Prediction-powered Inference for Clinical Trials: application to linear covariate adjustment

Pierre-Emmanuel Poulet, Maylis Tran,
Sophie Tezenas du Montcel, Bruno Dubois, Stanley Durrleman,
Bruno Jedynak, for the Alzheimers Disease Neuroimaging Initiative*

[1]ARAMIS, Sorbonne Université, Paris Brain Institute (ICM Institut du Cerveau), INRIA, INSERM, AP-HP, Groupe Hospitalier Sorbonne Université, Paris, France.
[2]Department of Mathematics and Statistics, Portland State University, 1855 SW Broadway, Portland, 97201, Oregon, USA.
[3]Institut de Neurologie, Hôpital Salpêtrière Sorbonne-Université, 47 bd de l'hôpital, Paris, 75651, cedex 13, France.

*Corresponding author(s). E-mail(s): Bruno.jedynak@pdx.edu;
Contributing authors: p.epoulet@gmail.com;
maylis.tran@icm-institute.org; bruno.dubois@psl.aphp.fr;
stanley.durrleman@icm-institute.org; sophie.tezenas@aphp.fr;

Prediction-powered inference (PPI) [1] and its subsequent development called PPI++ [2] provide a novel approach to standard statistical estimation, leveraging machine learning systems, to enhance unlabeled data with predictions. We use this paradigm in clinical trials. The predictions are provided by disease progression models, providing prognostic scores for all the participants as a function of baseline covariates. The proposed method would empower clinical trials by providing untreated digital twins of the treated patients while remaining statistically valid. The potential implications of this new estimator of the treatment effect in a two-arm randomized clinical trial

*Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`

(RCT) are manifold. First, it leads to an overall reduction of the sample size required to reach the same power as a standard RCT. Secondly, it advocates for an imbalance of controls and treated patients, requiring fewer controls to achieve the same power. Finally, this technique directly transfers any disease prediction model trained on large cohorts to practical and scientifically valid use. In this paper, we demonstrate the theoretical properties of this estimator and illustrate them through simulations. We show that it is asymptotically unbiased for the Average Treatment Effect and derive an explicit formula for its variance. We then compare this estimator to a regression-based linear covariate adjustment method. An application to an Alzheimer's disease clinical trial showcases the potential to reduce the sample size.

1 Introduction

Clinical trials are essential to demonstrate the benefits of a new treatment. Evaluating whether an intervention affects a measure of disease severity is one of the focus of causal inference [3]. Randomized clinical trials (RCT) provide a gold standard in this field. A key component in designing an RCT is estimating the treatment effect correctly and providing an uncertainty measure to perform a statistical test.

A primary lever to decrease uncertainty in a clinical trial's outcome is to increase the sample size. However, recruiting more subjects has a cost, both financial and ethical, since we cannot provide the treatment to every participant. Moreover, enrolling enough patients for a rare disease in a short time window can be extremely challenging.

Some approaches have been proposed to use additional information from patients in other cohorts, like the placebo arm of another RCT or patients in observational cohorts. Among such methods, we can cite Matching Adjusted Indirect Comparisons (MAIC) [4], which adjusts for covariates distribution shifts between the compared cohort and the clinical trial of interest, or Bayesian Dynamical Borrowing, which uses data from the previous study under the form of a Bayesian prior [5]. They lead to increased power in clinical trials. However, these techniques usually lack control over the type I error rate.

Simultaneously, machine-learning techniques have been developed in almost every field of medicine. Specifically, we have witnessed a sharp increase in disease progression models (DPM) performance over the last few years. Whether they are event-based models such as SuStaIn [6], non-linear mixed effect models such as Disease Course Mapping [7] or dynamical models based on differential equations [8, 9], these models have shown great potential in predicting the evolution of a disease over time. Our approach will thus rely on such systems to learn the disease progression patterns from historical data. They can then help select or stratify patients for a clinical trial [10] and improve clinical trial efficacy [11]. In our case, we will use the trained models to predict the evolution of the patients in the clinical trial from their baseline information (and, when available, pre-inclusion information). This will yield a prognostic score that will be included in our analysis. This is akin to the digital twins' methods [12].

By accounting for variability associated with prognostic factors, the precision of the estimated treatment effect can be improved, leading to narrower confidence intervals. Prior developments have been made to include prognostic factors in the statistical

analysis of an RCT, among which the prognostic covariate adjustment method [13] which has been approved by regulatory agencies [14]. Even though this method is quite recent, the idea to adjust for prognostic scores or other composite scores is not new [15]. Our approach builds on the recent concept of prediction-powered inference (PPI) [1], and its subsequent extension PPI++ [2], a method which bridges machine learning prediction systems and statistical estimators by enhancing the latter. More specifically, the method guarantees a decrease in the variance of the estimators, meaning narrower confidence intervals and potential sample size reduction. An application to medical data was proposed in [16]. We extend this framework to RCT with disease progression models as prognostic factors.

Even though prognostic factor adjustment is essential to improve the validity and efficacy of RCTs by increasing their power or reducing sample size, they can be prone to overfitting when too many factors are included. Our approach relies on a single prognostic score, also denoted a super-covariate in [17], which, on the downside, requires this score to be accurate. Finally, every part of the method is pre-specifiable for an RCT analysis and can be used to determine the sample size needed at the design stage.

Our contributions in this paper are the following: we propose adapting the prediction-powered estimators to clinical trials and derive their mathematical properties. We compare to ANCOVA [13], a linear covariate adjustment method. We perform basic simulations with small samples and unbalanced designs to contrast the different methods. We then perform a re-analysis of a randomized clinical trial for Alzheimer's disease. First, using simulations, and then using the actual trial data. In each case, we contrast different estimators of the average treatment effect: a reference estimator, or baseline, PPI, and ANCOVA.

2 Method

Let us introduce our notations. Let N be the total number of patients (or participants) in the clinical trial. The patients, denoted by index $i \in [1, N]$, are randomly assigned to the treated arm and the control arm. We use the words " treatment and control " generically while acknowledging that the control arm is often associated with a default treatment. For simplicity, we do not consider covariate-adaptive randomization in our analysis, such as stratified permuted block design [18] or Pocock-Simon's minimization [19]. Without loss of generality, and for convenience of notation, we assume the first m patients to be in the treated arm while the $n = N - m$ others are controls or placebo. The values of m and n can be adjusted in the trial's design, allowing for an imbalance.

We denote $(X_i)_{1 \leq i \leq N} \in \mathcal{X}^N$ the baseline variables of the patients, $(Y_i)_{1 \leq i \leq N} \in \mathbb{R}^N$ the outcome of the clinical trial, and $(T_i)_{1 \leq i \leq N} \in \{0, 1\}^N$ the treatment variable, such that for $1 \leq i \leq m, T_i = 1$ and for $m + 1 \leq i \leq N, T_i = 0$. We will often use Y_i^c (resp. Y_i^t) to denote $Y_{i|T_i=0}$ (resp. $Y_{i|T_i=1}$). We assume throughout that given the randomization, the patients' random variables are independent samples from a common distribution $(X_i, Y_i|T_i) \sim \mathcal{D}$. Adequate randomization ensures that $X_i \perp T_i$.

The average treatment effect is our estimand, which is defined as follows:

Definition 1 (ATE)**.** *Using previously introduced notations, let the average treatment effect:*

$$ATE = \mathbb{E}(Y_i|T_i = 1) - \mathbb{E}(Y_i|T_i = 0)$$

It is the difference between $\theta_1 = \mathbb{E}(Y_i|T_i = 1)$ and $\theta_0 = \mathbb{E}(Y_i|T_i = 0)$.

We assume that we are given a prediction function $f : X \in \mathcal{X} \mapsto f(X) \in \mathbb{R}$. In theory, this function can be any black-box predictor; in practice, it is designed to work best with a prognostic score such that $f(X_i)$ is highly correlated with Y_i^c. A special case of interest is when f is a machine-learning predictor. In this case, we require that f be trained with data that is disjoint from that of the trial.

We also introduce notations for the variance and covariance of the variables mentioned previously: $\sigma_t^2 = \mathbb{V}(Y_i^t)$, $\sigma_c^2 = \mathbb{V}(Y_i^c)$, $\sigma_f^2 = \mathbb{V}(f(X_i))$, $\rho_t = \frac{Cov(f(X_i),Y_i^t)}{\sigma_f \sigma_t}$, $\rho_c = \frac{Cov(f(X_i),Y_i^c)}{\sigma_f \sigma_c}$.

2.1 Reference estimator

The simplest estimator for the average treatment effect, which we will call the reference estimator, is the following:

Definition 2 (Reference estimator)**.** *Let the reference estimator for ATE:*

$$\widehat{ATE}_N^{ref} = \frac{1}{m}\sum_{i=1}^{m} Y_i^t - \frac{1}{n}\sum_{i=m+1}^{N} Y_i^c$$

We often neglect the subscript N in the following to simplify notations. The Central Limit Theorem provides the proof of the next proposition:

Proposition 1. *Assuming samples $(Y_i, X_i, T_i)_i$ are i.i.d, then:*

$$\mathbb{E}(\widehat{ATE}_N^{ref}) = ATE$$

$$\mathbb{V}(\widehat{ATE}_N^{ref}) = \frac{\sigma_c^2}{n} + \frac{\sigma_t^2}{m}$$

and the estimator is asymptotically normal:

$$\sqrt{N}(\widehat{ATE}_N^{ref} - ATE) \xrightarrow[N\to\infty]{d} \mathcal{N}(0, V^{ref})$$

where $V^{ref} = \frac{\sigma_t^2}{\pi_t} + \frac{\sigma_c^2}{\pi_c}$ with $\lim_{N\to\infty} \frac{n}{N} = \mathbb{P}(T_i = 0) = \pi_c$ and $\lim_{N\to\infty} \frac{m}{N} = \mathbb{P}(T_i = 1) = \pi_t$.

From the variance formula we can derive an optimal partition between controls and treated patients: $n^*_{ref} = N\frac{\sigma_c}{\sigma_c+\sigma_t}$ which yields $\mathbb{V}(\widehat{ATE}^{ref}; n*_{ref}) = \frac{(\sigma_c+\sigma_t)^2}{N}$. Designing a trial with this optimal partition implies that one has access to approximate values for the standard deviations of the outcome in the control arm and the treated arm, respectively σ_c and σ_t. It can be done using a pilot study or a previous clinical trial. However, it is often assumed that $\sigma_t = \sigma_c$ for simplicity, alleviating the estimation burden as the variance of the outcome under no intervention is often easier to estimate, for instance, using an observational cohort. Note that the reference estimator is seldom used when covariates are available.

2.2 Prediction-powered estimator

Prediction-powered inference (PPI) [2] relied on the intuition that it might be better to estimate θ_0 from predictions $(f(X_i))_{1\leq i\leq N}$ rather than relying on the control observations $(Y_i)_{m<i\leq N}$. However $\hat{\theta}_0^{naive} = \frac{1}{N}\sum_{i=1}^{N} f(X_i)$ has no guarantee to be unbiased. Therefore a correction to the predictions is obtained by computing the empirical bias between control patients and their prediction: $\hat{\Delta} = \frac{1}{n}\sum_{i=m+1}^{N}(Y_i^c - f(X_i))$, a term which is called the *rectifier*. Combining with the empirical mean for the estimator of θ_1, we get:

Definition 3 (PPI estimator)**.** *Using previous notations, let the prediction-powered inference estimator:*

$$\widehat{ATE}_N^{PPI} = \hat{\theta}_1 - (\hat{\theta}_0^{naive} + \hat{\Delta}) = \frac{1}{m}\sum_{i=1}^{m}(Y_i^t - f(X_i)) - \frac{1}{n}\sum_{i=m+1}^{N}(Y_i^c - f(X_i))$$

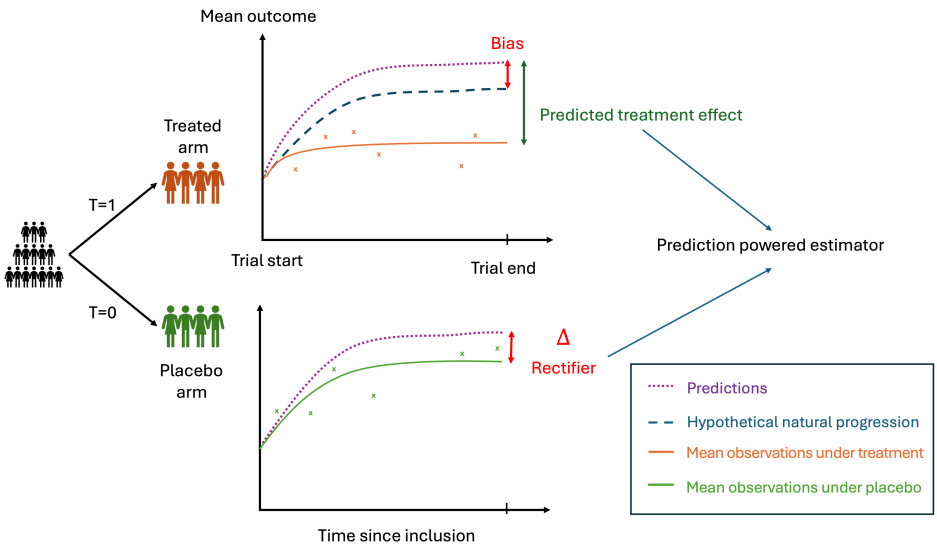

Fig. 1 General scheme of the prediction-powered Inference estimator for Average Treatment Effect in a randomized clinical trial. The estimator leverages a prognostic model predicting the natural disease history of a patient using only information available at baseline. The estimator combines two terms: the predicted treatment effect, which is the difference between the observed outcome of treated patients and their predicted natural progression, and the rectifier, which accounts for the bias of the predictions (including both model error and placebo effect). The result is a new unbiased estimator whose variance depends on the quality of the predictions.

Using the CLT, we obtain:

Proposition 2.

$$\mathbb{E}(\widehat{ATE}_N^{PPI}) = ATE$$

$$\mathbb{V}(\widehat{ATE}_N^{PPI}) = \frac{\sigma_t^2 + \sigma_f^2 - 2\sigma_f\sigma_t\rho_t}{m} + \frac{\sigma_c^2 + \sigma_f^2 - 2\sigma_f\sigma_c\rho_c}{n}$$

and the estimator is asymptotically normal:

$$\sqrt{N}(\widehat{ATE}_N^{PPI} - ATE) \xrightarrow[N\to\infty]{d} \mathcal{N}(0, V^{PPI})$$

where $V^{PPI} = \frac{\sigma_t^2+\sigma_f^2-2\sigma_f\sigma_t\rho_t}{\pi_t} + \frac{\sigma_c^2+\sigma_f^2-2\sigma_f\sigma_c\rho_c}{\pi_c}$.

Note firstly that the prognostic models need not be unbiased. It is rectified by the observation gathered at the end of the trial. This is the most salient feature of PPI. Secondly, compared with the variance of the reference estimator, see proposition 1, the variance of the PPI estimator depends on the quality of the prediction function f through the terms ρ_c, ρ_t measuring the correlation of the prognostic score with the outcome within each arm. Qualitatively, if the prognostic score strongly correlates with the outcome, the PPI estimator has a smaller variance than the standard one. If the prognostic score strongly negatively correlates with the outcome, one can reverse its sign and obtain the same result. However, a larger variance is obtained when the prognostic score weakly correlates with the outcome. This last case is highly undesirable as we do not want to risk losing power in a clinical trial setting.

2.3 Empowered estimators

2.3.1 PPI++ estimator

In the extension of the initial PPI paper [2], the authors proposed circumventing the PPI method's limitations by introducing a weighted contribution of the predictions and optimizing. The new estimator resulting from this operation is called the PPI++ estimator:

Definition 4 (PPI++ estimator)**.** *Let* $\lambda \in \mathbb{R}$*, then:*

$$\widehat{ATE}_N^{PPI\lambda} = \frac{1}{m}\sum_{i=1}^{m}(Y_i^t - \lambda f(X_i)) - \frac{1}{n}\sum_{i=m+1}^{N}(Y_i^c - \lambda f(X_i))$$

Notice that $\widehat{ATE}_N^{PPI\lambda} = (1-\lambda)\widehat{ATE}^{ref} + \lambda\widehat{ATE}^{PPI}$. Therefore, this estimator is unbiased. Again, the CLT provides the following:

Proposition 3.

$$\mathbb{E}(\widehat{ATE}_N^{PPI\lambda}) = ATE$$

$$\mathbb{V}(\widehat{ATE}_N^{PPI\lambda}) = \frac{\sigma_t^2 + \lambda^2\sigma_f^2 - 2\lambda\sigma_f\sigma_t\rho_t}{m} + \frac{\sigma_c^2 + \lambda^2\sigma_f^2 - 2\lambda\sigma_f\sigma_c\rho_c}{n}$$

and the estimator is asymptotically normal:

$$\sqrt{N}(\widehat{ATE}_N^{PPI\lambda} - ATE) \xrightarrow[N\to\infty]{d} \mathcal{N}(0, V^{PPI\lambda})$$

where $V^{PPI\lambda} = \frac{\sigma_t^2+\lambda^2\sigma_f^2-2\lambda\sigma_f\sigma_t\rho_t}{\pi_t} + \frac{\sigma_c^2+\lambda^2\sigma_f^2-2\lambda\sigma_f\sigma_c\rho_c}{\pi_c}$.

2.3.2 PPCT estimator

We now define the Prediction-Powered for Clinical Trial (PPCT) estimator. Since λ is a design parameter, we can optimize it. We choose to minimize $V^{PPI\lambda}$, which is equivalent to maximizing the power of the statistical test of the clinical trial. This fact will be discussed in section 2.4.2.

Definition 5 (PPCT estimator)**.** *Let λ^* be the value minimizing the variance of the PPI++ estimator. The PPCT estimator is then:*

$$\widehat{ATE}^{PPCT} = \widehat{ATE}^{PPI\lambda^*}$$

This definition can be made further explicit:

Proposition 4 (Optimal λ^* and variance of PPCT)**.** *The PPCT estimator is unbiased for the ATE and:*

$$\lambda^* = \frac{\pi_c \sigma_t \rho_t + \pi_t \sigma_c \rho_c}{\sigma_f}$$

$$\mathbb{V}(\widehat{ATE}_N^{PPCT}) = \frac{\sigma_t^2}{m} + \frac{\sigma_c^2}{n} - (\lambda^*)^2 \sigma_f^2 \left(\frac{1}{n} + \frac{1}{m}\right)$$

with the asymptotical variance:

$$\mathbb{V}(\widehat{ATE}^{PPCT}) = \frac{\sigma_t^2}{\pi_t} + \frac{\sigma_c^2}{\pi_c} - (\lambda^*)^2 \sigma_f^2 \left(\frac{1}{\pi_c} + \frac{1}{\pi_t}\right)$$

Comparing with Proposition 1, we obtain:

Corollary 5 (PPCT efficiency)**.** *The PPCT estimator is always more efficient than the reference estimator for ATE.*

When computing the estimator, we use consistent estimators of the standard deviations and correlation coefficients to derive a consistent estimator of λ^*. In the following applications, we will use the empirical standard deviation and correlation computed from the data. The optimal value of λ indicates the quality of the predictor as well: the closer λ^* is to 1, the better the predictions. From a theoretical perspective, $\lambda^* = 0$ corresponds to the situation where $\rho_c = \rho_t = 0$, so that the predictor is uncorrelated with the outcome. In this case, the PPCT estimator is equal to the reference estimator. On the other hand, $\lambda^* = 1$ corresponds to $\rho_c = \rho_t = 1$. The prognostic score perfectly correlates with the outcome, and the PPCT estimator equals the PPI one.

Solving explicitly for an optimal partition of controls and treated patients is also an option: $n^*_{PPI\lambda} = N \frac{1}{1+\sqrt{\frac{\sigma_t^2+\lambda^2\sigma_f^2-2\lambda\sigma_f\sigma_t\rho_t}{\sigma_c^2+\lambda^2\sigma_f^2-2\lambda\sigma_f\sigma_c\rho_c}}}$. This expression can roughly be interpreted as follows: if the prognostic score correlates better with the outcome in the control arm than in the treated arm, then we increase the proportion of treated patients (and conversely). However, when designing the protocol, it is challenging to obtain reliable estimates of the prognostic score correlation with the outcome under no treatment ρ_c and, more especially, the correlation with the outcome after intervention ρ_t, hence limiting the potential application of our method to the design of unbalanced arms trials.

2.3.3 Parallel with prognostic covariate adjustment

We refer to the paper by [13] for more details about the formulas of prognostic covariate adjustment, which has received approval from both FDA and EMA [14] as a method to reduce clinical trial sample size. This method performs an ANCOVA using the prognostic score as a covariate.

Definition 6 (ANCOVA estimator)**.** *Let us introduce the design matrix* $\mathbf{Z}^T = [\mathbf{1}, \mathbf{T}, f(\mathbf{X}), \mathbf{T}f(\mathbf{X})]$*. The outcome is modeled as a linear function of the design matrix:*

$$\mathbf{Y} = \beta^T \mathbf{Z} \tag{1}$$

where β *are the coefficients of the regression and are estimated through ordinary least squares:* $\hat{\beta} = \mathbf{Z}^{\dagger} Y$*. The ANCOVA estimator of ATE,* $\widehat{ATE}^{ANCOVA}$*, corresponds to the second coordinate of* $\hat{\beta}$*.*

The asymptotic variance of this estimator provided by [13] is:

Proposition 6. *Assuming samples* $(Y_i, X_i, T_i)_i$ *are i.i.d, then:*

$$\sqrt{N}(\widehat{ATE}_N^{ANCOVA} - ATE) \xrightarrow[N\to\infty]{d} \mathcal{N}(0, V^{ANCOVA})$$

where in the scalar case, $V^{ANCOVA} = \frac{\sigma_c^2}{\pi_c} + \frac{\sigma_t^2}{\pi_t} - \left(\frac{1}{\pi_c \pi_t}\right) \xi_*^T V \xi_*$ *with* $\xi_* = \pi_c \rho_t \sigma_t \sigma_f + \pi_t \rho_c \sigma_c \sigma_f$ *and* $V = \mathbb{V}(f(X))$*.*

This variance is the same as $\mathbb{V}(\widehat{ATE}^{PPCT})$. This should not come as a surprise as both are linear estimators adjusted with a prognostic score with the aim of minimizing variance. If we actually solve explicitly (1) in the case where $\dim(f(\mathbf{X})) = 1$ (proof in supplementary), we obtain:

Proposition 7 (Unidimensional ANCOVA estimator)**.** *If* $\dim(f(\boldsymbol{X})) = 1$*, then:*

$$\widehat{ATE}^{ANCOVA} = \frac{1}{m}\sum_{i=1}^{m}(Y_i^t - \lambda_t f(X_i)) - \frac{1}{n}\sum_{i=m+1}^{N}(Y_i^c - \lambda_c f(X_i))$$

where $\lambda_c = \frac{\hat{\rho}_c \hat{\sigma}_c}{\hat{\sigma}_{fc}}$ *and* $\lambda_t = \frac{\hat{\rho}_t \hat{\sigma}_t}{\hat{\sigma}_{ft}}$*, with* $\hat{\rho}_c$ *being the empirical correlation between* Y *and* $f(X)$*,* $\hat{\sigma}_c$ *the empirical standard deviation of* Y *and* $\hat{\sigma}_{fc}$ *the empirical standard deviation of* $f(X)$ *under* $T = 0$ *(similarly superscript* t *applies with* $T = 1$*).*

This expression has many similarities to the $\widehat{ATE}^{PPCT}$ estimator, which instead of two different coefficients λ_c and λ_t has only one λ^*. However, being both consistent linear estimators, PPCT and ANCOVA are asymptotically equivalent, both in their pointwise ATE estimator and their variance estimator.

One noteworthy difference is that the variance of the PPCT estimator for a sample of size N has an analytic expression as a function of the model parameters, providing a more explicit and, thus, possibly more explainable estimator compared to ANCOVA. Another advantage of the PPI++ formulation is the possibility to generalize it to a non-linear setting, e.g., time-to-event data, see section 2.5.

2.4 Clinical Trial enhancement

2.4.1 Statistical testing

Using the asymptotically normal distributions of the aforementioned estimators, we can perform two-sided tests for the null hypothesis H_0: $ATE = 0$ vs the alternative H_0: $ATE \neq 0$. Namely:

Proposition 8. *Let $\alpha \in (0,1)$ be the desired type I error rate of the test. The null hypothesis is rejected when*

$$\widehat{ATE} \notin [-\Phi^{-1}(1-\frac{\alpha}{2})\hat{V}, +\Phi^{-1}(1-\frac{\alpha}{2})\hat{V}]$$

where Φ is the cumulative distribution of the normal distribution and $\widehat{ATE}$ is an ATE estimator among $\widehat{ATE}^{ref}$, $\widehat{ATE}^{PPI}$, $\widehat{ATE}^{PPI\lambda}$, $\widehat{ATE}^{PPCT}$, $\widehat{ATE}^{ANCOVA}$ and $\hat{V}$ is a consistent estimator of its asymptotic variance.

As for standard clinical trial design, the normality assumption is valid only for large sample sizes, which is often considered to be the case when $n, m \geq 20$. An alternative, valid for all the estimators described, is to use a permutation test [20] for assessing the success of the trial.

2.4.2 Power and sample size computation

To compute the statistical test's power, we introduce the treatment effect Δ_H under the alternative hypothesis.

Definition 7 (Power)**.** *Let β the power of the statistical test in proposition 8:*

$$1-\beta = \mathbb{P}(\widehat{ATE} \notin [-\Phi^{-1}(1-\frac{\alpha}{2})\hat{V}, +\Phi^{-1}(1-\frac{\alpha}{2})\hat{V}] | H_1)$$

Under normality assumptions, the formula linking the power β of a trial to the sample size based on the hypothesized treatment effect size Δ_H and type-I error rate α is, therefore:

Proposition 9. *Let $\alpha, \beta \in (0,1)$ and $\Delta_H \in \mathbb{R}$. The minimum sample size required for the statistical test in proposition 8 to have a level of α and a power of $1-\beta$ is:*

$$N = \mathbb{V}(\widehat{ATE}) \left(\frac{\Phi^{-1}(1-\frac{\alpha}{2}) + \Phi^{-1}(1-\beta)}{\Delta_H} \right)^2$$

This shows how variance reduction directly affects the required sample size, with a proportional effect. A similar analysis would show the same proportional effect for a one-sided test when we are interested in the power to show a positive effect of active treatment, not just a difference (that could be negative). We undergo the routinely made hypotheses: $\sigma_c = \sigma_t$ and $\rho_c = \rho_t = \rho$ (which holds for a constant treatment effect, for instance), we obtain the following relationship for asymptotic variances of the estimators:

$$\mathbb{V}(\widehat{ATE}^{ANCOVA}) = \mathbb{V}(\widehat{ATE}^{PPI\lambda^*}) = (1-\rho^2)\mathbb{V}(\widehat{ATE}^{ref}) \tag{2}$$

where ρ^2 is actually R^2, the out-of-sample R-squared.

This is particularly useful in clinical trial design as a rule to reduce sample size based on the estimated R^2 of the prognostic factor. This formula implies that for the same power and type-I error control, the relative reduction in CT sample size is R^2. We will verify whether this principle holds in our experiments on both simulated and real clinical trial data.

However, we want to emphasize that such simplifying assumptions are not likely to hold, especially $\rho_c = \rho_t$, which assumes that the prognostic score explains as much variance of the treated arm as the control/placebo arm. This is very bold in the context of disease progression models, which are trained to predict the evolution under no intervention. We thus provide an analysis with more conservative assumptions.

We are interested in computing the power gain obtained using the PPCT estimator compared to the reference estimator as a function of the estimator's performance. We focus here on reducing the size of the control arm.

Consider a standard design with a control and a treatment arm of the same size, notated m. The test statistic in proposition 8 uses $\widehat{ATE}^{ref}$. We are interested in comparing this design to an alternative one where the test statistic is the PPCT ATE estimator and the groups are unbalanced. The treatment group is of size m, the same as the standard design, while the control group is of size $n_* \leq m$. We make the following assumptions:

1. We know σ_c. We have access to a trained predictor f. We know ρ_c. Without loss of generality, we assume that $\rho_c \geq 0$. If this were not the case, we would replace f with $-f$;
2. $\sigma_t = \sigma_c := \sigma$ This is a standard assumption in the design of clinical trials;
3. $\rho_t \geq 0$. This expresses that the signs of ρ_c and ρ_t are the same. Note that this is a minimal but still restrictive condition.

We constrain the PPCT to have at least the same power as the reference estimator. This is obtained by having the variance of the PPCT estimator be less or equal to the variance of the reference estimator. This provides the following

$$\left(\frac{1}{m} + \frac{1}{n_*}\right)(\sigma^2 - \lambda_*^2 \sigma_f^2) \leq \frac{2\sigma^2}{m} \tag{3}$$

with

$$\lambda^* = \frac{\sigma}{\sigma_f(n_* + m)}(n_* \rho_t + m\rho_c) \tag{4}$$

Simplifying, we obtain that

$$(n\rho_t + m\rho_c)^2 \geq m^2 - n_*^2 \tag{5}$$

Using the fact that $\rho_c \geq 0$ and $\rho_t \geq 0$, we obtain that a sufficient condition is

$$(m\rho_c)^2 \geq m^2 - n_*^2 \tag{6}$$

or equivalently,

$$n_* \geq m\sqrt{1-\rho_c^2} \tag{7}$$

Thus, we can choose

$$n_* = m\sqrt{1-\rho_c^2} \tag{8}$$

and obtain a design with the same power as the standard design. Note the following from (8)

1. when $\rho_c = 0$, $n^* = m$ and the PPCT design is the same as the standard design;
2. when $\rho_c = 1$, $n_* = 0$, there is no need for a control group.

A range of values for n_*/m are provided as a function of ρ_c in Table 1. A prediction highly correlated with the outcome (ρ_c large) would allow for a drastic reduction in the relative size of the control group compared to a standard balanced RCT

ρ_c	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\frac{n_*}{m}$	1.00	0.99	0.98	0.95	0.92	0.87	0.80	0.71	0.60	0.44	0.00

Table 1 Ratio of the size of the control group (n_*) to the size of the treatment group (m) for obtaining the same power as a standard balanced RCT. This ratio is shown as a function of the correlation between the classifier and the outcome (ρ_c).

2.5 General prediction-powered for clinical trial estimator

The PPI method can be generalized to other estimators of interest in the context of clinical trials. We define the largest possible application framework, as in PPI++ [2], by focusing on a class of estimators resulting from the maximization of an objective function. Equivalently, such an estimator can be defined using a loss function $\ell_\theta : (X,Y) \in \mathcal{X} \times \mathcal{Y} \mapsto \ell_\theta(X,Y) \in (R)$ parametrized by $\theta \in \Theta$, from which we define the population loss $L(\theta) = \mathbb{E}[\ell_\theta(X,Y)]$, and the prediction loss $L^f(\theta) = \mathbb{E}[\ell_\theta(X, f(X))]$. The estimand is $\theta^* \in \arg\min_{\theta\in\Theta} L(\theta)$. We introduce the relations to the data observations with the empirical losses:

$$L_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell(\theta, Y_i) \tag{9}$$

$$L_n^f(\theta) = \frac{1}{n}\sum_{i=m+1}^{N} \ell(\theta, f(X_i^c)) \tag{10}$$

$$L_m^f(\theta) = \frac{1}{m}\sum_{i=1}^{m} \ell(\theta, f(X_i^t)) \tag{11}$$

$$\tag{12}$$

From there we define the PPCT loss:

$$L_N^{PP}(\theta) = L_N(\theta) + \lambda(L_m^f(\theta) - L_n^f(\theta)) \tag{13}$$

with $\lambda \in \mathbb{R}$. Using the fact that $X \perp T$, we have $\mathbb{E}[L_m^f(\theta)] = \mathbb{E}[L_n^f(\theta)] = L^f(\theta)$, thus $\mathbb{E}[L_N^{PP}(\theta)] = L(\theta)$. Then:

Definition 8 (General PPCT estimator)**.** *Let the general PPCT estimator:*

$$\hat{\theta}^{PP} \in \arg\min_{\theta \in \Theta} L_N^{PP}(\theta)$$

The PPI++ analysis [2] can be generalized to show that under some regularity assumptions, $\hat{\theta} \overset{N\to\infty}{\longrightarrow} \theta^*$ strongly. Many estimators fall under this generic framework, including linear regression (for instance, adjusted for covariates in an ANCOVA), logistic regression, and Cox regression. Details will be provided elsewhere.

3 Experiments and results

We will showcase the performance of the PPCT estimator on both synthetic data and real-life clinical trial data. Our application is focused on Alzheimer's disease (AD) but is by no means restricted to AD or neurodegenerative diseases. The interest lies in the length of such chronic diseases, where the follow-up of patients has allowed the development of disease progression models. Such models have shown promising results for predicting the future evolution of the patients, for instance, in the TADPOLE challenge [21].

3.1 Simulations in unbalanced settings

In this first experiment, we present a simple simulation setting to illustrate the differences between the estimators under study. We particularly examine the performance of treatment effect estimators under unbalanced arms settings by varying the number of control patients and investigating how the quality of the prognostic score affects both the pointwise ATE estimators and their variance estimators. Let us describe the simulation model :

$$Y_c \sim \mathcal{F} \tag{14}$$

$$Y_t = Y_c + \Delta \tag{15}$$

$$f(X) = Y_c + \epsilon \tag{16}$$

$$\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2) \tag{17}$$

where ϵ is the noise of the predictor, with $\sigma_\epsilon^2 = \sigma_c^2 \frac{1-\rho_c^2}{\rho_c^2}$. Notice that $\rho_c = \rho_t$ since we consider a constant treatment effect. We fix the treatment effect $\Delta = 1$. We evaluate two distributions, first the Normal (or Gaussian) distribution $\mathcal{F} = \mathcal{N}(0, \sigma_c^2)$ with $\sigma_c = 1$ and second the Chi-square distribution with 5 degrees of freedom to have a skewed distribution. We set the number of treated patients to $m = 100$ while varying the number of controls $n = 2, 5, 10, 50, 100$ to investigate how a strong predictor can facilitate the reduction in the number of control patients in a clinical trial. The

results over 1000 simulations in each setting are shown in the tables 2 and C1 in the supplementary material.

ρ	n	**Reference**	**PPI**	**PPCT**	**PPCT2**	**ANCOVA**
Normal						
0.00	2	$\mathbf{0.99 \pm 0.71}$	0.99 ± 1.0	0.98 ± 0.95	$\mathbf{0.99 \pm 0.71}$	0.45 ± 20
	5	$\mathbf{1.0 \pm 0.48}$	1.0 ± 0.67	1.0 ± 0.53	$\mathbf{1.0 \pm 0.48}$	1.0 ± 0.57
	10	$\mathbf{0.99 \pm 0.32}$	0.99 ± 0.47	0.99 ± 0.33	$\mathbf{0.99 \pm 0.32}$	0.99 ± 0.34
	50	$\mathbf{1.0 \pm 0.17}$	1.0 ± 0.25	1.0 ± 0.18	1.0 ± 0.18	1.0 ± 0.18
	100	$\mathbf{1.0 \pm 0.14}$	1.0 ± 0.21	$\mathbf{1.0 \pm 0.14}$	$\mathbf{1.0 \pm 0.14}$	$\mathbf{1.0 \pm 0.14}$
0.50	2	0.99 ± 0.72	1.0 ± 1.2	1.0 ± 0.89	$\mathbf{1.0 \pm 0.62}$	1.1 ± 5.2
	5	0.99 ± 0.46	1.0 ± 0.81	1.0 ± 0.45	$\mathbf{1.0 \pm 0.41}$	1.0 ± 0.47
	10	1.0 ± 0.34	1.0 ± 0.58	1.0 ± 0.30	$\mathbf{1.0 \pm 0.29}$	1.0 ± 0.30
	50	1.0 ± 0.18	0.99 ± 0.30	$\mathbf{1.0 \pm 0.15}$	$\mathbf{1.0 \pm 0.15}$	$\mathbf{1.0 \pm 0.15}$
	100	1.0 ± 0.14	0.99 ± 0.24	$\mathbf{0.99 \pm 0.13}$	$\mathbf{0.99 \pm 0.13}$	$\mathbf{0.99 \pm 0.13}$
0.95	2	$\mathbf{0.99 \pm 0.73}$	1.0 ± 1.0	0.98 ± 0.98	1.0 ± 1.0	-0.34 ± 32
	5	1.0 ± 0.46	1.0 ± 0.46	$\mathbf{0.99 \pm 0.41}$	1.0 ± 0.46	1.0 ± 0.45
	10	$\mathbf{0.99 \pm 0.33}$	1.0 ± 0.38	$\mathbf{1.0 \pm 0.33}$	$\mathbf{1.0 \pm 0.33}$	$\mathbf{1.0 \pm 0.33}$
	50	1.0 ± 0.17	$\mathbf{0.99 \pm 0.15}$	$\mathbf{0.99 \pm 0.15}$	$\mathbf{1.0 \pm 0.15}$	$\mathbf{1.0 \pm 0.15}$
	100	1.0 ± 0.14	$\mathbf{1.0 \pm 0.13}$	$\mathbf{1.0 \pm 0.13}$	$\mathbf{1.0 \pm 0.13}$	$\mathbf{1.0 \pm 0.13}$
Chi-square						
0.00	2	$\mathbf{1.0 \pm 2.2}$	1.1 ± 2.3	0.87 ± 3.2	$\mathbf{1.0 \pm 2.2}$	-2.6 ± 67
	5	$\mathbf{0.95 \pm 1.5}$	$\mathbf{0.96 \pm 1.5}$	0.94 ± 1.6	$\mathbf{0.95 \pm 1.5}$	0.96 ± 1.7
	10	$\mathbf{0.99 \pm 1.0}$	1.0 ± 1.1	1.0 ± 1.1	$\mathbf{0.99 \pm 1.0}$	1.0 ± 1.1
	50	$\mathbf{1.0 \pm 0.55}$	1.0 ± 0.58	1.0 ± 0.56	$\mathbf{1.0 \pm 0.55}$	1.0 ± 0.56
	100	$\mathbf{0.98 \pm 0.45}$	0.99 ± 0.48	$\mathbf{0.98 \pm 0.45}$	$\mathbf{0.98 \pm 0.45}$	$\mathbf{0.98 \pm 0.45}$
0.50	2	0.98 ± 2.3	0.98 ± 1.2	1.6 ± 2.1	$\mathbf{0.96 \pm 1.1}$	2.8 ± 52
	5	0.98 ± 1.4	1.0 ± 0.79	1.3 ± 0.85	$\mathbf{1.0 \pm 0.69}$	1.1 ± 0.78
	10	0.97 ± 1.1	1.0 ± 0.59	1.2 ± 0.61	$\mathbf{1.0 \pm 0.52}$	1.1 ± 0.55
	50	1.0 ± 0.54	1.0 ± 0.30	$\mathbf{1.0 \pm 0.26}$	$\mathbf{1.0 \pm 0.26}$	$\mathbf{1.0 \pm 0.26}$
	100	1.0 ± 0.45	1.0 ± 0.24	$\mathbf{1.0 \pm 0.21}$	$\pm 1.0 \pm 0.21$	$\mathbf{1.0 \pm 0.21}$
0.95	2	0.93 ± 2.3	$\mathbf{1.0 \pm 0.24}$	1.7 ± 2.1	0.99 ± 0.25	1.2 ± 7.3
	5	0.96 ± 1.5	$\mathbf{1.0 \pm 0.15}$	1.4 ± 0.87	0.99 ± 0.16	1.0 ± 0.21
	10	0.99 ± 1.0	$\mathbf{1.0 \pm 0.11}$	1.2 ± 0.33	$\mathbf{1.0 \pm 0.11}$	1.0 ± 0.12
	50	0.99 ± 0.57	$\mathbf{1.0 \pm 0.06}$	1.0 ± 0.07	$\mathbf{1.0 \pm 0.06}$	$\mathbf{1.0 \pm 0.06}$
	100	1.0 ± 0.47	$\mathbf{1.0 \pm 0.05}$	$\mathbf{1.0 \pm 0.05}$	$\mathbf{1.0 \pm 0.05}$	$\mathbf{1.0 \pm 0.05}$

Table 2 Mean ATE estimation $\pm$ one standard deviation estimated with 1,000 Monte-Carlo simulations. **Bold**: estimators with the lowest standard deviation per row. Red: estimators that are problematic, being either very biased or having an overly high variance.

Table 2 shows the values of five point-wise treatment effect estimators $\widehat{ATE}$: the reference, the PPI, the PPCT, and a variant denoted PPCT2, and the ANCOVA estimators. These estimators are described in the method section. The PPCT estimator uses $\widehat{\lambda^*} = \frac{n\hat{\sigma_t}\hat{\rho_t} + m\hat{\sigma_c}\hat{\rho_c}}{N\hat{\sigma_f}}$ with $\hat{\sigma}$ being the empirical standard deviation (with one degree of freedom to be unbiased) and $\hat{\rho}$ being the empirical correlation. This yields a consistent estimator for λ^*. However, note that $\hat{\rho}_c$ is being multiplied by m while $\hat{\rho}_t$ is multiplied by n, which means that in extremely unbalanced settings with $n << m$, the variance of this estimator is high due to the dominant term being $\hat{\rho}_c$ which is estimated over the few n subjects. Therefore, we proposed a variant, denoted PPCT2, for

λ^*, which relies on a constant treatment effect assumption such that $\rho_c = \rho_t$, i.e., the covariance of the predictor with the outcome is the same under treatment or without it. In this case, $\widehat{\lambda^*}$ is estimated with $\frac{\widehat{Cov}(Y_i, f(X_i))}{\widehat{\sigma_f^2}}$ where the numerator is the unbiased pooled covariance estimator and the denominator is the standard unbiased estimator of the variance.

Table 2 highlights in red the estimators that are problematic, being either highly biased or exhibiting an overly high variance. It is the case systematically for ANCOVA for $n = 2$, that is, for a control group of size two, which is an extreme example. PPCT is also problematic for $n = 2$ when the response is Chi-squared distributed with 5 degrees of freedom. The Reference, PPI, and PPCT2 are always stable.

Excluding the case $n = 2$, we observe that under normally distributed data, none of the estimators appear to be biased. Then, in the case where the predictor is weakly correlated with the outcome ($\rho \leq 0.5$), the PPCT estimators are competitive, and more specifically, the PPCT2 estimator almost always provides the least variable estimator.

In the case of balanced arms and even at $n = 50$, ANCOVA, PPCT and PPCT2 perform well and quite similarly.

With a skewed distribution, here the Chi-square with 5 degrees of freedom, we can observe deviations from the targeted $ATE = 1$ value at low control group sample sizes ($n \leq 5$) for the PPCT estimator. However, PPCT2 fixes this issue and consistently outperforms the PPCT and ANCOVA. Lastly, in a highly correlated scenario ($\rho = 0.95$), the PPI estimator performs the best. However, this estimator is relatively poor if the predictor is less correlated with the outcome.

When using the Normal approximation of the test statistic, the pointwise ATE estimators are not enough to decide whether the treatment effect is significant or not; we also require the variance of these estimators. We use the following variance estimators: for the reference $\hat{\mathbb{V}}(\hat{ATE}_{ref}) = \frac{\hat{\sigma_c}^2}{n} + \frac{\hat{\sigma_t}^2}{m}$, for the PPCT and PPCT2 $\hat{\mathbb{V}}(\widehat{ATE}^{PPCT}) = \frac{\hat{\sigma_t}^2 + \hat{\lambda^*}^2 \hat{\sigma_f}^2 - 2\hat{\lambda^*}\hat{\sigma_f}\hat{\sigma_t}\hat{\rho_t}}{m} + \frac{\hat{\sigma_c}^2 + \hat{\lambda^*}^2 \hat{\sigma_f}^2 - 2\hat{\lambda^*}\hat{\sigma_f}\hat{\sigma_c}\hat{\rho_c}}{n}$ and we use the sandwich estimator for the least square regression of the ANCOVA. The results for the variance estimators are shown in Table C1 in the supplementaries.

From the results in the normal setting, we gather that low sample sizes in the control arm, i.e., $n = 2, 5$, tend to lead to a relatively high variance for the ANCOVA estimator. We also notice that in those extreme cases, the PPCT estimator tends to largely underestimate its true variance, which would lead to a high false discovery rate (FDR). Correcting with a better λ^* estimator as in PPCT2 yields an unbiased variance estimator. The PPCT2 variance estimator also consistently has lower variance than the PPCT and ANCOVA at lower sample sizes. At large sample sizes, however, PPCT, PPCT2, and ANCOVA have similar variance estimators, which match the best possible estimators (i.e., reference in the random predictor case $\rho = 0$ and PPI in the quasi-perfect predictor case $\rho = 0.95$).

Those results suggest that for balanced designs, PPCT and ANCOVA behave similarly. However, for designs enabled by the advent of prognostic models, the settings with low samples in the control arm are interesting. Here, some differences between the estimators appear. Namely, we see that PPCT would lead to a high FDR, while ANCOVA requires at least five patients in the control arm to be unbiased. However,

the choice of estimator for λ^* in PPCT is of importance as showcased by PPCT2, which performs the best in terms of variance estimation at low n values while remaining unbiased. However, remember that the experimental design is here limited to $\sigma_c = \sigma_t$.

3.2 Simulated clinical trial data

In those simulation experiments, we used the adsim R package [22]. The model used in this library [23] has been calibrated on AD patient data and provides close to real-life clinical trial setting simulations. The outcome of interest is the ADAS-Cog-13 for AD Assessment Scale- cognitive, a clinical score between 0 and 70, increasing with cognitive decline. The generative model for this score is Beta distributed:

$$Y_i/70 \sim Beta(\alpha = \psi_i\tau, \beta = (1-\psi_i)\tau), \text{ where } \psi_i = \Big(1+\exp\big(-(\eta_i+\alpha_i\times t+\delta\times T_i)\big)\Big)^{-1} \tag{18}$$

where η_i and α_i are random effects with normal distributions (generated by latent covariates), t is the length of the trial, δ is the effect of the treatment on the logit of the ADAS-Cog-13, ψ_i is the estimated underlying progression model for the ADAS-Cog-13, and τ is the precision of the Beta noise added to the model to generate the observations. In the following experiments, t is fixed at 1 year, τ and the hidden fixed effects are set at the mean values available in the adsim package. We use a 50/50 randomization between the treated and placebo arms. The parameters are fixed to be as close as possible to real life, and small perturbations did not lead to meaningful changes in the experiments. A sample of trajectories, both under the control and treatment arm is shown on figure 2 with the parameters used in the simulation and described below.

The only variables in the simulation model that we vary are the total number of subjects N and the precision of the Beta-distributed noise on the outcome τ.

To simulate a predictor, for each individual i, we use the underlying natural model $f(X_i) = \psi_i(\eta_i, \alpha_i; T_i = 0) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon)$ allows us to control the predictor's performance. Adding a constant bias did not influence the results since both PPI/PPCT and ANCOVA are adjusted linearly, so they account for any systematic bias.

We first simulated trials with $N = 100$ patients, with a treatment size being the one calibrated on the Donepezil in Alzheimer's [23], $\delta = 0.1382$ which is equivalent to an average treatment effect of minus two points over one year on the outcome scale: $ATE \simeq -1.96$. We simulated 1000 trials and gradually increased σ_ϵ to make predictions more noisy. We computed the power as the mean number of trials where the null hypothesis was rightly rejected. We also derived the theoretical sample size required for a trial equally powered as the standard design, i.e., using the reference estimator. The results are presented in Figure 3.

We observe a gain of power for the empowered estimators PPCT and ANCOVA, which equates to a smaller sample size for the clinical trial. When the predictor is too noisy, the PPI estimator performs worse than the standard design. The PPCT and ANCOVA estimators are never significantly different, indicating their highly similar performance. They are always preferred to the standard design for this metric. The

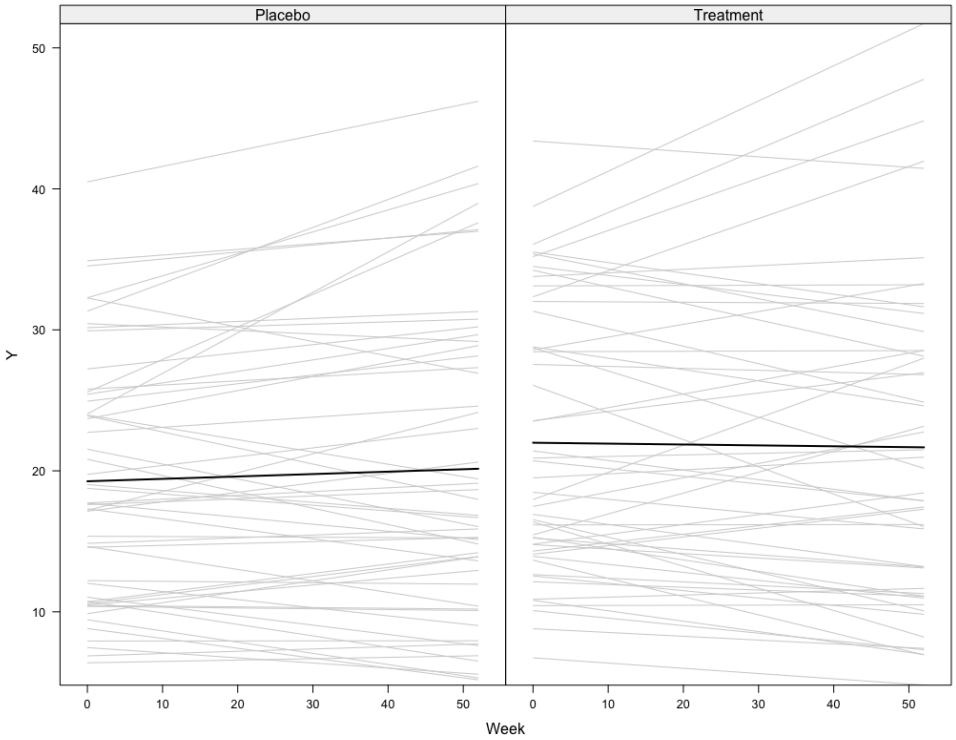

Fig. 2 Trajectories of the ADAS-Cog-13 score progression, simulated for a sample of 100 placebo and treated patients across time, with a treatment size of $\delta = 0.2$.

power increase and the sample size reduction are more pronounced when the predictions correlate more with the outcome. All these phenomena are consistent with the theory presented in the previous sections.

The PPCT with $\hat{\lambda}^*$ and ANCOVA estimators are consistent but not necessarily unbiased for a finite sample. Indeed, for PPCT, note that λ^* is estimated with the trial data. We provide an additional simulation experiment to evaluate the bias of the estimators. We varied the size of the clinical trial simulations, with a fixed noise level of $\sigma_\epsilon = 2.0$ for the predictions. Figure 4 shows boxplots for all 4 methods at the sample sizes $N = 10, 100, 200, 1000$. In all cases, the bias is negligible.

Simulations showcase the two main applications of the PPCT estimator: increase the power of a clinical trial or reduce the sample size at the design stage. Experiments with close to real-life data show that the estimated treatment effect has no bias.

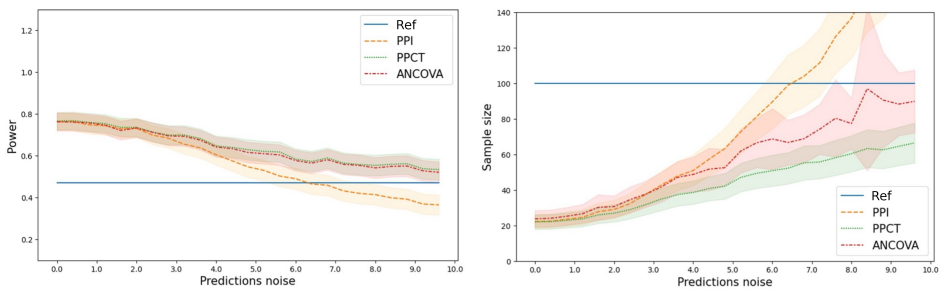

Fig. 3 Left: Power of the estimators as a function of the prediction noise. Right: Sample size required for a clinical trial based on the different estimators as a function of the noise of the predictions. Noise is added on the native scale of the outcome (0-70). When the prediction noise is lower resp. higher than about 6.0, the PPI estimator is better resp. worse than the reference estimator. In contrast, PPCT and ANCOVA are always as or more powerful as the reference estimator. The reduction in sample size is more pronounced when the prediction is better (lower prediction noise).

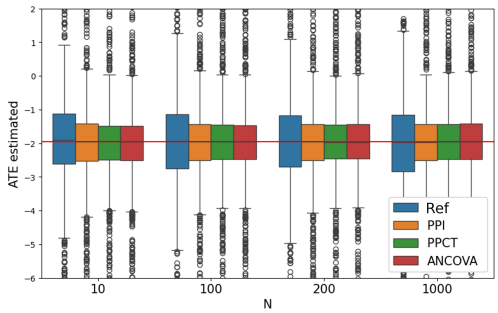

Fig. 4 Average Treatment Effect estimated by 4 methods as a function of the trial size over 1000 simulations. Each method provides negligible bias. PPI, PPCT, and ANCOVA show a reduced variance and thus improve upon the reference estimator.

3.3 Application to clinical trial data

3.3.1 Dataset overview

We now apply the estimators to a real-life AD clinical trial: the Hippocampus Study [24]. This one-year clinical trial showed that Donepezil significantly reduced the annualized rate of atrophy in the hippocampus. Even though the primary endpoint of the trial was the atrophy of the hippocampus measured by MRI and automatically segmented by a pre-defined pipeline (Freesurfer [25]), the trial also recorded secondary outcomes, among which the volume of the ventricles and two clinical assessments: the Mini-Mental State Examination (a clinical assessment of cognitive abilities) and the ADAS13 (which was also the outcome in the previous simulations). After data processing, the values were available for $N = 167$ subjects, with $n = 89$ placebo and $m = 78$ treated patients. A descriptive summary of the variables can be found in Table 3.

3.3.2 Predictors

For the prognostic scores, we used Disease Progression Models (DPM) calibrated on a natural disease history observational cohort.

Dataset:

We used the Alzheimers Disease Neuroimaging Initiative (ADNI) database. This longitudinal cohort records a wide range of modalities, from MRI to biomarkers to clinical assessments. Namely, all the markers from the Hippocampus Study are included in this database, meaning that we can learn the typical progression patterns of those features. We will then be able to predict the progression of the trial's patients using a well-calibrated DPM. The data's descriptive statistics can be found in Table 4. We included all patients with an AD diagnosis and at least two visits, which boils down to 730 subjects for a total of 4636 visits.

	MMSE		ADAS13		MRI Hippocampus ICV		MRI Ventricles ICV	
	Baseline	End	Baseline	End	Baseline	End	Baseline	End
count	167	167	167	167	167	167	167	167
mean	26.0	25.3	20.2	21.8	0.004	0.004	0.026	0.027
std	3.1	3.0	5.1	6.4	0.001	0.001	0.010	0.011
min	0	17	9	7	0.003	0.003	0.006	0.006
25%	25	23.5	16	18	0.004	0.004	0.019	0.020
50%	27	26	20	22	0.004	0.004	0.024	0.026
75%	28	27	24	25	0.005	0.005	0.030	0.033
max	30	30	35	47	0.006	0.006	0.067	0.071

Table 3 Summary of the Hippocampus Study values at baseline and end of trial. MMSE: Mini-Mental State Examination, cognitive assessment ranging from 0 (worse) to 30 (healthy); ADAS13: AD Assessment Scale, cognitive score ranging from 0 (healthy) to 70 (worse); MRI: Magnetic Resonance Imagery; ICV: intracranial volume, i.e. the ratio of the volume of the considered region (hippocampus or ventricles) over the total brain volume.

Disease Progression Models:

We calibrated three different models to predict the joint progression of the four selected markers (volume of hippocampus, volume of ventricles, MMSE and ADAS13):

1. Disease Course Mapping [7]: a non-linear mixed effect model based on Riemannian geometry. This model assumes a common pattern of progression in the population, determined by the population parameters, and that individuals deviate from it according to random effects captured by their parameters (for instance, an acceleration factor to distinguish between fast and slow progressors). We used the logistic curves Disease Course Mapping model, as it is well-suited for cognitive scores, such as the MMSE and ADAS13. The implementation of the model has been done with the Python library Leaspy
2. ODE-augmented-Lagrangian [8] is an algorithm that uses the augmented Lagrangian method of optimization for learning a non-parametric ordinary differential equation (ODE) of the form $\dot{x} = f(x)$. x is a vector of biomarkers - assuming

	MMSE	ADAS13	MRI Hippocampus ICV	MRI Ventricles ICV
count	3799	3623	3395	3392
mean	23.789	27.018	0.004	0.032
std	4.706	11.816	0.001	0.014
min	0	0	0.001	0.005
25%	22	19	0.003	0.022
50%	25	26	0.004	0.030
75%	27	33.330	0.004	0.040
max	30	85	0.007	0.090

Table 4 Summary of the ADNI variables. MMSE: Mini-Mental State Examination, cognitive assessment ranging from 0 (worse) to 30 (healthy); ADAS13: AD Assessment Scale, cognitive score ranging from 0 (healthy) to 70 (worse); MRI: Magnetic Resonance Imagery; ICV: intracranial volume, i.e. the ratio of the volume of the considered region (hippocampus or ventricles) over the total brain volume.

all the biomarkers are real-valued, $\dot{x}$ is the time-derivative of this vector, and f is a non-parametric function belonging to a reproducing kernel Hilbert space.

3. ODE-OCK [9] is the same model as for the ODE-augmented-Lagrangian optimized using the occupation kernel (OCK) formulation.

We then compute individualized predictions for the patients of the Hippocampus Study, using their values available at baseline and deducing the evolution at one year. However, this is a challenging task, as we only have one observation (at baseline), even though it is 4-dimensional, which is usually not enough to infer the underlying rate of progression. The prediction at one year is therefore less accurate than if we had visits for the patients prior to inclusion in the trial.

Results:

We considered all four markers as potential trial endpoints. The outcome value Y_i for a patient i was obtained as the difference between their value at the end of the trial and their value at baseline. As the theoretical study shows, the predictor works best when the prognostic factor is highly correlated with the outcome. The prognostic factors were therefore adjusted to be the predicted change in score over one year, meaning that we took the difference between the predictions of the DPM and the value of the patient at baseline.

We computed the estimators and their variance, and deduced the required sample size to have an equally powered clinical trial using formula provided in Proposition 9. The results are shown in table 5.

We observe a sample size reduction for the PPCT and ANCOVA estimators in the case of the MMSE and ADAS13 scores, as well as the MRI Ventricles. The best-performing models are the DCM for the MMSE and ADAS13 and the OCK for the MRI ventricles. The DCM model reaches an R^2 of 15% on the MMSE. This number means that we can either increase the power of the trial or reduce the sample size to obtain an equally powered trial. The DCM predictions' correlation is better if we only look at controls, $\rho_c = 0.625$, whereas it does not capture as well the evolution of treated patients, $\rho_t = 0.125$. This is coherent with the purpose of a DPM, as it

	MMSE	ADAS13	MRI Hippocampus	MRI Ventricles
Reference	100	100	100	100
PPI (DCM)	94	172	572	2639
PPCT (DCM)	84	**97**	100	100
ANCOVA (DCM)	**76**	99	103	103
PPI (OCK)	133	187	597	2307
PPCT (OCK)	100	100	100	**88**
ANCOVA (OCK)	103	101	103	91
PPI (Lagrangian)	127	204	599	2565
PPCT (Lagrangian)	100	98	100	100
ANCOVA (Lagrangian)	107	99	104	103

Table 5 Sample size required for an equally powered clinical trial using the estimators with three predicting models on the four endpoints of the Hippocampus Study. DCM: Disease Course Mapping model; OCK: occupation kernel model; Lagrangian: augmented Lagrangian ODE model. MMSE: Mini-Mental State Examination; ADAS13: AD Assessment Scale; MRI: Magnetic Resonance Imagery. Boldface highlights the minimum of each column when it is less than 100

accurately predicts the progression of placebo patients, while the model is not meant to describe the trajectories of patients under treatment.

The OCK method shows a sample size reduction on the ventricles, with an R^2 around 10%, $\rho_c = 0.337$ and $\rho_t = 0.375$. In practice, the correlation of the predictions with the outcome is greatly limited by the noise inherent to the measure. For the MRI measures, since the variability of the measure far outweighs the scale of the changes over one year, the R^2 of the predictions with the outcome is thus capped. The R^2 obtained did not differ much from 0 on the hippocampus, even with the models based on ODEs, which are less constrained than the DCM and have an easier time learning the progression of the imagery features.

From a statistical test standpoint, the reduction of the variance did not lead to a significant enough confidence interval reduction to change the conclusions of the trial on the modification of the MMSE nor the ADAS13, as seen in Figure 5. On the specific instance of the MMSE, we observe that the reduction in the confidence interval is counterbalanced by the estimator being closer to zero. For the hippocampus, the PPCT and ANCOVA do not change the estimator, thereby leading to the same conclusion as the primary analysis of the trial. The reduction of variance on the ventricles brought the estimator closer to statistical significance. We can imagine that with an improved prediction, for instance, if we had more points per patient prior to the baseline, the test could have rejected the null hypothesis.

4 Discussion

The newly introduced prediction-powered estimator for Clinical Trials (PPCT) is a powerful tool to enhance clinical trials. This method leverages historical data through a prognostic score and refines the estimator for the Average Treatment Effect. We have demonstrated that it is unbiased and decreases the variance of the estimator compared to the reference method. These properties have been validated on simulated clinical trial data. Our method has mainly two applications: either increasing the power of

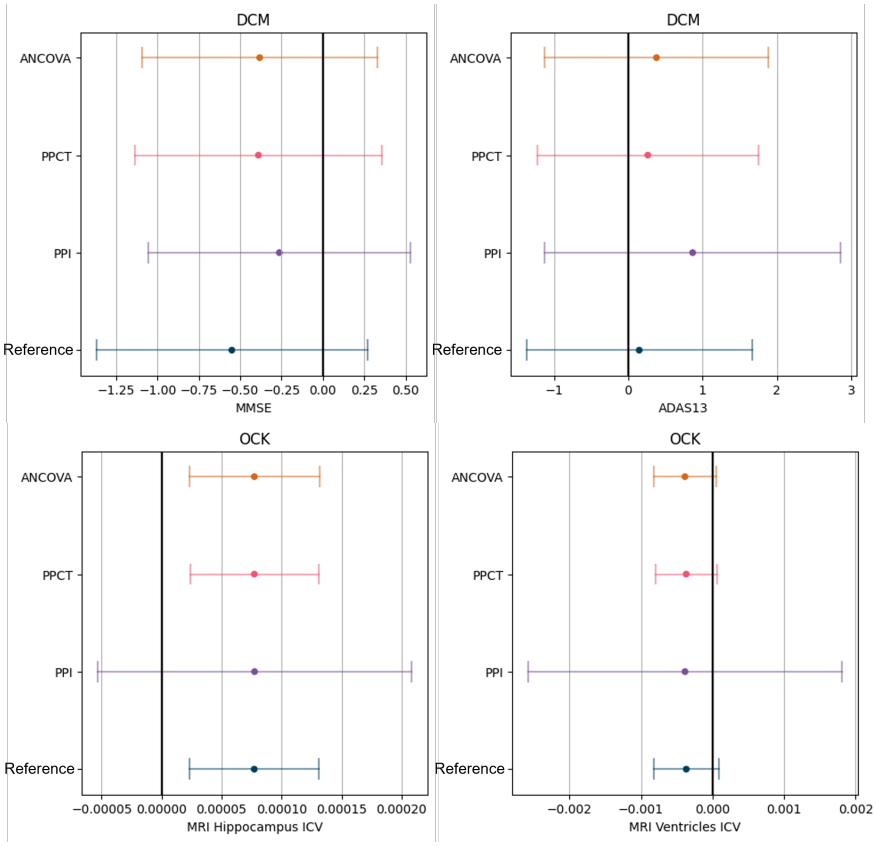

Fig. 5 Estimators with their 95% confidence interval on the 4 outcomes of the Hippocampus Study. The predictor used for the prediction-powered estimators and the ANCOVA is the Disease Course Mapping model for the two cognitive scores (MMSE and ADAS13) whereas we showed the results with the occupation kernel ODE method for the imagery markers (hippocampus and ventricles).

the clinical trial while maintaining the same type I error rate, or reducing the sample size at the design stage while retaining an equally powered RCT.

Our theoretical study shows that if we fix the tuning hyperparameter λ, the estimator is unbiased even at small sample sizes and has a known variance. When estimating this parameter using the trial data, the bias is shown empirically to be negligible. If we use assumption-free tests such as a permutation test, we do not require the estimator's asymptotic distribution to decide the trial's result. This can be particularly interesting when the trial size cannot guarantee the normality assumption. The estimator is very similar to the prognostic covariate adjustment method at large sample sizes. We derived the explicit formulation of the latter to show the differences, even if they are minor.

One point that needs to be addressed is whether the parameter λ should be fixed in advance of the statistical analysis, based on prior knowledge of the outcome and prognostic score correlations (from a pilot study or historical data), or computed "on

the fly" with a consistent estimator of the optimal λ as we did in our experiments. This allows for a better estimator but loosens the theoretical guarantees about unbiasedness. Our simulations show that the bias is negligible. However, this point needs further exploration. Other methods to determine the parameter λ^* might be better and should be investigated in the future. As shown in the simulations, different estimators of the λ^* parameters lead to very different results. For instance, the direct consistent estimator leads to underestimation of the variance in low sample sizes, which could lead to a high false discovery rate. On the other hand, a more adequate λ^* estimator yields a much better ATE estimator.

We also provided an analysis of the variance estimators of the different ATE estimators, which are essential to determining whether the trial showed a significant effect. Under large sample sizes, we showed that the variance of the estimators is all consistent, with the variance of the PPCT estimator being dependent on the choice of $\hat{\lambda}^*$. In the best case, the variance of the PPCT variance estimator can be lower than that of ANCOVA, which could lead to a higher power.

We chose to focus on the specific estimator for the Average Treatment Effect with a continuous outcome, as it encompasses the larger part of the RCT. However, we briefly introduced the general PPCT estimator, which is much more powerful. Further developments will allow us to apply this method to other critical estimators, for instance M-estimators, Cox regression, or even being able to adjust for covariates in the ATE estimation. This shows great promise for generalisation, as opposed to prognostic covariate adjustment, which can only be used in linearly adjusted estimation of the treatment effect on a continuous outcome.

The performance of the predictions is crucial, as a prognostic score highly correlated with the outcome can lead to large sample size reductions. Even without making too strong assumptions, we are able to derive sample size formulas to help decrease the sample size with accurate predictions. We showed that the prediction-powered framework favors unbalanced arms when the predictor is reliable. Indeed, one can interpret the predictions as synthetic control patients, thus lessening the need for actual controls in the trial. However, the method also performs sample size reduction in balanced trials, and this involves less risk at the design stage.

In our application to a real-life RCT, we obtained around $15 - 20\%$ variance (or equivalently, sample size) reduction on secondary endpoints. The estimator of the treatment effect on the primary endpoint was not affected, as the prediction models were not able to forecast the outcome precisely at the individual level. This shows that the method can have practical implications, but also that it is hard to obtain strong prognostic factors. The challenge is complex, as we need a model able to predict the change in score of a patient over one year. This can be translated as: we need to capture the slope of each patient, not solely its intercept. Furthermore, the prognostic score needs to correlate with individual variability, meaning that imputing the mean slope is not enough. Doing so with only baseline covariates is a daunting task.

We can think of various ways to improve on the proposed predictors, for instance, using ensemble methods such as Super Learner [26]. Another route would be to increase the information available for each patient, namely past observations, in order to have more reliable predictions. It is easier to learn the slope from several points.

One perk of the technique is that the estimator and the prognostic factors can be engineered separately, allowing the trial designer to make use of any complex machine learning system built by another as a "black-box" predictor. This is all the more relevant as supervised learning techniques are developing rapidly, and are otherwise complicated to use in clinical practice. For instance, we can mention Deep Learning models, which are notoriously hard to interpret and are thus shunned by clinicians.

Our proposed PPCT estimator has desirable properties for a RCT: smaller variance, unbiasedness, and a simple explicit formula. The method can either be used to reduce the cost of a clinical trial by shrinking its size or in a safer way by increasing the power of the trial.

Overall, this paper proposes an application of the prediction-powered inference technique to clinical trials, considering the case of linear covariate adjustment. The derived estimators of the average treatment effect are shown to be different but closely related to existing ones, including the ANCOVA estimator, providing an alternative point of view, but also possible advantages in unbalanced trials and long-tailed outcomes. PPI and PPI++ were originally presented in the context of M-estimators. We foresee that bringing this research to the setting of clinical trials will provide benefits in non-linear settings, including qualitative and time-to-event outcomes.

5 Declarations

5.1 Ethics approval and consent to participate

The ADNI data has been used in conformity with the ADNI terms of use detailed at `https://adni.loni.usc.edu/terms-of-use/`. The protocol of the 'Hippocampus study' and informed consent forms were approved by the ethics committee of Salpêtriere Hospital. The data has been used in conformity with the recommendations of this ethics committee.

5.2 Consent for publication

Not applicable

5.3 Availability of data and materials

The data used in this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu), launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD). The versions of the dataset used for this experiment were ADNI 1, 2, 3 and ADNI GO.

The data of the Hippocampus Study were obtained thanks to Bruno Dubois and collaborators and were first published in [24].

5.4 Competing interests

The authors do not declare any conflict of interest with the present study. The funding sources had no involvement in the study design, in the collection, analysis, and interpretation of data, in the report's writing, and in the decision to submit the article for publication.

5.5 Funding

The research leading to these results has received funding from the program "Investissements davenir" ANR-10-IAIHU-06. This work was also funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). The work at Portland State University was partly funded by the National Institute of Health RO1AG021155, R01EY032284, R01AG027161, and National Science Foundation #2136228. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimers Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimers Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

5.6 Authors' contributions

An author contribution table is given in table 6.

5.7 Acknowledgements

Not applicable

Contribution	PEP	MT	ST	SD	BMJ	BD
Conception	✓			✓	✓	
Design	✓			✓	✓	
Acquisition						✓
Analysis	✓	✓		✓	✓	
Interpretation	✓		✓	✓	✓	
Software creation	✓					
Draft or revised	✓	✓	✓	✓	✓	✓

Table 6 Contributions. PEP: Pierre-Emmanuel Poulet, MT: Maylis Tran, ST: Sophie Tezenas du Montcel, SD: Stanley Durrleman, BMJ: Bruno Michel Jedynak, BD: Bruno Dubois

References

[1] Angelopoulos, A.N., Bates, S., Fannjiang, C., Jordan, M.I., Zrnic, T.: Prediction-powered inference **382**(6671), 669–674 https://doi.org/10.1126/science.adi6000 . Publisher: American Association for the Advancement of Science. Accessed 2023-11-30

[2] Angelopoulos, A.N., Duchi, J.C., Zrnic, T.: PPI++: Efficient Prediction-Powered Inference. arXiv. https://doi.org/10.48550/arXiv.2311.01453 . http://arxiv.org/abs/2311.01453 Accessed 2024-01-19

[3] Maldonado, G., Greenland, S.: Estimating causal effects **31**(2), 422–429 https://doi.org/10.1093/intjepid/31.2.422 . Accessed 2024-10-15

[4] Signorovitch, J.E., Sikirica, V., Erder, M.H., Xie, J., Lu, M., Hodgkins, P.S., Betts, K.A., Wu, E.Q.: Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research **15**(6), 940–947 https://doi.org/10.1016/j.jval.2012.05.004

[5] Kaplan, D., Chen, J., Yavuz, S., Lyu, W.: Bayesian dynamic borrowing of historical information with applications to the analysis of large-scale assessments **88**(1), 1–30 https://doi.org/10.1007/s11336-022-09869-3 . Accessed 2024-10-15

[6] Young, A.L., Marinescu, R.V., Oxtoby, N.P., Bocchetta, M., Yong, K., Firth, N.C., Cash, D.M., Thomas, D.L., Dick, K.M., Cardoso, J., Swieten, J., Borroni, B., Galimberti, D., Masellis, M., Tartaglia, M.C., Rowe, J.B., Graff, C., Tagliavini, F., Frisoni, G.B., Laforce, R., Finger, E., Mendona, A., Sorbi, S., Warren, J.D., Crutch, S., Fox, N.C., Ourselin, S., Schott, J.M., Rohrer, J.D., Alexander, D.C.: Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference **9**(1), 1–16 https://doi.org/10.1038/s41467-018-05892-0 . Number: 1 Publisher: Nature Publishing Group. Accessed 2020-04-21

[7] Schiratti, J.-B., Allassonnire, S., Colliot, O., Durrleman, S.: A bayesian mixed-eects model to learn trajectories of changes from repeated manifold-valued observations, 33

[8] Lahouel, K., Wells, M., Rielly, V., Lew, E., Lovitz, D., Jedynak, B.M.: Learning nonparametric ordinary differential equations from noisy data **507**, 112971 https://doi.org/10.1016/j.jcp.2024.112971 . Accessed 2024-10-16

[9] Rielly, V., Lahouel, K., Lew, E., Wells, M., Haney, V., Jedynak, B.: Learning High-Dimensional Nonparametric Differential Equations Via Multivariate Occupation Kernel Functions. https://arxiv.org/abs/2306.10189v1 Accessed 2024-10-16

[10] Maheux, E., Koval, I., Ortholand, J., Birkenbihl, C., Archetti, D., Bouteloup, V., Epelbaum, S., Dufouil, C., Hofmann-Apitius, M., Durrleman, S.: Forecasting individual progression trajectories in alzheimers disease **14**, 761 https://doi.org/10.1038/s41467-022-35712-5 . Accessed 2024-10-16

[11] Ezzati, A., Lipton, R.B., Alzheimers Disease Neuroimaging Initiative: Machine learning predictive models can improve efficacy of clinical trials for alzheimer's disease **74**(1), 55–63 https://doi.org/10.3233/JAD-190822

[12] Burman, C.-F., Hermansson, E., Bock, D., Franzn, S., Svensson, D.: Digital twins and bayesian dynamic borrowing: Two recent approaches for incorporating historical control data https://doi.org/10.1002/pst.2376

[13] Schuler, A., Walsh, D., Hall, D., Walsh, J., Fisher, C., Disease, f.t.C.P.f.A., Initiative, t.A.D.N., Study, t.A.D.C.: Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score **18**(2), 329–356 https://doi.org/10.1515/ijb-2021-0072 . Publisher: De Gruyter. Accessed 2024-01-23

[14] Adjustment for Baseline Covariates in Clinical Trials - Scientific Guideline | European Medicines Agency (EMA). https://www.ema.europa.eu/en/adjustment-baseline-covariates-clinical-trials-scientific-guideline Accessed 2024-10-16

[15] Li, L.: Prognostic factor analyses. In: Piantadosi, S., Meinert, C.L. (eds.) Principles and Practice of Clinical Trials, pp. 1771–1787. Springer. https://doi.org/10.1007/978-3-319-52636-2_121 . https://doi.org/10.1007/978-3-319-52636-2_121 Accessed 2024-10-16

[16] Schultz, J., Prince, J.L., Jedynak, B.M.: Predictive powered inference for healthcare; relating optical coherence tomography scans to multiple sclerosis disease progression. In: Machine Learning for Healthcare Conference (2024). PMLR

[17] Holzhauer, B., Adewuyi, E.T.: super-covariates: Using predicted control group outcome as a covariate in randomized clinical trials. Pharmaceutical Statistics **22**(6), 1062–1075 (2023)

[18] Zelen, M.: The randomization and stratification of patients to clinical trials **27**(7), 365–375 https://doi.org/10.1016/0021-9681(74)90015-0 . Accessed 2024-07-03

[19] Pocock, S.J., Simon, R.: Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial **31**(1), 103–115 https://doi.org/10.2307/2529712 . Publisher: [Wiley, International Biometric Society]. Accessed 2024-07-03

[20] Good, P.: Permutation Tests: a Practical Guide to Resampling Methods for Testing Hypotheses. Springer, ??? (2013)

[21] Marinescu, R.V., Oxtoby, N.P., Young, A.L., Bron, E.E., Toga, A.W., Weiner, M.W., Barkhof, F., Fox, N.C., Golland, P., Klein, S., Alexander, D.C.: TADPOLE challenge: Accurate alzheimer's disease prediction through crowdsourced forecasting of future data **11843**, 1–10 https://doi.org/10.1007/978-3-030-32281-6_1

[22] Polhamus, D., Rogers, J., Gillespie, B., French, J., Gastonguay, M.: From evidence synthesis to trial optimization: The adsim package for model-based simulation in alzheimers disease. In: PAGE. Abstracts of the Annual Meeting of the Population Approach Group in Europe. ISSN, vol. 6032, p. 2012 (1871)

[23] Rogers, J.A., Polhamus, D., Gillespie, W.R., Ito, K., Romero, K., Qiu, R., Stephenson, D., Gastonguay, M.R., Corrigan, B.: Combining patient-level and summary-level data for alzheimer's disease modeling and simulation: amyloid beta regression meta-analysis **39**(5), 479–498 https://doi.org/10.1007/s10928-012-9263-3

[24] Dubois, B., Chupin, M., Hampel, H., Lista, S., Cavedo, E., Croisile, B., Louis Tisserand, G., Touchon, J., Bonafe, A., Ousset, P.J., Ait Ameur, A., Rouaud, O., Ricolfi, F., Vighetto, A., Pasquier, F., Delmaire, C., Ceccaldi, M., Girard, N., Dufouil, C., Lehericy, S., Tonelli, I., Duveau, F., Colliot, O., Garnero, L., Sarazin, M., Dormont, D., Group, .S.: Donepezil decreases annual rate of hippocampal atrophy in suspected prodromal alzheimer's disease **11**(9), 1041–1049 https://doi.org/10.1016/j.jalz.2014.10.003 . _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1016/j.jalz.2014.10.003. Accessed 2024-02-29

[25] Fischl, B.: FreeSurfer **62**(2), 774 https://doi.org/10.1016/j.neuroimage.2012.01.021 . Publisher: NIH Public Access. Accessed 2024-10-16

[26] Pirracchio, R., Petersen, M.L., Laan, M.: Improving propensity score estimators' robustness to model misspecification using super learner **181**(2), 108–119 https://doi.org/10.1093/aje/kwu253