set that is  
 820 independent from the study data and then apply the prognostic score as covariate in an ANCOVA  
 821 model for the actual data analysis.

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

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

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

	<b>Standard approach</b>	<b>PROCOVA</b>
<i>Design Stage</i>	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
<i>Sample Size considerations</i>	Sample size is estimated based on $\alpha$ , $\beta$ , difference to be detected and variability based on historical studies  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	Sample size is estimated based on $\alpha$ , $\beta$ , difference to be detected and variability, as well as $\rho$ (correlation coefficient between prognostic index and outcome) based on historical studies  Uncertainty in variability and prognostic ability is accounted for (using parameters $\lambda$ and $\gamma$ )
<i>Analysis</i>	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

832

833 Overall, there are two major differences between the conventional approach and PROCOVA:  
834 - the method to evaluate the robustness of the sample size estimate, which will be addressed in the  
835 answer to Questions 3 and 5

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

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

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

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

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

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

## 866 **Question 2**

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

### 870 ***Applicant's position***

871 Theorem 2 proves that a prognostic score, i.e., the prediction of a potential outcome under  
872 control conditions for all trial participants conditioned on their observed baseline covariates, is  
873 the optimal covariate to adjust for in ANCOVA. Theorem 2 is presented and further discussed in  
874 Section 3.1.2.2, Appendix 2 and Appendix 3.

875 Our simulation results described in Section 3.2 and specifically in Table 1 and Table 2, as well as  
876 in Appendix 4, demonstrate that the higher the correlation between the prognostic score and the

877 observed control outcomes, the greater the reduction in the variance of treatment effect  
878 estimates. This finding held when PROCOVA™ was applied alone (Table 1) or combined with  
879 adjustment for baseline covariates (Table 2).

880 Additional evidence is provided by our empirical demonstration presented in Section 3.3, with further  
881 technical details included in Appendix 5, Appendix 6, and Appendix 7. Specifically, the results in Table  
882 4 and Table 5 show that greater reductions in variance can be achieved when the prognostic score is  
883 more highly correlated with the observed outcome.

#### 884 **CHMP answer**

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

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

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

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

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

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

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

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

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

### 948 **Question 3**

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

### 952 ***Applicant's position***

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

### 957 **CHMP answer**

958 It can be agreed that applying adjustment for the PROCOVA prognostic score or a set of covariates for  
959 ANCOVA in general could lead to a smaller minimum sample size to achieve a desired level of power.  
960 As outlined by the Applicant, the minimum sample size is a function of the Pearson correlation  
961 coefficient between observations and predictions of the prognostic model. During sample size planning  
962 an investigator may take into account explained variation due to covariates, such as the prognostic  
963 score in PROCOVA, which will result in smaller sample size than assuming an unadjusted analysis or  
964 zero correlation between covariates and outcome. However, overly optimistic assumptions on the  
965 effect of covariates may result in too low sample sizes and inconclusive studies. It is noted that the  
966 Applicant recommends using a separate data set independent from the training data to estimate the

967 correlation coefficient and thus avoid overestimation of the correlation; this is supported. Please see  
968 the answer to Question 5 for further considerations and more detailed comments regarding sample size  
969 planning.

## 970 **Questions on the Context-of-Use**

### 971 **Question 4**

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

974 **a. PROCOVA™ is a special case of ANCOVA;**

975 **b. ANCOVA is an acceptable statistical method to estimate treatment effects in phase 2 and**  
976 **3 clinical trials with continuous responses under current regulatory guidance.**

### 977 ***Applicant's position***

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

985 Our overview of PROCOVA™ (Section 3.1.1) and our mathematical results (Section 3.1.2)  
986 establish that PROCOVA™ is a special case of ANCOVA with a particular choice of adjustment  
987 covariate. As such, PROCOVA™ inherits the statistical properties of ANCOVA described above,  
988 and these statistical properties hold for PROCOVA™ when used in conjunction with any  
989 prognostic model, regardless of the approach to modeling or the data used to inform the  
990 model. Therefore, PROCOVA™ is also acceptable and should be recommended for use to  
991 estimate treatment effects in pre-specified analyses of pivotal/registration trials.

### 992 **CHMP answer**

993 As outlined in the answer to question 1, CHMP agrees that the proposed method is a special case of  
994 ANCOVA. Therefore, similar to other ANCOVA models adjusted for a prognostic score, the proposed  
995 method will be acceptable to estimate the treatment effect and perform statistical inference on it in  
996 randomized trials. The proposed PROCOVA procedure can be considered an acceptable formal  
997 presentation of approaches that were used in clinical trial settings before when prognostic covariates  
998 were included in analysis models, e.g. by imaging based risk scores in oncology or covariate based risk  
999 scores in cardiovascular diseases.

1000 Regarding use of linear models for estimation, it is noted that from a regulatory perspective for a  
1001 primary estimand and analysis, application of a linear ANCOVA model with covariate adjustment would  
1002 be acceptable even if the linear model does not model the relationship between treatment, covariates  
1003 and outcomes correctly if an average treatment effect for a population-level summary is targeted. It is  
1004 though acknowledged that an improved modelling of the true relationship between treatment, (a larger  
1005 set of) covariates and outcome can be beneficial and can improve the precision of the estimator and  
1006 could potentially also allow better understanding of conditional treatment effects if relevant in a  
1007 particular disease setting.

1008 The Applicant proposes to perform statistical inference on the treatment effect using large sample  
1009 normal approximations to the respective test statistic. While this approach is asymptotically valid, it  
1010 neglects the variability of the estimate for the residual variance nuisance parameter. It is therefore



1011 recommended to use t-distributions (which take into account this variability under the assumption of  
1012 normally distributed residuals) to avoid too liberal test decisions. This is particularly emphasized as the  
1013 sample a size may be small in phase II, and even phase III studies. The Applicant agreed during the  
1014 discussion meeting that using the t-distribution is a reasonable, conservative approach for trials with  
1015 smaller sample sizes.

1016 The Applicant further proposes to use robust "sandwich" variance estimation in inferential procedures.  
1017 This is acceptable, however certain properties of the robust variance estimator need to be taken into  
1018 account: Using a bias-adjusted estimator is required as the small sample bias of the unadjusted robust  
1019 variance estimator may be considerable. The bias adjustment proposed by the Applicant is acceptable.  
1020 The robust estimator has larger variability than the model-based estimator. Hence it may not be  
1021 suitable with small sample sizes. In any case, hypothesis tests and confidence intervals should be  
1022 based on t-distributions as discussed above. In the discussion meeting, the Applicant pointed out that  
1023 there is no definite way for choosing the degrees of freedom for a reference t-distribution when using  
1024 robust variance estimation. This is acknowledged, however using an approximate number of degrees of  
1025 freedom is considered acceptable. E.g., the work by Lipsitz, Ebrahim and Parzen 1999 on a respective  
1026 Satterthwaite approximation may be considered (Lipsitz, S. R., Ibrahim, J. G., & Parzen, M. (1999). A  
1027 degrees-of-freedom approximation for a t-statistic with heterogeneous variance. Journal of the Royal  
1028 Statistical Society: Series D (The Statistician), 48(4), 495-506).

1029 The following further specific concerns may need to be addressed in an actual application:

1030 1) Since the prognostic score is trained under control conditions, it is possible that its correlation to the  
1031 outcome is larger under control than under treatment. This could result in unequal residual variance in  
1032 the two groups, which may lead to inflation of the type I error rate in trials with unequal group sizes.  
1033 The robust variance estimation as proposed by the Applicant is an acceptable remedy of this issue.

1034 2) A score that includes complex transformations of the considered variables may be prone to result in  
1035 skewed distribution with some outliers, even if the included variables have unsuspecting distributions at  
1036 their original scale. Outliers in the prognostic score may turn out to be influential points in fitting the  
1037 analysis model, which may raise concerns regarding the robustness of results. It is recommended that  
1038 the PROCOVA analysis should be supported by appropriate model diagnostics to assess the robustness  
1039 of the analysis results with respect to deviations in single observations.

1040 3) The Applicant claims that with recent methodological developments a prognostic score with  
1041 considerable correlation can be obtained for a variety of continuous responses in multiple therapeutic  
1042 areas. Correlation values around 0.4 are considered in the empirical examples and values up to 0.8 are  
1043 considered in the theoretical sections. Considering the conventional approach, a strong prognostic  
1044 index with a correlation of such a magnitude would usually be accounted for in the study planning, e.g.  
1045 using stratified randomisation. The Applicant clarified during the discussion meeting that the prognostic  
1046 score to be used in the PROCOVA analysis is not intended to be used for stratification. As the  
1047 prognostic score is derived from a potentially large set of variables, it is not considered practical to be  
1048 implemented in the randomization procedure. This aspect was further addressed in a second list of  
1049 issues, and the updated handbook developed by the Applicant instructs trial statisticians to consider (a  
1050 limited number of) the strongest prognostic factors for stratified randomization taking into account  
1051 that (some of) these candidate stratification factors could already be included in the prognostic score.

1052 4) It is expected that data on all variables included in the prognostic score will be collected in the  
1053 randomised trial. Concerning incomplete data on covariates for prognostic score adjustment, there are  
1054 be several options and a missing data imputation scheme should be pre-specified. Missing data was  
1055 further addressed in the second list of issues. Additional instructions were provided for situations  
1056 where significant differences in data completeness exist between the new trial and the validation  
1057 dataset. The correlation coefficient R may be lower in a new trial if one or more important variables are

1058 expected to be missing frequently (or with a different pattern of missingness). While the prediction  
1059 model would be able to generate prognostic scores for all subjects, regardless of missing data, the  
1060 advantage of PROCOVA may be decreased. Generally, if the proportion of missing data is low and  
1061 imputation is considered, multiple imputation could be preferable and imputations should not depend  
1062 on data of post-baseline measurements in the target trial. It is acknowledged that baseline covariates  
1063 cannot be impacted by intercurrent events.

1064 5) While it is understood that the prognostic score adjustment targets an average treatment effect for  
1065 a trial population, subgroup analysis based on covariates could be relevant for characterisation of the  
1066 treatment effect. This would be of particular relevance in case of (expected) differential treatment  
1067 effects. The Applicant provided further instructions on how such situations should be addressed at the  
1068 design and analysis stage when using PROCOVA. Please refer to the answer to Question 2.

#### 1069 **Question 5**

1070 **Does the EMA agree that it is acceptable to account for the adjustment of the prognostic**  
1071 **score using PROCOVA™ during sample size estimation for a phase 2 and 3 clinical trials with**  
1072 **continuous responses?**

#### 1073 ***Applicant's position***

1074 We have provided three lines of evidence demonstrating that the use of PROCOVA™ can reduce  
1075 variance of the treatment effect estimates: mathematical results (Section 3.1.2), simulations  
1076 (Section 3.2 and specifically Table 1 and Table 2) and empirical examples (Section 3.3 –  
1077 Experiment 2 and Table 4).

1078 In addition, we have shown that the same power can be delivered with a smaller sample size and  
1079 lower variance (reduced via application of PROCOVA™), as with a larger sample size and higher  
1080 variance. This was established in our simulations described in Section 3.2 and in empirical  
1081 demonstration presented in Section 3.3 (see Experiment 2) and Table 5.

1082 The technical details for our mathematical results are provided in Appendix 2 and Appendix 3; for our  
1083 simulations – in Appendix 4, for empirical demonstrations – in Appendix 5 and Appendix 6, and for  
1084 sample size estimation – in Appendix 7.

#### 1085 **CHMP answer**

1086 As stated in the answer to Question 3, it is agreed that taking into account explained variation due to  
1087 covariates, such as the prognostic score in PROCOVA, results in reduced residual variance and hence  
1088 will result in smaller sample size than assuming an unadjusted analysis.

1089 Nonetheless, for such a planning approach to be acceptable potential uncertainties in the assumption  
1090 on the variance explained by the prognostic score need to be taken into account. Overly optimistic  
1091 assumptions on the effect of covariates may result in too low sample sizes and inconclusive studies.  
1092 Most trials are planned conservatively without taking into account possible gains in power due to  
1093 adjusting for covariates and the actual power may then be larger than the planning assumption of, e.g.  
1094 80% or 90%. Also in usual sample size planning, different assumptions regarding the variance and  
1095 other relevant parameters are explored to assess the impact of deviations from the made assumptions  
1096 on the resulting power.

1097 As a first step, an attainable advantage over using ANCOVA with single covariate adjustment should be  
1098 justified. The Applicant demonstrates that this should be the case if the prognostic score is able to  
1099 capture a nonlinear relationship between covariates and outcomes of interest. This is discussed in  
1100 Schuler et al. (Schuler et al., arXiv:2012.09935v2 2021), and there would be no gain in efficiency  
1101 when adjusting with a prognostic score assuming a linear relationship between covariates and

1102 outcome. During the discussion meeting, the Applicant further elaborated on the attainable advantage  
1103 of the PROCOVA procedure over ANCOVA with single covariates. The potential sample size reductions  
1104 using PROCOVA depend on the ratio of the correlation between a single baseline covariate (or a linear  
1105 combination of the covariates that would typically be considered in the analysis) and the outcome and  
1106 the correlation between the single prognostic (PROCOVA) score and the outcome. The gain in sample  
1107 size (or likewise in power or precision of the estimates) can then be evaluated (graphically) and should  
1108 also take the optimism due to prognostic model building into account. The relative pros and cons of  
1109 using PROCOVA or ANCOVA are compared to make a final determination to choose one of the three  
1110 paths: no adjustment, ANCOVA with one or more pre-specified covariates, or PROCOVA. This issue was  
1111 raised in a second list of issues and was addressed by the Applicant in a handbook for trial statisticians  
1112 guiding the application of PROCOVA. The handbook provides guidance to help the trial statistician  
1113 make an informed choice among the three paths with step-by-step instructions.

1114 In the original procedure described by the Applicant, an inflation parameter ( $\gamma$ ) for standard deviation  
1115 in the control arm, as well as a deflation parameter ( $\lambda$ ) for prognostic correlation in both arms need to  
1116 be selected. The latter has been set to  $\lambda=0.9$  in the analysis of the Alzheimer data set. A clear  
1117 rationale for that choice was not provided. In an actual application, it needs to be carefully considered  
1118 how  $\lambda$  and  $\gamma$  are chosen. Evaluation of the robustness of the sample size or power estimate with  
1119 respect to deviations from assumptions, as outlined above, seems generally more informative than  
1120 relying on the two modifying parameters. At the discussion meeting and in the written responses to  
1121 CHMP's first list of issues, the Applicant outlined rules of thumb for the choice of the deflation factor  $\lambda$   
1122 for the correlation coefficient. The choice is proposed to depend on the extent of model validation. The  
1123 value may be close to 1 if there was extensive validation using external data sets, it may be chosen  
1124 conservatively (e.g.  $\lambda=0.5$ ) if the model was developed and validated on the same data set, or it may  
1125 be decided to not use PROCOVA at all. It was considered important by SAWP to provide the practitioner  
1126 with such rules of thumb but also to advise conduct of sensitivity analyses to prevent under-powered  
1127 trials. The updated handbook provides guidance for the choice of the deflation factor  $\lambda$ , and for the  
1128 conduct of sensitivity analyses taking into account a potential over-optimism of the prognostic model  
1129 and the fact that the correlation of the prognostic score with the outcome may be smaller under  
1130 experimental treatment. It should still be kept in mind that the approach using  $\lambda$  and  $\gamma$  may not cover  
1131 the range of all parameters relevant for assessing the robustness of the sample size and should not be  
1132 understood as prescriptive by sponsors to account for all uncertainties.

1133 Establishing external validity of historical data was raised as an issue in the second list of issues and  
1134 the Applicant addressed this with the updated guidance documents. The handbook provides definitions  
1135 and instructions to validate the prognostic model. Instructions include recommendations to collaborate  
1136 with model developers to establish the external validity of historical validation data sets. Specific  
1137 comments are provided on how to match the validation dataset to the trial population, on how to  
1138 account for the potential changes in the SOC, and how to address different extent of missing data  
1139 between the validation dataset and the trial data. These instructions are acknowledged. Prognostic  
1140 model validation using a data set that is independent from the historical training data and from the  
1141 study data, as proposed by the Applicant, is certainly endorsed to avoid too optimistic estimates of the  
1142 correlation coefficient. However, the feasibility of this step may be limited by the availability of  
1143 additional validation data that have similar properties as the planned study data.

1144 Moreover, it should be kept in mind that the sample size of a clinical trial should in most cases be  
1145 sufficient not only for the primary hypothesis test but also for providing a sufficiently large safety  
1146 database or, in some cases, to address more than one endpoint or the precision in important  
1147 subgroups (see Q4).

1148 With regard to the scenarios addressed with the empirical application of PROCOVA provided with the  
1149 briefing document, these are considered to be of relevance and the results of Experiment 1 and 2

1150 support the application of the proposed procedures. It is noted that data from patients who dropped  
1151 out of the study were not included in the analysis (p. 21, briefing document). This would not be  
1152 acceptable for regulatory purposes. It is also noted that the empirical applications mention two  
1153 outcomes of interest (ADAS-Cog11 and CDR at 18 months). While the sample size in the example  
1154 cases was calculated for ADAS-Cog11, analyses for CDR are also reported. With respect to co-primary  
1155 endpoints, the Applicant states in Section 3.1.1 "If there are multiple outcomes of interest, such as co-  
1156 primary endpoints, each with a desired power level and target effect size, then this procedure must be  
1157 repeated for each outcome, and the largest sample size should be selected." This approach is not in  
1158 general appropriate as it may result in insufficient power to reject all co-primary endpoints  
1159 simultaneously. Instead, the conjunctive power should be the basis for sample size calculations with  
1160 co-primary endpoints. However, it is agreed that in case of multiple endpoints of interest using  
1161 multiple prognostic models or a multivariate prognostic model may be necessary.

1162 The Applicant uses two-sided tests in the sample size and power calculations. Rejections due to  
1163 observed effects in both directions are counted as rejection of the null hypothesis. It is noted that from  
1164 a regulatory perspective, only one part of the comparisons may be relevant for study success. This  
1165 should usually be reflected in the hypothesis testing. With respect to considering the expected dropout  
1166 rate  $d$ , accounting for dropouts in sample size considerations as proposed by the Applicant using  
1167  $n_d = n / (1 - d)$  is generally reasonable. However, typically all randomised subjects should be included in  
1168 the primary analysis and a strategy to address post-randomisation events affecting the outcome as  
1169 well as missing data handling should be taken into account.

1170 In summary, the assumed reduction in residual variance due to a prognostic score may in principle be  
1171 taken into account to reduce sample size, if it can be ensured that the calculation is conservative with  
1172 respect to uncertainties in the assumptions made, and if the resulting sample size is large enough to  
1173 meet other relevant purposes apart from the primary hypothesis test.

#### 1174 **Question 6**

1175 **Does the EMA agree that PROCOVA™, combined with a predictive prognostic model and if**  
1176 **implemented as described, could enable increases in power and/or decreases in minimum**  
1177 **sample sizes in phase 2 or 3 clinical trials with continuous responses?**

#### 1178 ***Applicant's position***

1179 Our approach is designed to prospectively decrease the uncertainty, or variance, in treatment effect  
1180 estimates from RCTs without compromising strict type-1 error rate control in the large-sample setting.  
1181 We achieve this by combining curated historical control arm data, highly predictive modeling, and  
1182 covariate adjustment for the prognostic score generated through modeling.

1183 Our mathematical results (Section 3.1.2, Appendix 2, and Appendix 3), simulations (Section 3.2  
1184 and specifically Table 1 and Table 2, as well as Appendix 4) and empirical examples (Section 3.3,  
1185 Appendix 5, Appendix 6, and Appendix 7) demonstrate that PROCOVA™ can reduce variance of  
1186 the treatment effect estimates in trials with continuous responses.

1187 This reduction in variance can be leveraged either by increasing analytical power without  
1188 increasing the sample size (Section 3.3, Experiment 1), or by reducing the minimum required  
1189 sample size while maintaining the power (Section 3.3, Experiment 2). The Sponsor can make  
1190 that choice depending on the circumstances of a particular trial but must prospectively pre-  
1191 specify the application of PROCOVA™ prior to unblinding, to avoid bias.

1192 In summary, our method is scientifically sound since it only adjusts for a single covariate (or  
1193 single additional covariate) derived from information collected at baseline/prior to randomization;  
1194 produces unbiased estimates for treatment effects; controls the type-I error rate; and leads to

1195 correct confidence interval coverage. It is also consistent with current FDA and EMA regulatory  
1196 guidance. As such, PROCOVA™ can be used to prospectively increase the power or reduce the  
1197 minimum required sample size in studies that support drug approvals, i.e., pivotal/confirmatory  
1198 Phase 3, and occasionally Phase 2, clinical trials.

1199 **CHMP answer**

1200 In principle, CHMP agrees that implementing PROCOVA as prognostic score adjustment using a  
1201 prognostic model derived from independent data and the proposed procedures could enable increases  
1202 in power and/or decreases in sample size in phase 2 and 3 clinical trials with continuous outcomes. The  
1203 presented mathematical properties, simulation exercises and empirical application support this use.  
1204 Regarding choice of sample size, the answers to Questions 3 and 5 should be considered to safeguard  
1205 that the selected sample size is suitable for the trial objectives.

1206 Regarding the mathematical properties of PROCOVA, as implemented the method can be regarded a  
1207 special case of ANCOVA sharing the properties of type I error control and asymptotically unbiased  
1208 estimates of the treatment effect with sufficiently large sample sizes. For the weaker assumptions the  
1209 Applicant uses the term 'technical' assumptions (Schuler et al., arXiv:2012.09935v2 2021), which may  
1210 be debated. However, it can be agreed that similar assumptions are required for a large variety of  
1211 parametric frequentist methods regularly applied and accepted from a regulatory perspective.  
1212 Therefore, the proposed prognostic covariate procedure is an acceptable statistical approach.

1213 The potential advantages of the PROCOVA procedure and prognostic score adjustment more broadly,  
1214 depend on the availability of appropriate historical data and the derivation of a non-linear predictive  
1215 model that would allow outcome prediction in a future clinical trial. The number of covariates that can  
1216 be included in the modelling approach is determined by the size and quality of the historical dataset.  
1217 However, it is clear that type I error control, unbiased effect estimation and confidence interval  
1218 coverage are not dependent on the choice or performance of the prognostic model. It is noted that  
1219 prognostic score adjustment can be used together with adjustment using single covariates. The  
1220 consequence of using the prognostic score together with additional prognostic covariates (one or more)  
1221 needs to be carefully considered. The impact of the potential multicollinearity on the precision of the  
1222 estimated coefficients may outweigh the proposed advantage of using PROCOVA and should thus be  
1223 investigated in advance in order to inform the parameterisation to be used in the final primary analysis  
1224 model (as well as subgroup analyses). Using PROCOVA together with individual covariates for stratified  
1225 randomisation was addressed in a second list of issues. Subgroup analyses based on covariates  
1226 included in the prognostic score are addressed in an updated handbook for application of the PROCOVA  
1227 method (see also the answer to Question 2). This includes subgroup analysis for covariates that could  
1228 be predictive of treatment effect. If the treatment effect is expected to differ between subgroups due  
1229 to predictive biomarkers as covariate (in contrast to a prognostic covariate) and precision of the  
1230 treatment effect is especially important, additional power calculations are recommended to ensure  
1231 sufficient power for subgroup analysis. Additionally, the need for pre-specification of the prediction  
1232 model may be a disadvantage in case of only a low number of covariates relevant for outcome  
1233 prediction that could instead be included in an ANCOVA as single covariates with potential advantages  
1234 in interpretation of results.

1235 **Qualification opinion statement and conclusion**

1236 The Applicant proposes the PROCOVA method for estimation and statistical inference on the treatment  
1237 effect in randomized controlled clinical trials measuring continuous outcomes. The procedure involves  
1238 developing a prognostic score for the outcome under control based on a historical data set that is  
1239 independent from the study data and then applying the prognostic score as covariate in an ANCOVA  
1240 model for the actual data analysis of a clinical trial.



1241 The methodology comprises three steps:

1242 Step 1: Training and evaluating a prognostic model to predict outcomes under the control  
1243 condition (generate prognostic score).

1244 Step 2: Accounting for the prognostic score while estimating the sample size required for a  
1245 prospective study.

1246 Step 3: Estimating the treatment effect from the completed study using a linear model while  
1247 adjusting for the control outcomes predicted by the prognostic model.

1248 CHMP qualifies PROCOVA as prognostic score adjustment and the proposed procedures as described in  
1249 a handbook for trial statisticians could enable increases in power or precision of treatment effect  
1250 estimates in phase 2 and 3 clinical trials with continuous outcomes. The presented mathematical  
1251 properties, simulation exercises and empirical application support this use. The assumed reduction in  
1252 residual variance due to a prognostic score may in principle be taken into account to reduce sample  
1253 size, if it can be ensured that the calculation is considering uncertainties in the assumptions made, and  
1254 if the resulting sample size is large enough to meet other relevant purposes of the clinical trial apart  
1255 from the primary hypothesis test and treatment effect estimation.

1256 Regarding the mathematical properties of PROCOVA, as implemented the method can be regarded a  
1257 special case of ANCOVA sharing the properties of type I error control and asymptotically unbiased  
1258 estimates of the treatment effect with sufficiently large sample sizes. The method uses a number of  
1259 assumptions that are similar to those required by a large variety of parametric frequentist methods  
1260 that are regularly applied and accepted from a regulatory perspective. Therefore, the proposed  
1261 prognostic covariate procedure is an acceptable statistical approach for primary analysis of clinical  
1262 trials.

1263 An attainable advantage over using ANCOVA with single covariate adjustment should be justified to  
1264 support application of the PROCOVA method. The Applicant demonstrates that this should be the case  
1265 if the prognostic score is able to capture a nonlinear relationship between covariates and outcomes of  
1266 interest. The potential sample size reductions using PROCOVA depend on the ratio of the correlation  
1267 between a single baseline covariate (or a linear combination of the covariates that would typically be  
1268 considered in the analysis) and the outcome and the correlation between the single prognostic  
1269 (PROCOVA) score and the outcome. The gain in sample size (or likewise in power or precision of the  
1270 treatment effect estimates) should be evaluated at the stage of planning a trial, also taking the  
1271 optimism due to prognostic model building into account. The relative pros and cons of using PROCOVA  
1272 or ANCOVA should be compared to make a final determination to choose one of the three paths: no  
1273 adjustment, ANCOVA with one or more pre-specified covariates, or PROCOVA. The PROCOVA handbook  
1274 provides step-by-step instructions for the trial statistician to make an informed choice among these  
1275 three paths.

1276 The potential advantages of the PROCOVA procedure and prognostic score adjustment in general  
1277 depend on the availability of appropriate historical data and the derivation of a prediction model that  
1278 would allow outcome prediction in a future clinical trial. The number of covariates that can be included  
1279 in the modelling approach is determined by the size and quality of the historical dataset(s).  
1280 Establishing external validity of historical data is of paramount importance when applying a prediction  
1281 model in a future clinical trial. Type I error control, unbiased effect estimation and confidence interval  
1282 coverage are not dependent on the choice or performance of the prognostic model. PROCOVA can be  
1283 used together with adjustment using single covariates and stratified randomisation, but the  
1284 consequence of using the prognostic score together with additional prognostic covariates (one or more)  
1285 needs to be carefully considered. Where such single covariates and/or stratification factors are already  
1286 included in the prognostic score, the impact of the potential multicollinearity on the precision of the



1287 estimated coefficients may outweigh the proposed advantage. Recommendations given in the  
1288 handbook for trial statisticians on subgroup analysis should be followed. CHMP notes that impact of  
1289 multicollinearity when applying PROCOVA is not fully understood and additional research is desirable.  
1290 In addition to the recommendations in the handbook for use of PROCOVA for a case that involves  
1291 adjusting for additional covariates, including stratification factors for trials with stratified  
1292 randomization, an alternative option to exclude these additional covariates from the prognostic score  
1293 model may be explored before application.

1294 CHMP cannot qualify a formalised procedure for prediction model development as part of the PROCOVA  
1295 method. Only specific settings were explored and it cannot be foreseen if successful outcome prediction  
1296 will be possible for the proposed very general context of use. There may be disease conditions for  
1297 which prediction of endpoints selected for clinical trials is not possible with a desired precision or was  
1298 not successful in previous settings. Outcomes from historical data may not allow prediction of control  
1299 arm outcomes of future trials in case of changes in the therapeutic landscape. In addition, CHMP  
1300 cannot issue a statement about the precision of prediction models in general and if these models would  
1301 allow meaningful improvement in power or reductions in sample size. However, it is noted that  
1302 prediction models could help understanding disease characteristics or even mechanistic properties.

1303 The chosen approach to prediction model development is according to the Applicant explicitly out of  
1304 scope of this qualification procedure. There are advances in statistical 'learning' methods, the ability to  
1305 handle high-dimensional data and progress with e.g. machine learning or deep learning methods.  
1306 However, derivation of a prediction model would require careful work by sponsors or independent  
1307 groups with access to appropriate data sets. Sponsors should be aware of the risk of overfitting when  
1308 using more complex predictive modelling approaches, including machine learning and artificial  
1309 intelligence methodology. Therefore, assessment of correlation between observations and outcomes  
1310 with data independent of training data would be of importance to avoid too optimistic estimates of this  
1311 correlation. The updated handbook provides guidance for the choice of a deflation factor  $\lambda$ , and for the  
1312 conduct of sensitivity analyses taking into account a potential over-optimism of the prognostic model  
1313 and the fact that the correlation of the prognostic score with the outcome may be smaller in a future  
1314 trial including experimental treatment.

1315 In simulations performed by the Applicant, potential differences between the historical population used  
1316 to derive the prognostic model and the trial population were addressed with simulations using a  
1317 'shifted population'. While this is acknowledged, the robustness of the planned PROCOVA approach  
1318 with regard to availability of covariates in historical and future data, data quality with regard to  
1319 misspecification or measurement error and missing or incomplete covariates need to be carefully  
1320 assessed.

1321 Approaches with non-linear models for analysis and direct comparisons to such models, as well as  
1322 models with treatment-by-covariate interactions are out of scope of this qualification procedure.

---

<sup>i</sup> All annexes mentioned under the Applicant's position refer to the documentation submitted with the request.