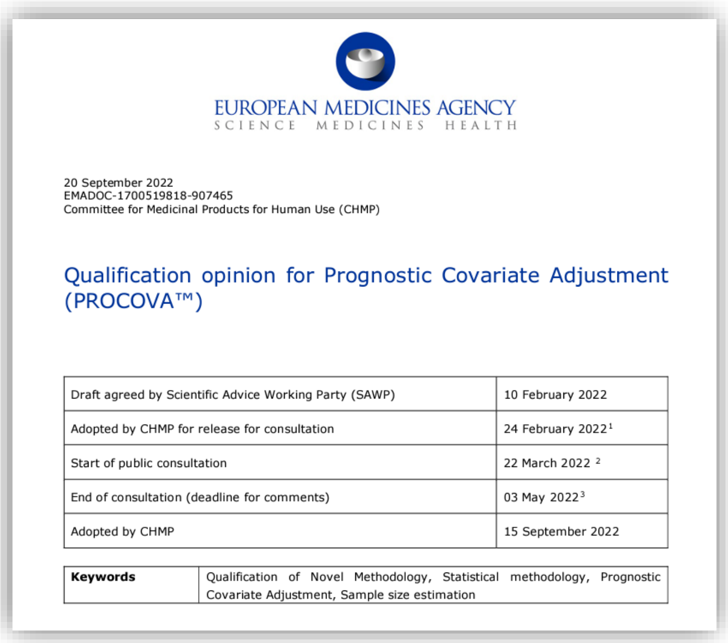
Enhancing Treatment Effect Estimation in Clinical Trials using Machine Learning: A Within-Study Prognostic Score Approach

Antigoni Elefsinioti (Bayer)
Maike Ahrens (Chrestos)
Sebastian Voss (Chrestos)
Karl Koechert (Sanofi)
Bohdana Ratitch (Bayer)

London, 11/06/2025

CHMP qualifies PROCOVA as prognostic score adjustment



EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

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

Qualification opinion for Prognostic Covariate Adjustment (PROCOVA™)

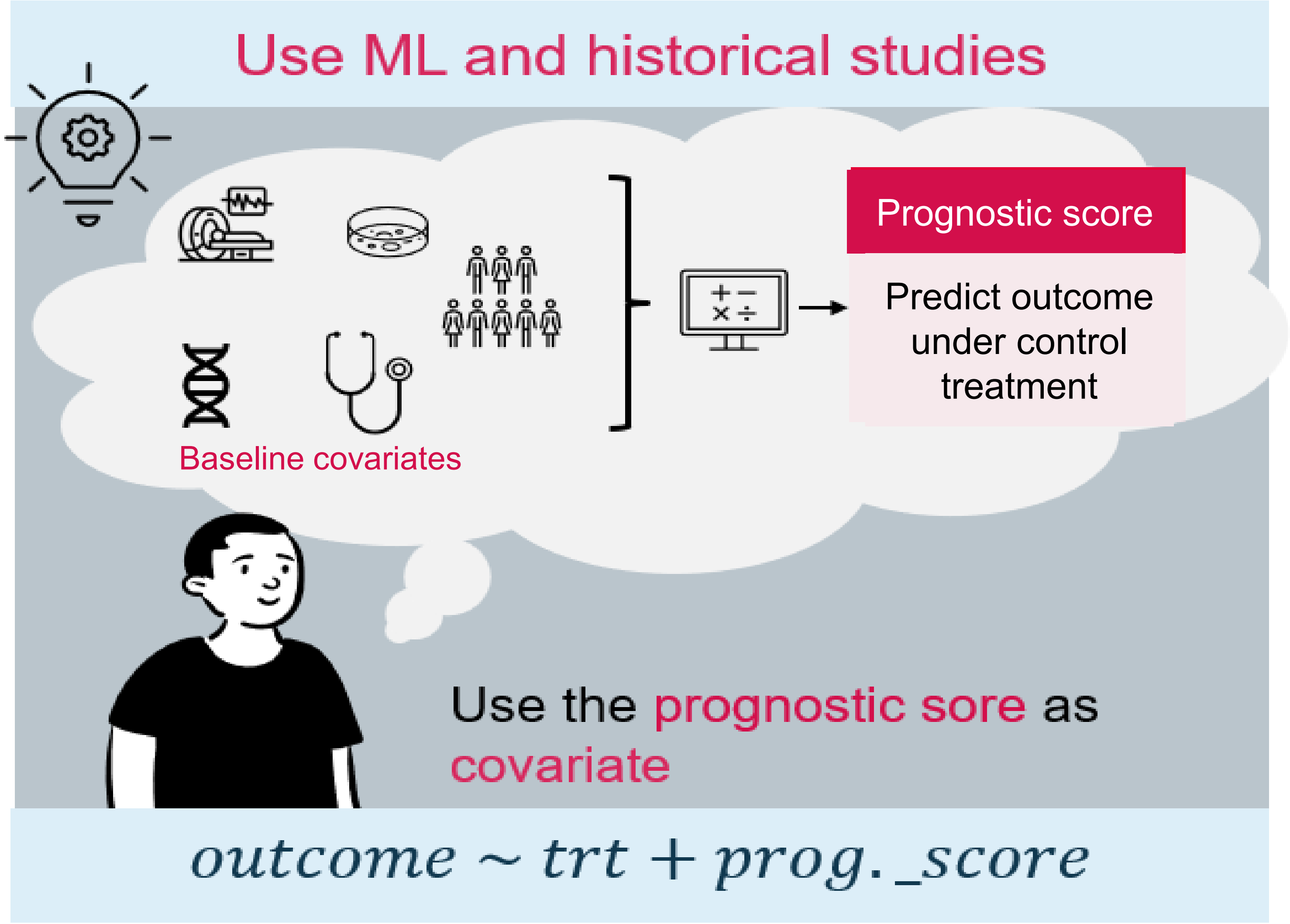
Draft agreed by Scientific Advice Working Party (SAWP)	10 February 2022
Adopted by CHMP for release for consultation	24 February 2022[1]
Start of public consultation	22 March 2022 [2]
End of consultation (deadline for comments)	03 May 2022[3]
Adopted by CHMP	15 September 2022

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

- Motivation **- efficiency gain** of **Phase 2/3** studies when estimating **treatment effects**
- Utilize **Machine Learning** techniques and **historical data** to develop **prognostic models**
- **Condense** the **prognostic information** from multiple **baseline covariates**

Focus: Linear regression without interactions

Condense the prognostic vc98c information from multiple baseline covariates



Use ML and historical studies

Baseline covariates

Prognostic score

Predict outcome under control treatment

Use the prognostic sore as covariate

$$outcome \sim trt + prog.\_score$$

Limitations of historical data use for prognostic modeling

- No matching historical datasets
- Difficulties in data harmonization
- Cost of acquiring data
- Population drift over time
- Quality of observational vs RCT historical data

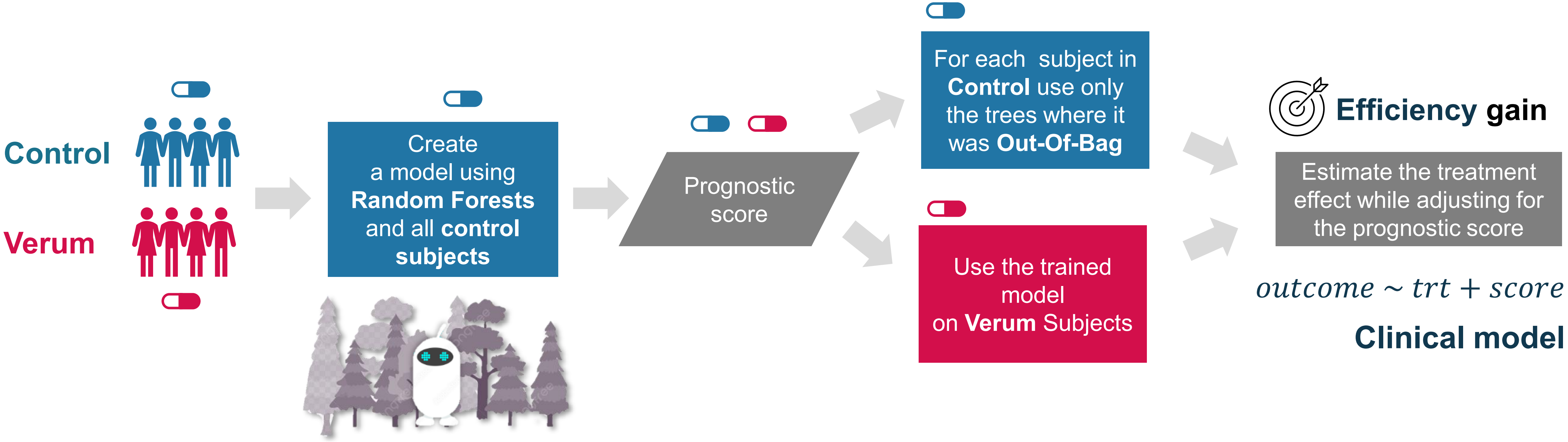
Limitations of historical data use for prognostic modeling

CHMP/EMA Qualification (2022):

- “**A drawback of PROCOVA**, however, is that the prognostic score must be prespecified including a scale factor, and weights used within the score **cannot be adjusted** to possible **differences** between the **training** setting and the **actual trial** setting.

- CHMP **Response** to a comment: “We agree that **further work comparing PROCOVA** to **methods that do not use historical data** (e.g. by using machine learning predictions on data for the control arm of the trial itself) would be an **interesting and valuable task.”**

Our approach focuses on estimating prognostic models from within-study data



Control

Verum

Create a model using **Random Forests** and all **control subjects**

Prognostic score

For each subject in **Control** use only the trees where it was **Out-Of-Bag**

Use the trained model on **Verum** Subjects

Efficiency gain

Estimate the treatment effect while adjusting for the prognostic score

$outcome \sim trt + score$

Clinical model

Prognostic scores for each participant — including those in the placebo arm — are **derived** from **models** trained on **independent datasets**, thus **mitigating biases** related to **model selection.**

Simulation study on prognostic score adjustment

GOAL

Assess the **properties of prognostic covariate adjustment** in a controlled environment for different scenarios of interest

SPECIAL FOCUS

Comparison of the **independent study** and the **within-study** approach

General effects on treatment effect estimation in linear models

- Change in precision
- Potential bias
- Effect on type I error

Simulation scenarios

Fixed simulation parameters

- **Number of subjects** in the target study (125 in placebo, 250 in treatment)
- **Prognostic factors** in the training study

Simulation scenarios

Varying simulation parameters

Simulation parameter	Level 1	Level 2	Level 3
Number of placebo subjects in the training study	2000	500	125
Performance of the ML model in the training study[1)]	high	mediocre	poor
Prognostic factors in the target study	same as in training	partially the same	completely different
Treatment effect size in the target study[2)]	80% power	50% power	no effect

1) controlled by the proportion of variability explained by the prognostic factors
2) effect size calibrated to the given power in the unadjusted model

Combinations of these parameters lead to **81 simulation scenarios**, each with **1000 iterations** for a total of **162.000 ML models** tuned and trained.

Simulation approach

Steps per simulation iteration

1 Draw synthetic data sets

2 Train a prognostic ML model (random forest)

(use control arms only)

3 Calculate prognostic score for the subjects in the target study

4 Fit linear models to the target study to assess the treatment effect →

Compared linear models

Score variations:

$y \sim \text{treatment} + \text{score}_{independent}$

$y \sim \text{treatment} + \text{score}_{within}$

Reference models:

$y \sim \text{treatment}$

$y \sim \text{treatment} + \text{covariates}_{oracle}$

The last model assumes perfect knowledge of the DGP and can be considered the best case for the respective scenario.

Simulation results

Figure introduction

Model

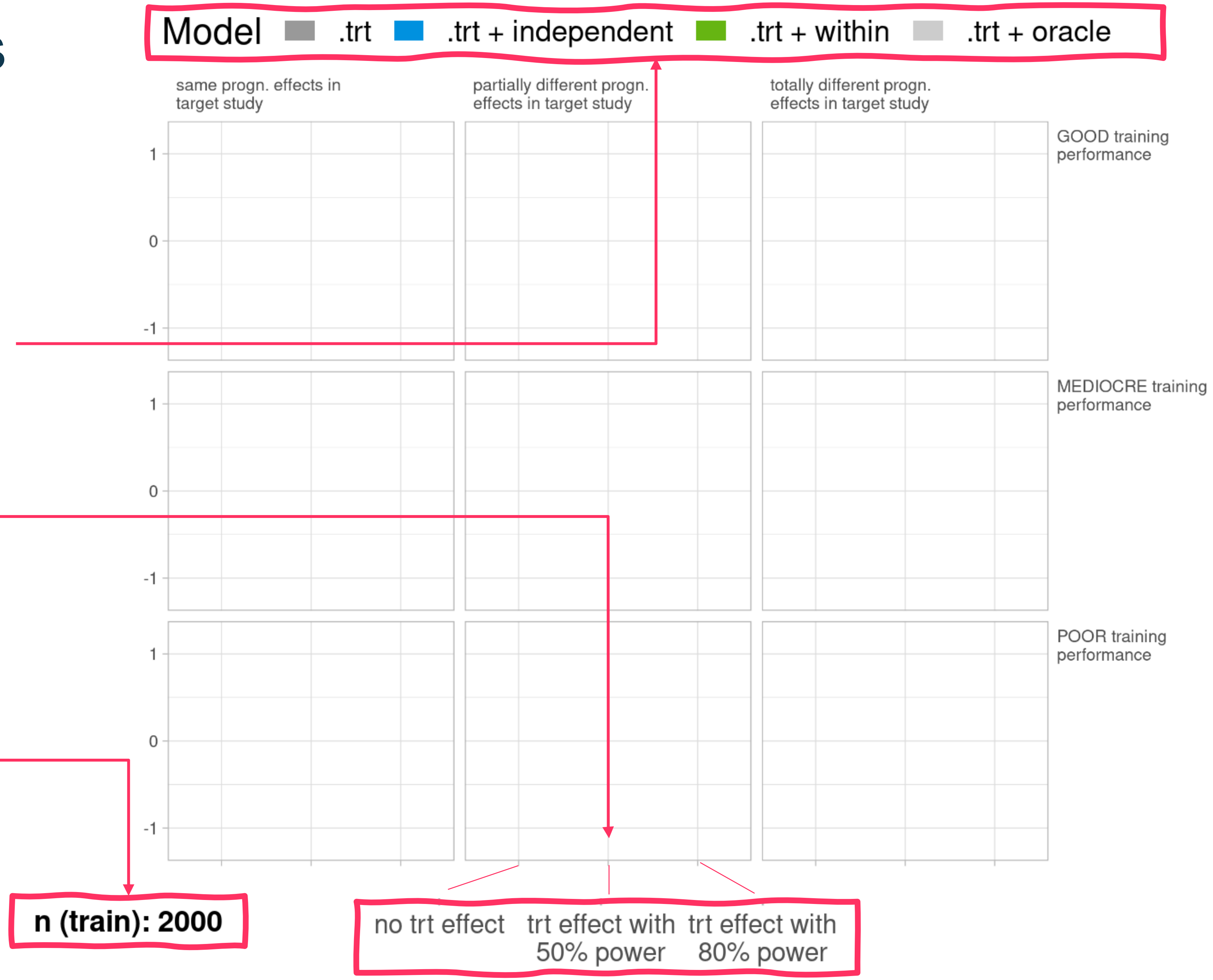
indicated by **color** in each panel

Treatment effect size

on the **x-axis** of each panel

Training population size

will be kept **fixed** per figure

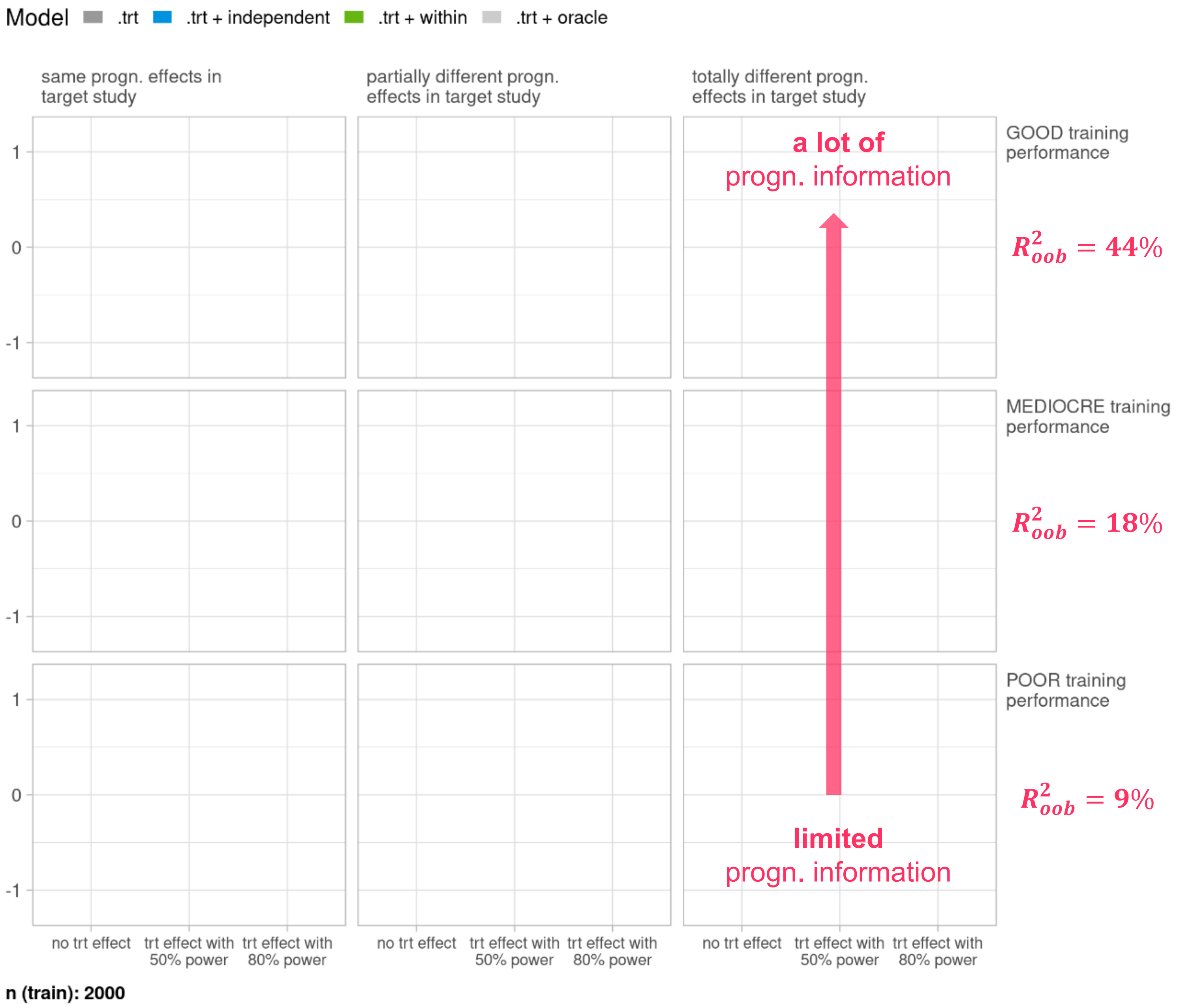


Simulation results

Figure introduction

ROWS

Amount of prognostic information
in the baseline data of the training study (reflected by training performance of the ML model)

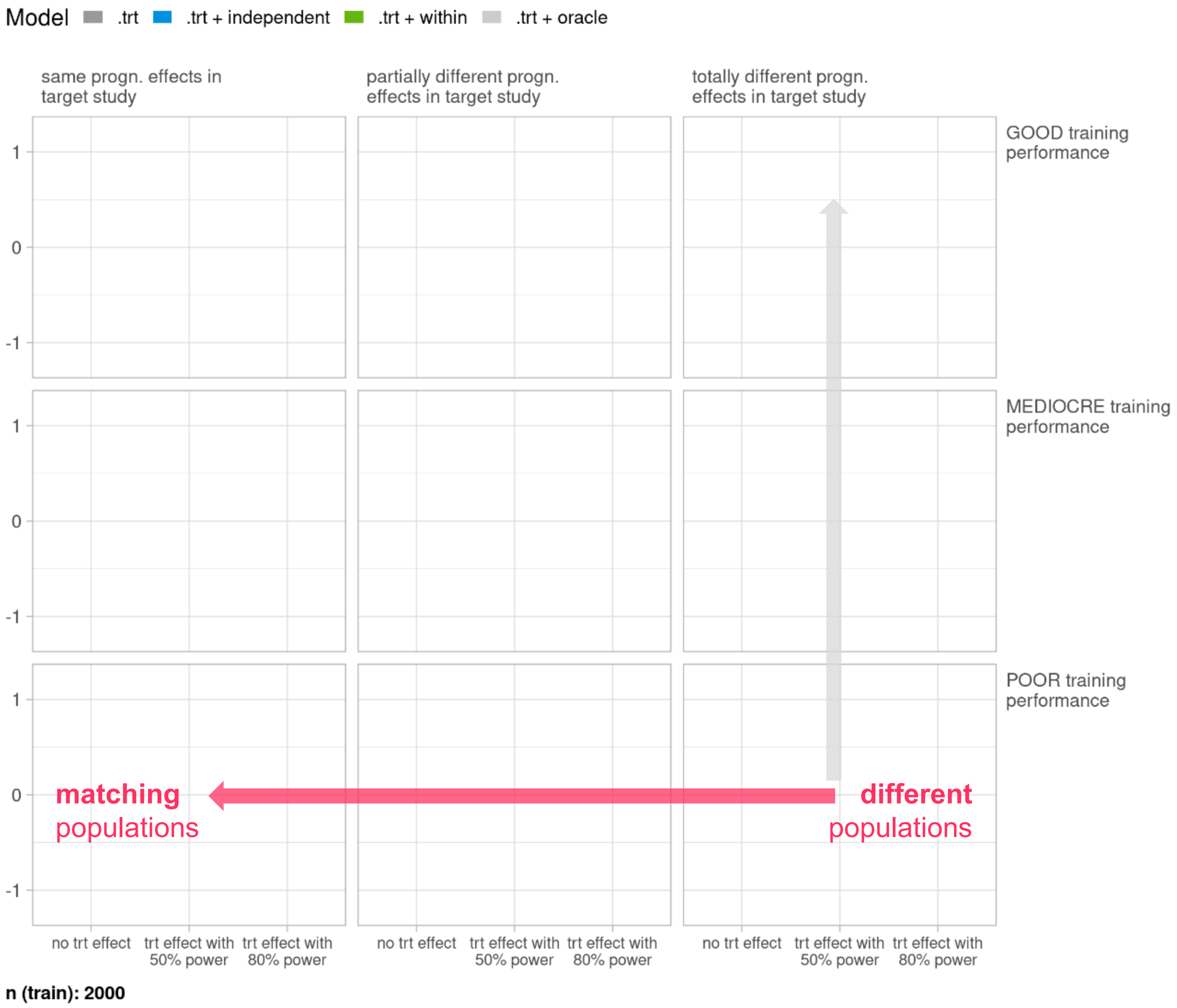


Simulation results

Figure introduction

COLUMNS

Similarity of populations in training and target study w.r.t. prognostic factors (only relevant for **independent study** approach)



Simulation results

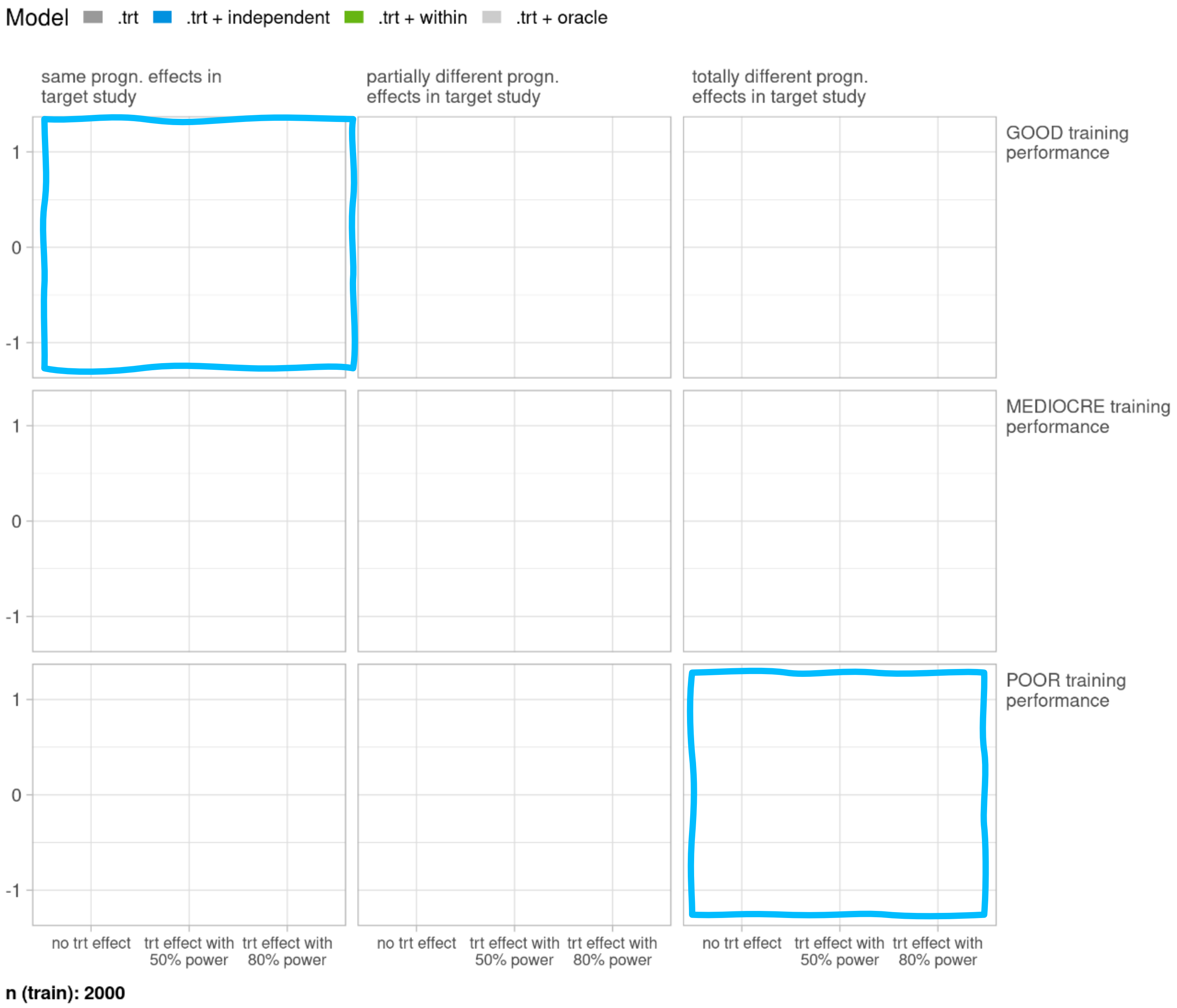
Figure introduction

Best case

- **good** training performance
- **similar** populations in training and target study

Worst case

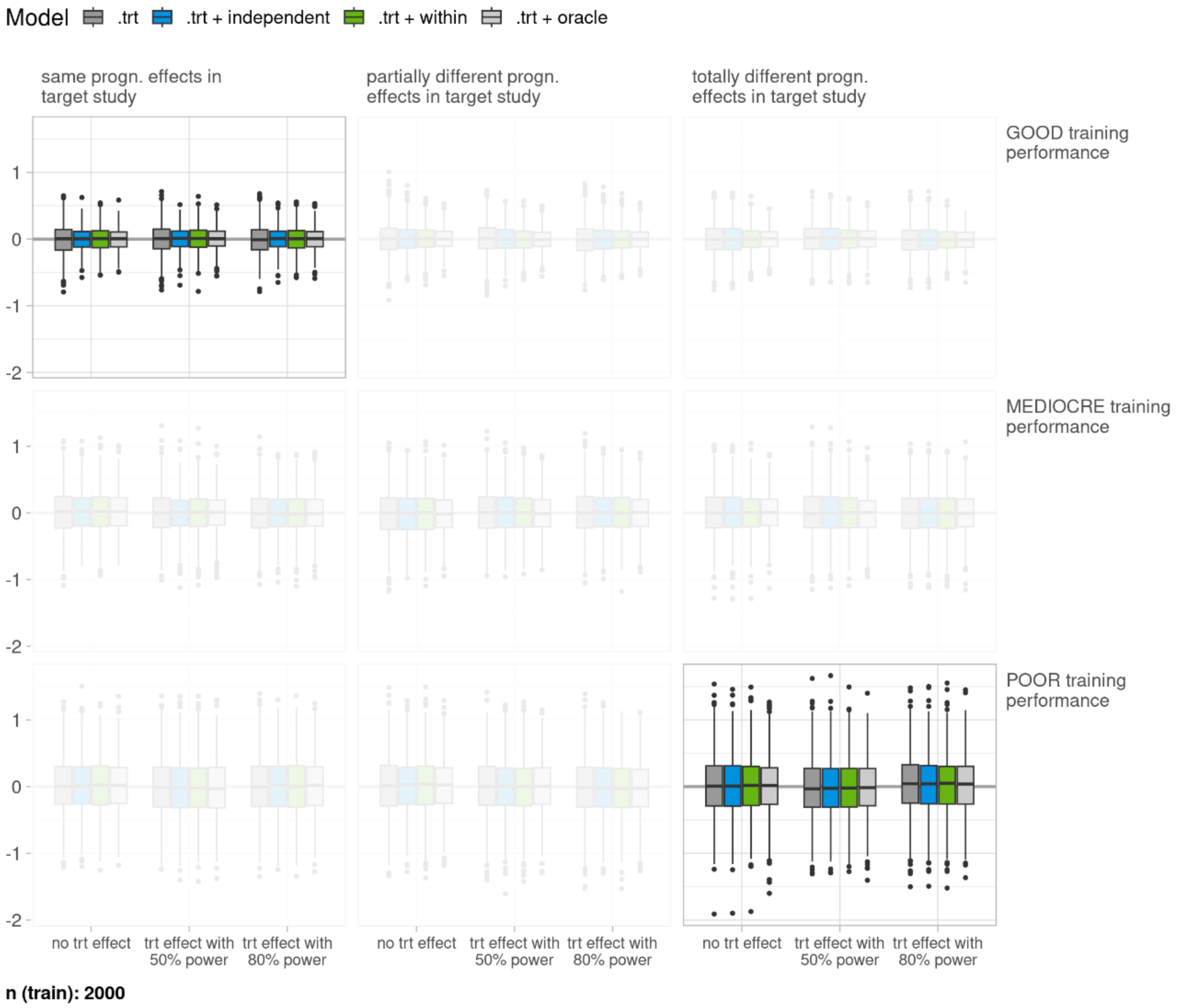
- **poor** training performance
- **different** populations in training and target study



Unbiasedness

Difference between estimated and real treatment effect

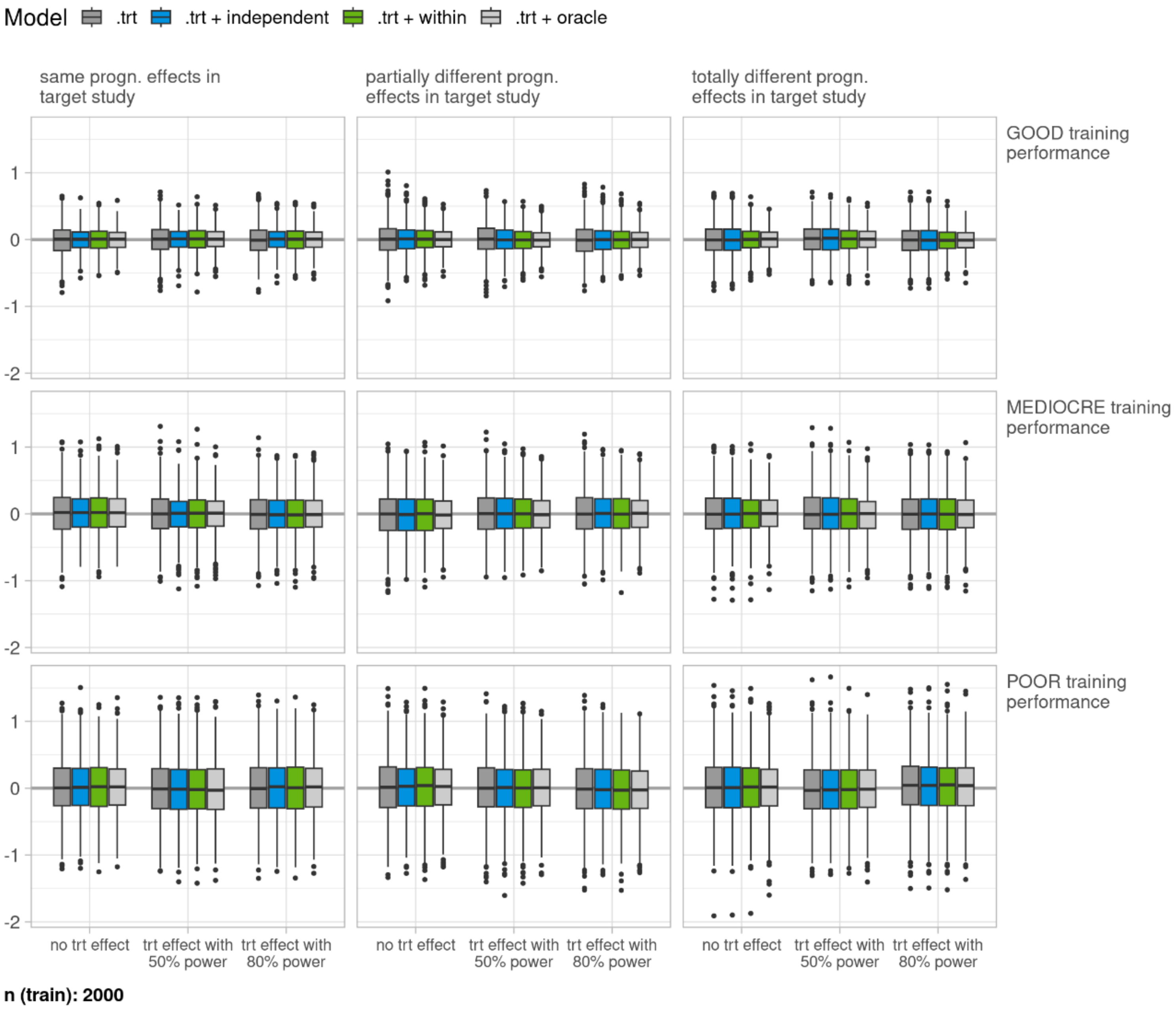
- Difference of estimator and true effect varies symmetrically around zero



Unbiasedness

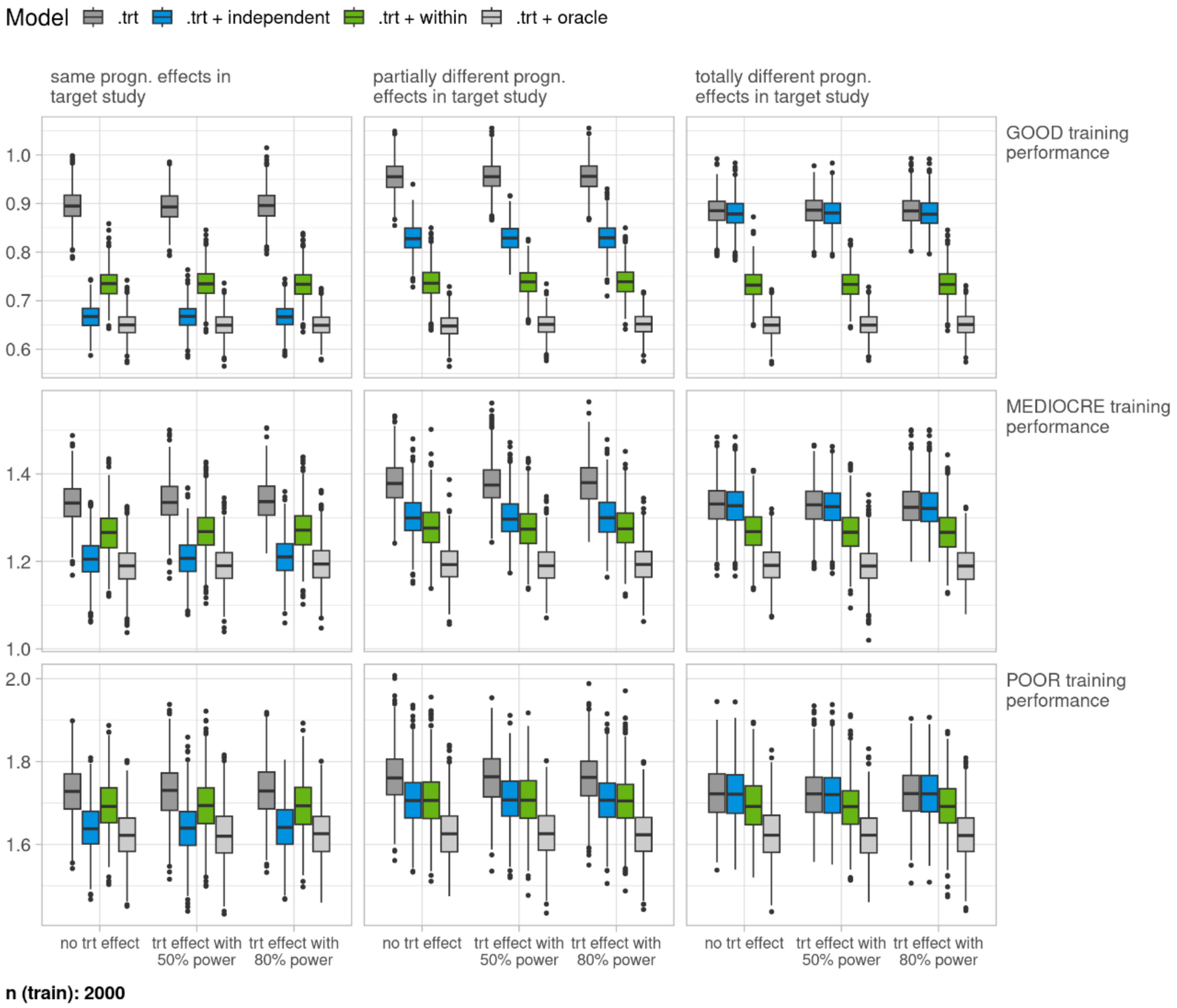
Difference between estimated and real treatment effect

- Difference of estimator and true effect varies symmetrically around zero
- No systematic bias in any of the scenarios



Precision

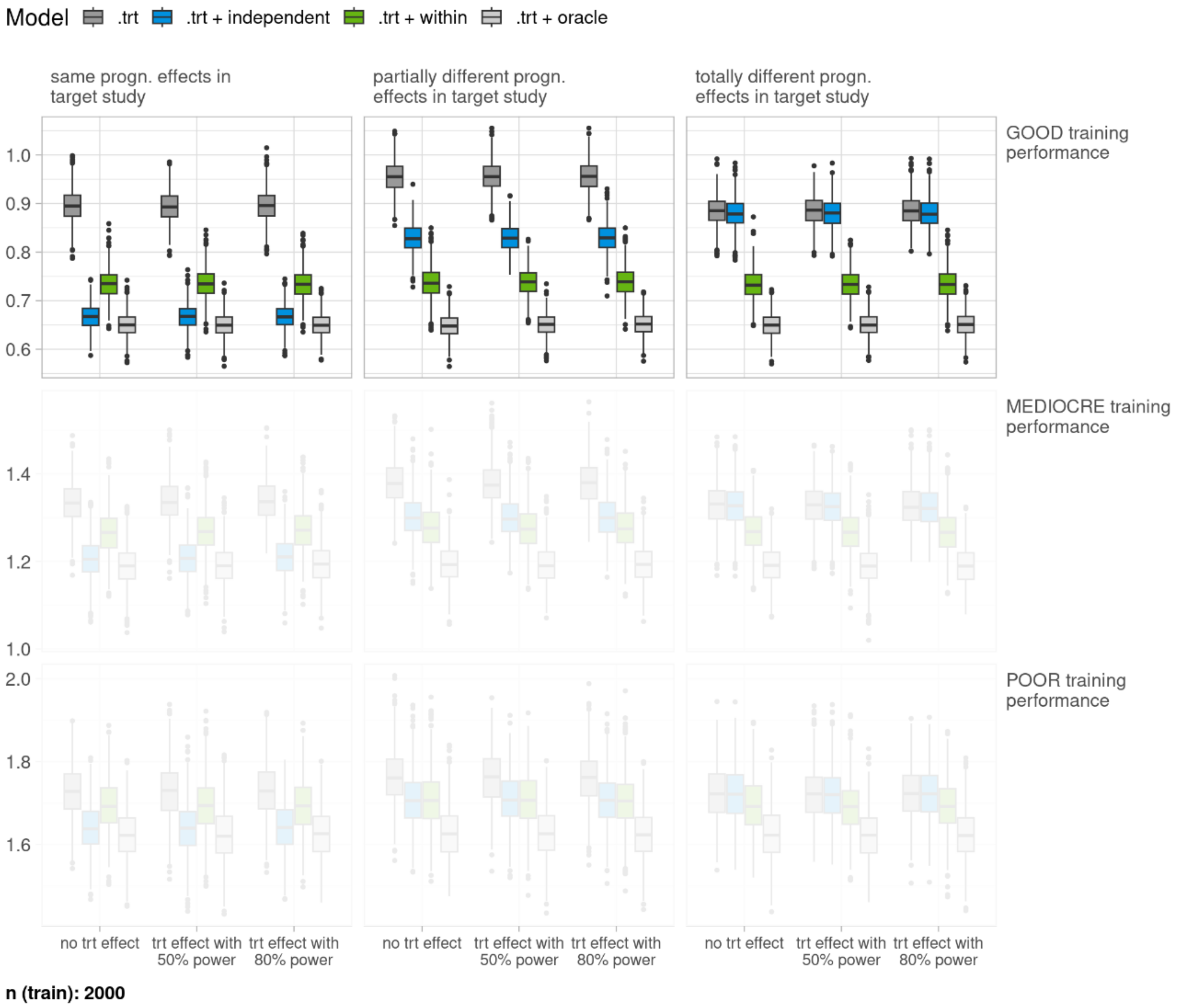
Widths of 95% CI for the treatment effect estimate



Precision

Widths of 95% CI for the treatment effect estimate

- Benefit of **independent study** score ranges **depending on similarity of populations**
- **Within-study** is **independent by design**



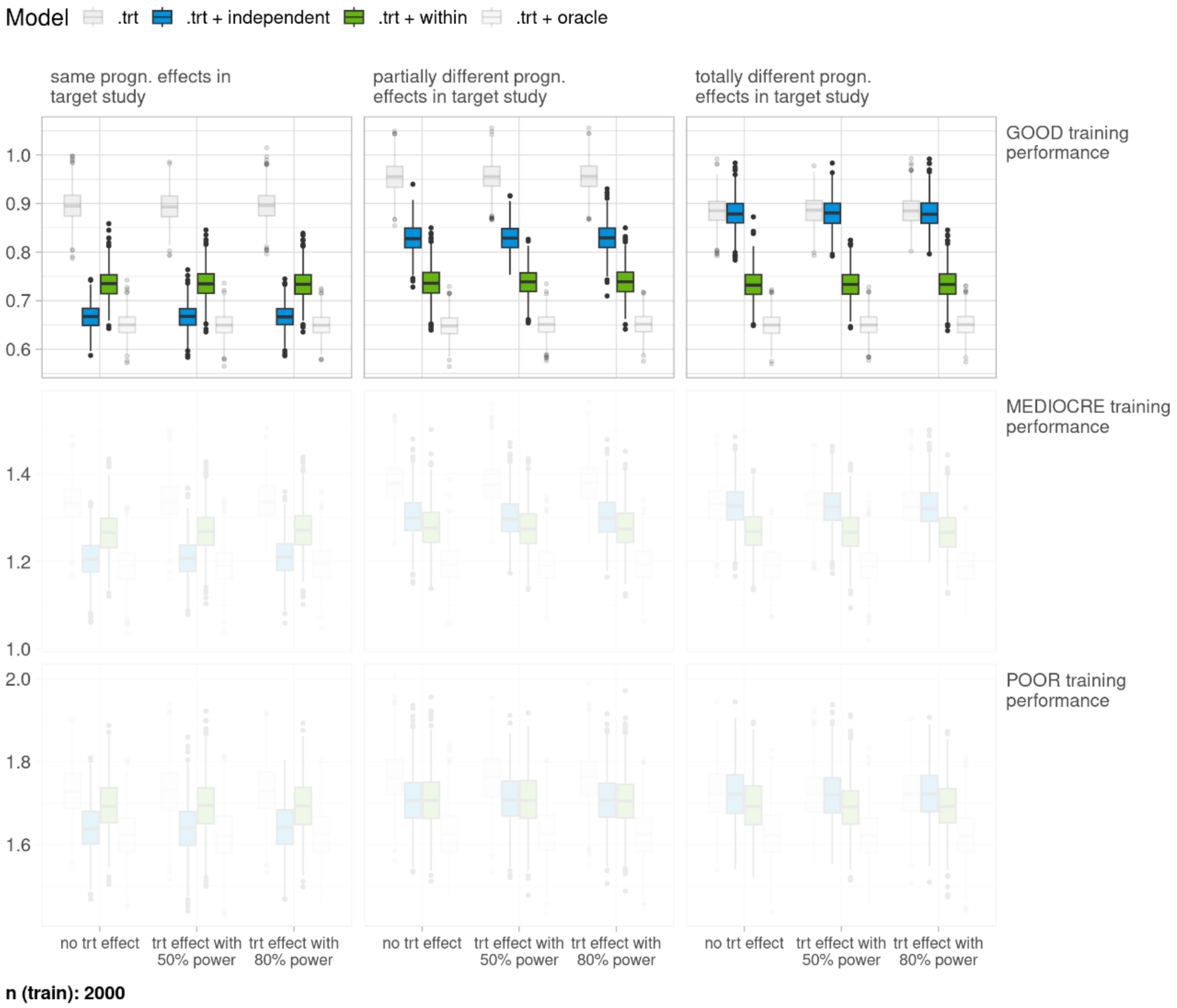
Precision

Widths of 95% CI for the treatment effect estimate

- Ranking of **within-study** and **independent study** approach depends on **similarity of studies:**

Smaller CI width with the

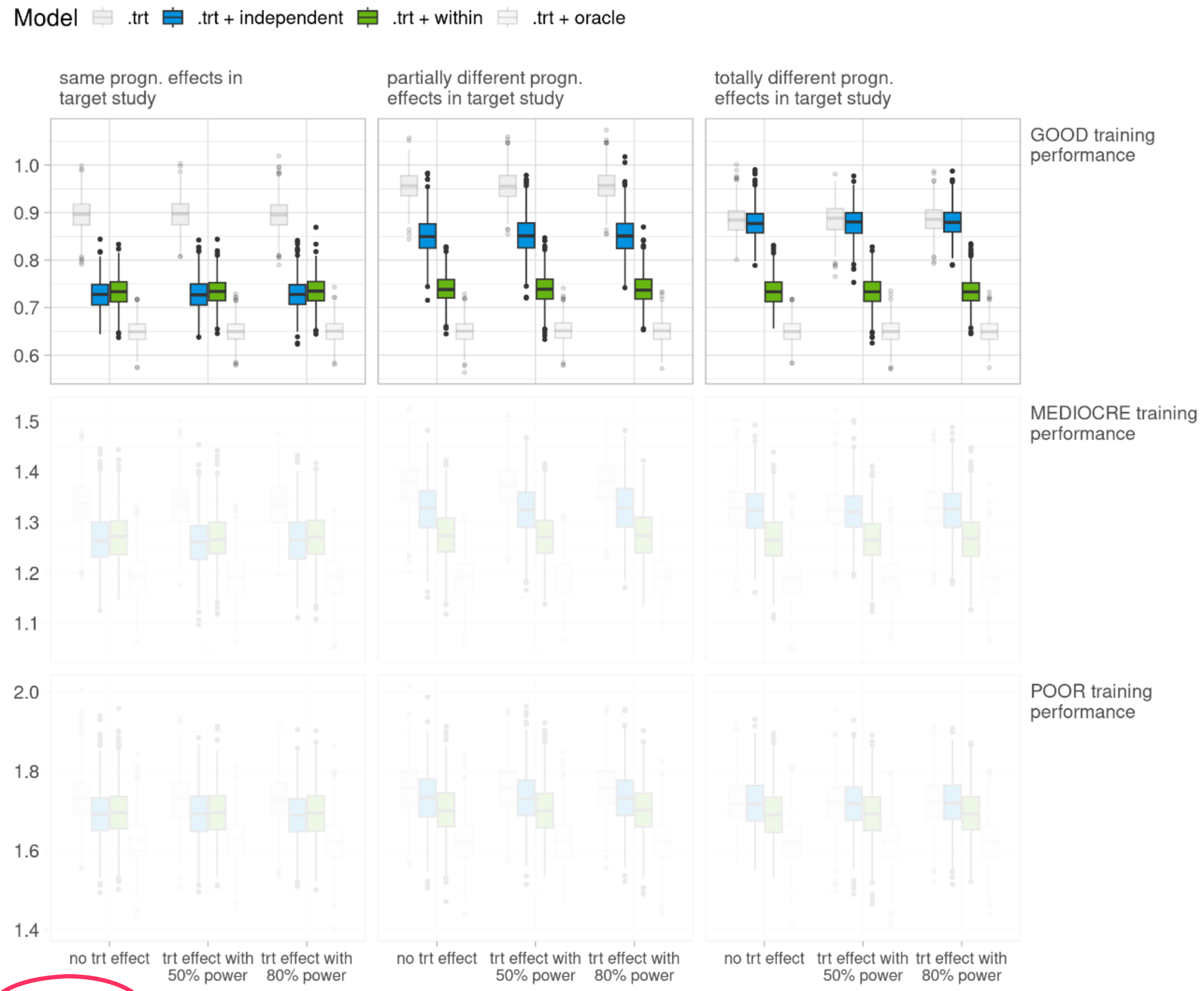
- **Independent study** score, if prognostic factors are the same
- **Within-study** score, if prognostic factors are at least partially different



Precision

Widths of 95% CI for the treatment effect estimate

- Size of historical data and control arm of target study is the same
- No Advantage of **independent study** score over **within-study** scor

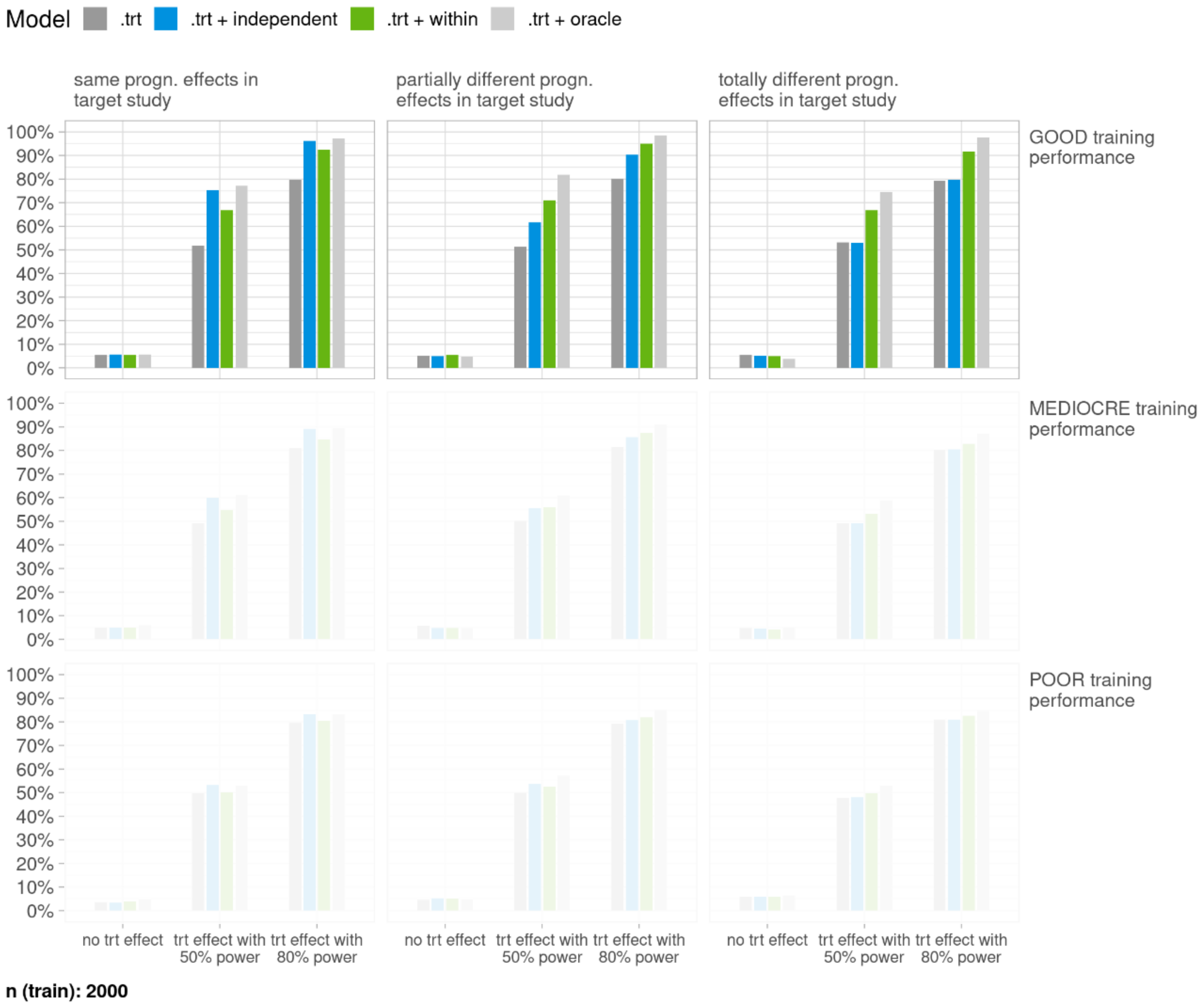


n (train): 125 = n (target, plc)

Power

Proportion of treatment effect estimate p-values ≤ 5%

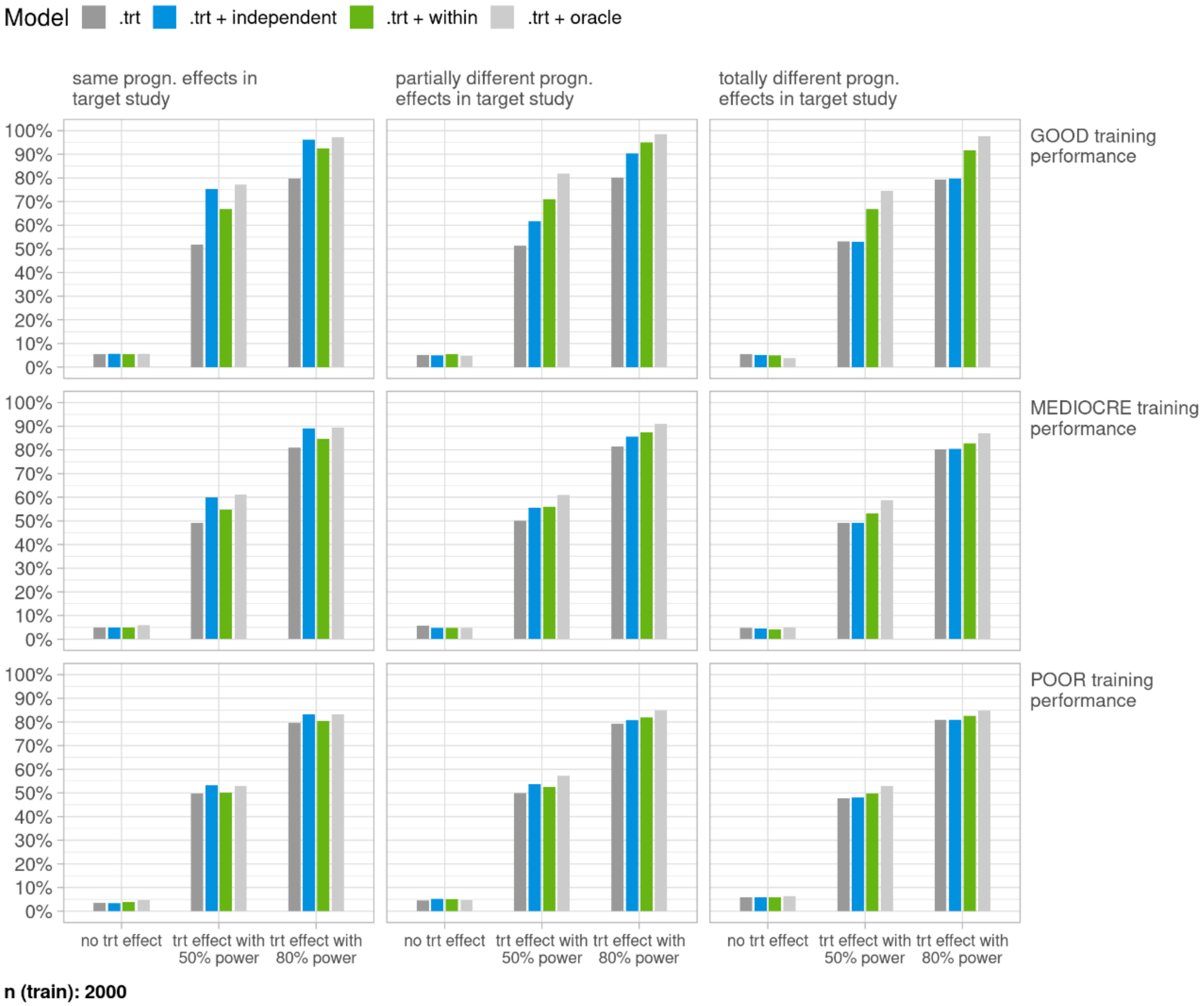
- Power either increases or stays the same, when adding any version of the prognostic score



Power

Proportion of treatment effect estimate p-values ≤ 5%

- No effect on type I error
- Slight differences from the targeted 5% type I error in some cases seem to be random and can be explained by the relatively low number of simulation iterations (1000)
- No loss of power



Conclusions from simulation study

Comparison of our **implementation of a within-study** vs **independent study** data led to the following conclusions.

- No systematic bias
- No effect on type I error
- No loss of power or precision

- **Each score version is preferrable in certain scenarios** in terms of larger benefits in the final analysis.

- **Important factors** to consider when choosing the appropriate score version
  - Similarity of populations (comparability of historical data set, at least partially matching)
  - Sample size (and availability) of historical data in comparison to target study

Health for all, Hunger for none

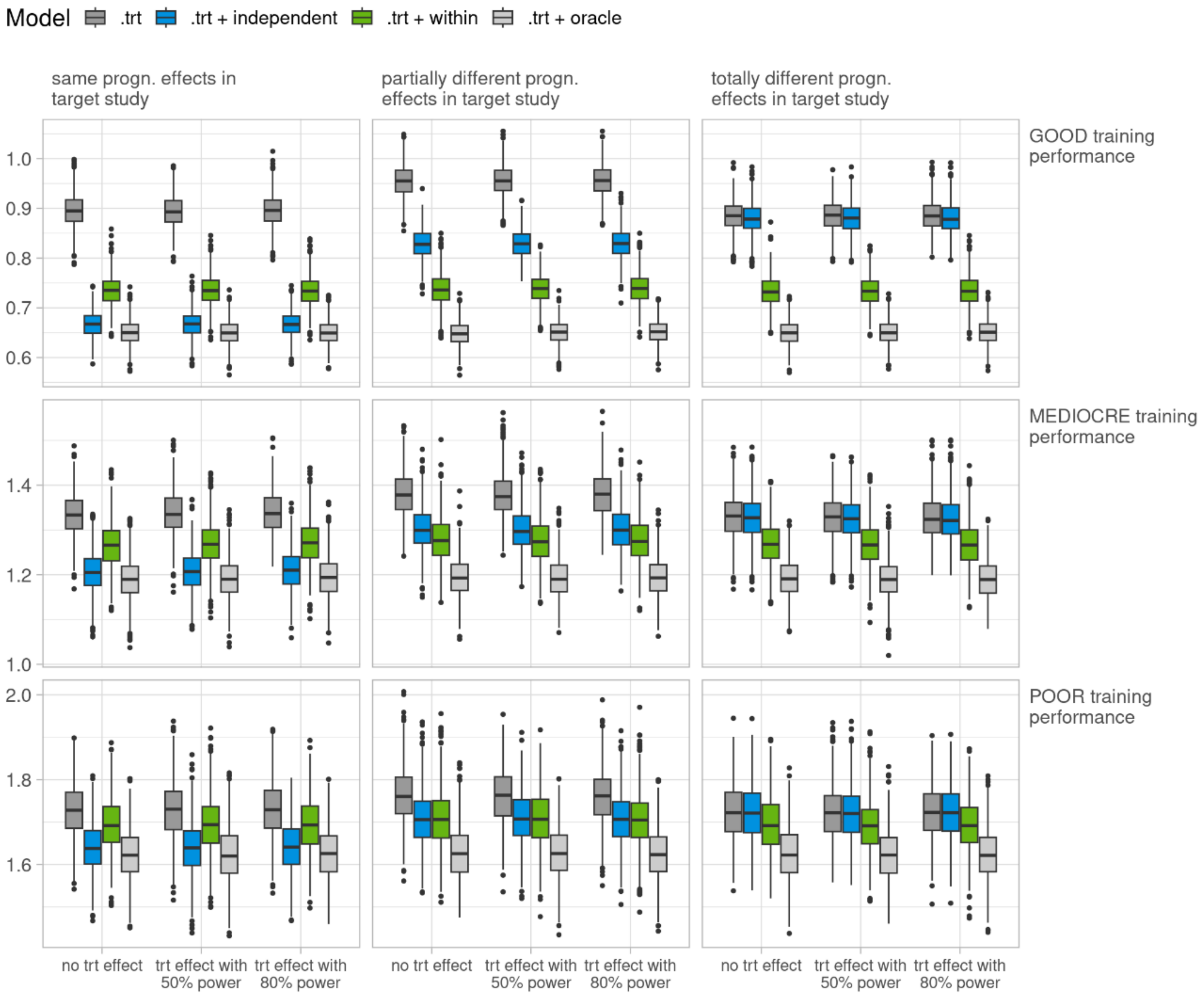
Thank you!

bayer.com

Precision

Widths of 95% confidence intervals for the treatment effect estimate

- Benefit of **independent study** score ranges between (close to) oracle and trt only model depending on similarity of populations, while **within-study** is independent by design



n (train): 2000

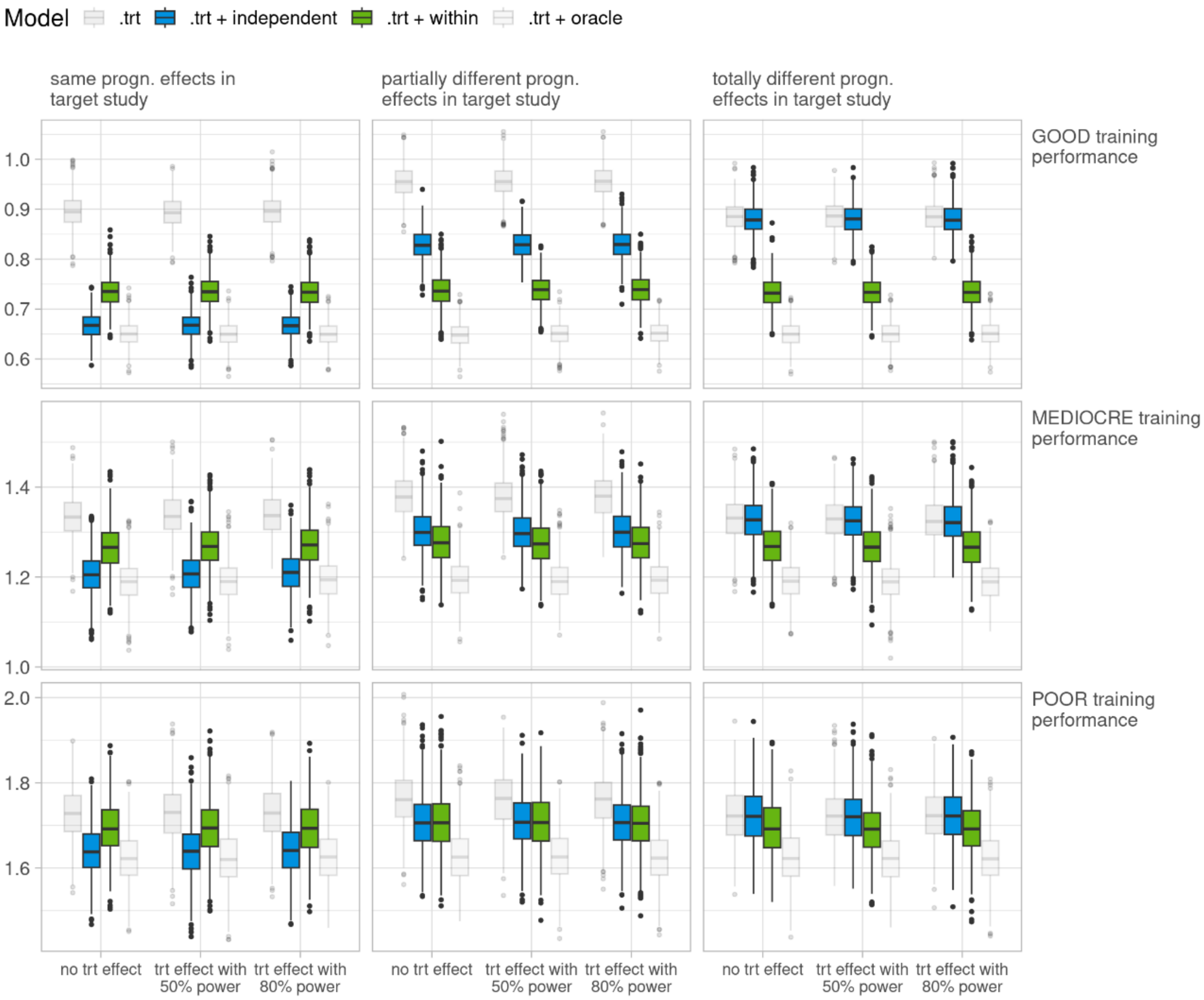
Precision

Widths of 95% confidence intervals for the treatment effect estimate

- Benefit of **independent study** score ranges between (close to) oracle and trt only model depending on similarity of populations, while **within-study** is independent by design
- Ranking of **within-study** and **independent study** approach depends on e.g. **similarity of studies:**

  Smaller CI width with the

  - **independent study** score, if prognostic factors are the same (left column, performance close to oracle)
  - **within-study** score, if prognostic factors are at least partially different (middle and right column)



n (train): 2000

Power

Proportion of treatment effect estimate p-values ≤ 5%

score vs. no adjustment

Independent study

~10 pct pts power increase

$R^2 = 18\%$ → ~20% sample size increase*

Within-study

~5 pct pts power increase

$R^2 = 9\%$ → ~10% sample size increase*

*according to ESSI formula