2026 ACIC Data Analysis Challenge

The year of the hydra

Sameer Deshpande
Jennifer Hill
George Perrett

Why a data challenge?

Plethora of causal inference methods!!!

How do we choose?

Don't we already have a mechanism for learning the strengths and weaknesses of methods? Refereed journal articles!

What could go wrong?

Why a data challenge?

Issues with methods papers

- Bias towards paper's proposed method due to
  - lack of expertise with/knowledge of competing methods
  - choice of competing methods and evaluation metrics
- Papers that assess performance based on simulations
  - typically cover a limited variety of scenarios
  - typically cover a limited number of competing methods
  - aren't calibrated well to data from real studies
- Papers that justify performance based on theory
  - may make assumptions that are far from realistic
  - can also overstate their performance advantages

Why a data challenge?

Issues with the overall landscape

- Refereeing systems is flawed
- Takes a while for new methods/approaches to be adopted
- Can be a disconnect between the method that is analyzed in a research paper and the one that is actually implemented in software
- Popularity of methods can depend more on the size of the megaphone than the quality of the method
- Reinforcement of "popular" methods exacerbated due to genAI (we've seen this in recent studies looking performance of genAI on causal inference tasks)

Why **this** data challenge?

Goal. This year the challenge focused on the complications that arise in situations with **multiple treatments**. WHY?

- Many existing software packages only handle binary treatments.
- Many causal inference researchers ignore (or avoid situations that require thinking about) multiple comparisons

Why **this** data challenge?

Setting: We focused only on **randomized experiments.** WHY?

- Challenging enough even in the randomized experiment setting.
- Didn't want to conflate issues caused by selection bias with those caused by a failure to properly account with uncertainty / multiplicity

Why **this** data challenge?

Accessibility: We wanted to make it **easier to participate**. HOW?

- Two submission tracks

- *The Works*: Submit results for all 9000 datasets.
- *Curation:* Focus on 18 datasets that represent the competition challenges

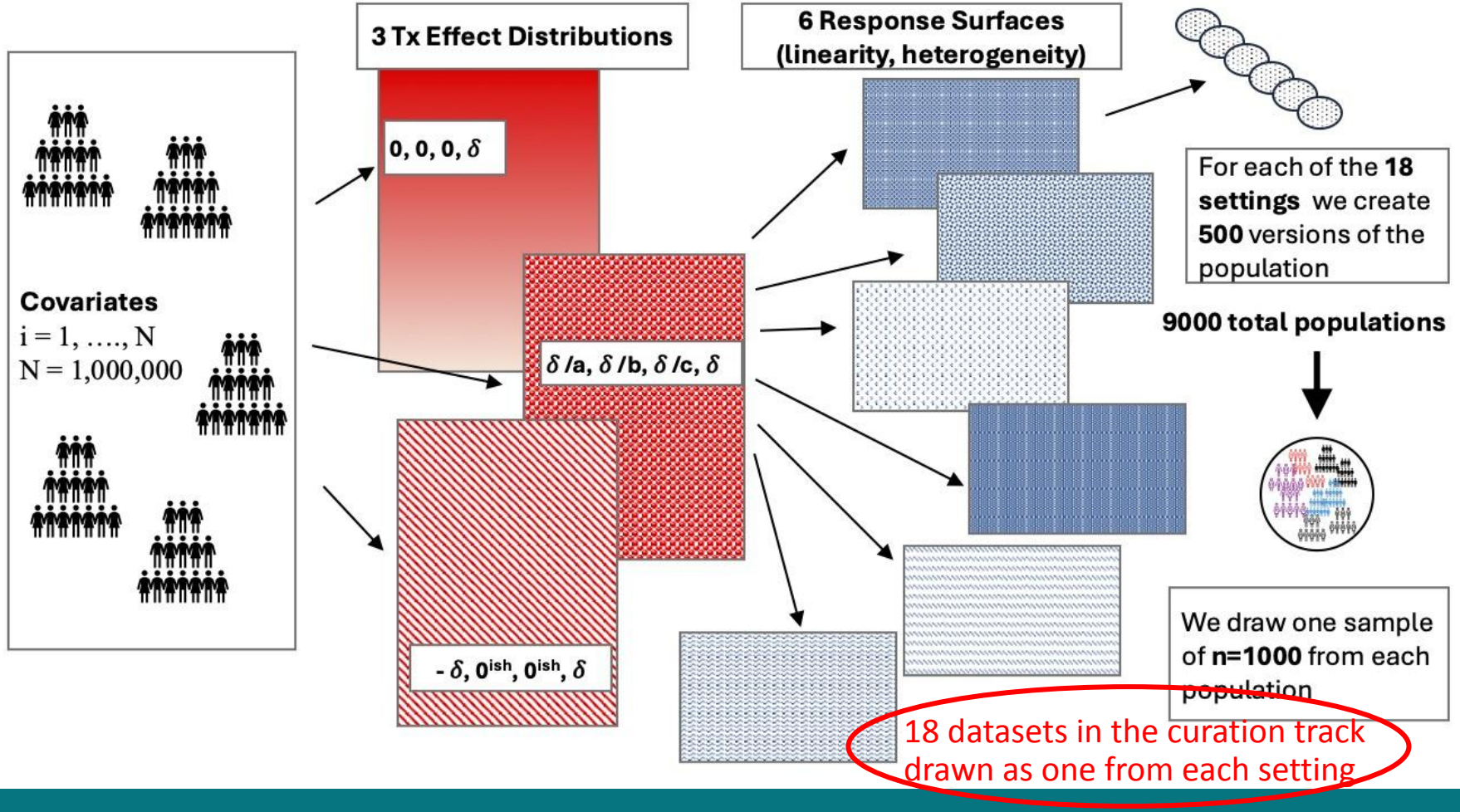
- Smaller datasets: n=1000

Everything could run on a laptop even if using a Bayesian machine learning algorithm.

data generating process

Overall data structure

- randomized experiment
- 5 treatment arms: Reference group A and 4 active arms (B,C, D, E)
- 3 different types of distributions of treatment effects sizes
- linear and non-linear response surfaces
- settings with both constant and varying treatment effects
- variation in the probability of being assigned to each arm


3 Tx Effect Distributions
6 Response Surfaces
(linearity, heterogeneity)
Covariates
i = 1, …., N
N = 1,000,000
0, 0, 0, δ
δ /a, δ /b, δ /c, δ
- δ, 0ish, 0ish, δ
For each of the 18 settings we create 500 versions of the population
9000 total populations
We draw one sample of n=1000 from each population
18 datasets in the curation track drawn as one from each setting

New considerations/subtleties

1) Threat of participants "learning" the DGP strategy and then gaming their strategy --> motivated creation of 9000 different populations

2) For what DGPs/estimands do sign errors make sense? do "false discoveries" make sense?

Estimands: individual and subgroup

$$\text{iCATE}(\mathbf{x}_i, z) = \mathbb{E}[Y_i(z) - Y_i(a) | \mathbf{X} = \mathbf{x}_i]$$

$$\text{subCATE}(z, x) = \left[\sum_{i=1}^{n} \mathbf{1}(x_{i,12} = x)\right]^{-1} \times \sum_{i=1}^{n} [\mathbb{E}[Y_i(z) - Y_i(a) | \mathbf{X} = \mathbf{x}_i] \times \mathbf{1}(x_{i,12} = x)]$$

Estimands: population and sample (conditional)

$$\mathrm{PATE}(z) = N^{-1} \sum_{i=1}^{N} Y_i(z) - Y_i(a)$$

$$\mathrm{sCATE}(z) = n^{-1} \sum_{i=1}^{n} \mathbb{E}[Y_i(z) - Y_i(a) | \mathbf{X} = \mathbf{x}_i]$$

submissions

Teams and Submissions

- 32 teams signed up
- 18 teams submitted a total of 63 different entries
- Majority of teams from academia but there were some from industry

Broad summary of types of methods

- BART: 12 submissions
  - Majority: single BART or separate BART for each arm
  - BCF variant: treatment effect ensemble had multivariate output
  - Varying coefficient BART: $E[Y \mid X, Z] = \mu(x) + \tau_A(x)*1(z = A) + \ldots + \tau_D(x)*1(z=D)$
- Generalized random forests: 4/62 submissions
- S-, X-, T-, or residual-learner: 12/62 submissions
- Prior fitted network / amortized inference: 5 of 62
- S-,X-, or T-learner: 12/62
- Other: ensembles (11), TMLE, prognostic score balancing
- UQ via posterior, bootstrap, or (DR) conformal

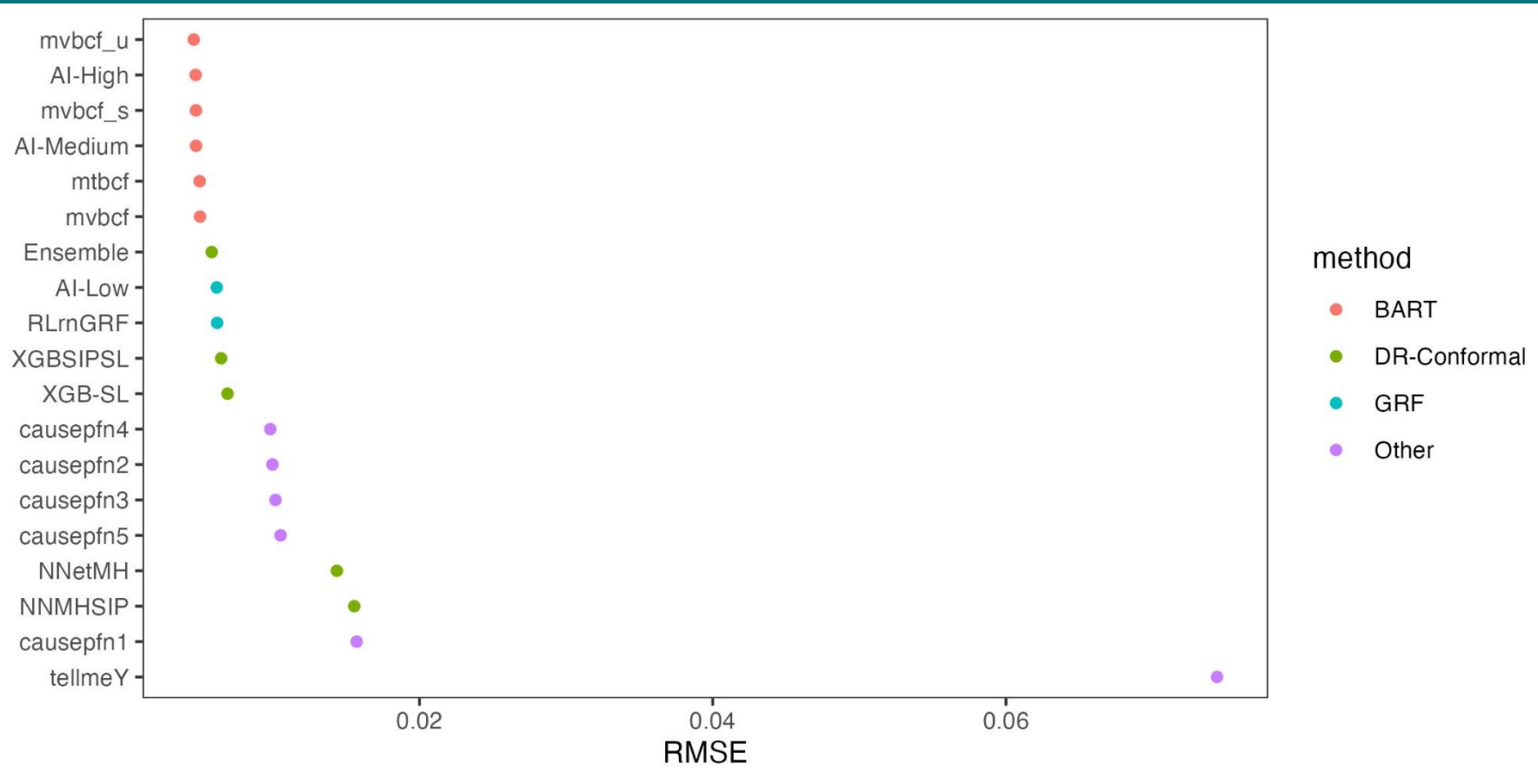
To AI or not to AI

- Several teams utilized generative AI to produce code & plan analyses
- Many had to manually tune & correct generated code
- Some teams use generative AI + agents to perform analysis
- One team used Claude for whole analysis w/ three levels of guidance
  - Minimal: picked and ran grf
  - Intermediate: compared several methods & selected BART
  - High: selected multi-chain BART and adjusted intervals based on synthetic data simulation

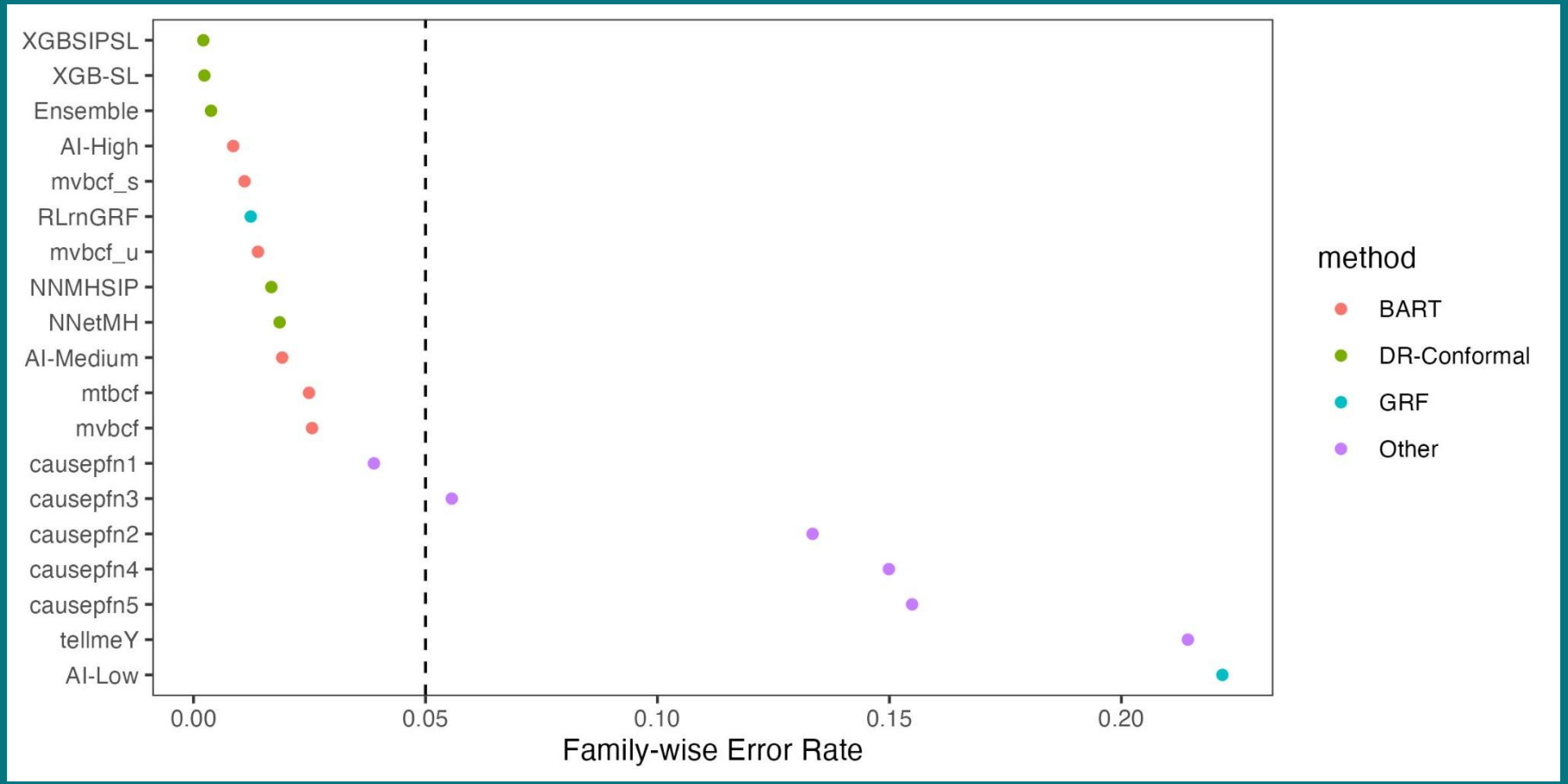
results

results (for "the works")

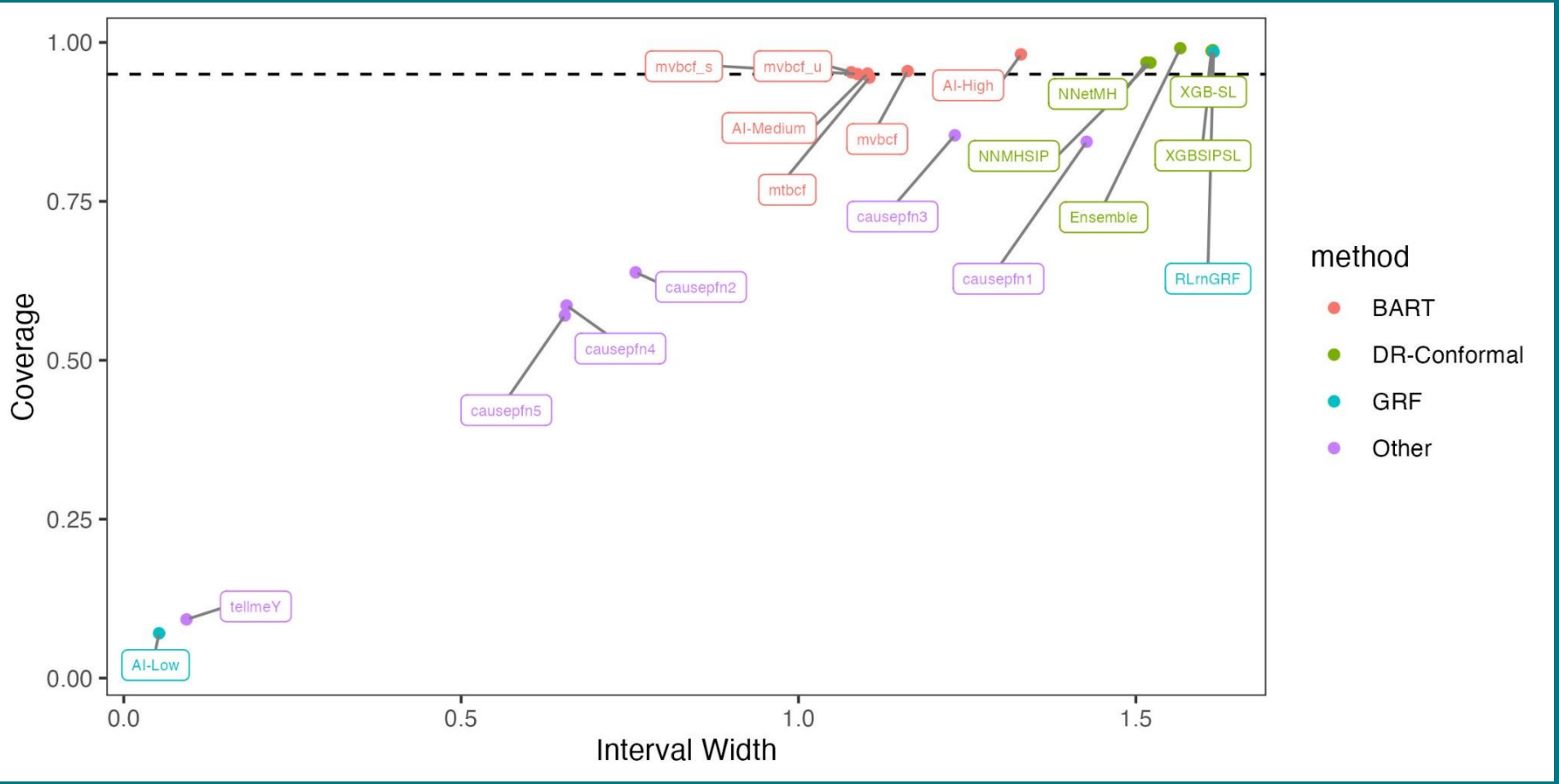
sample CATE - RMSE



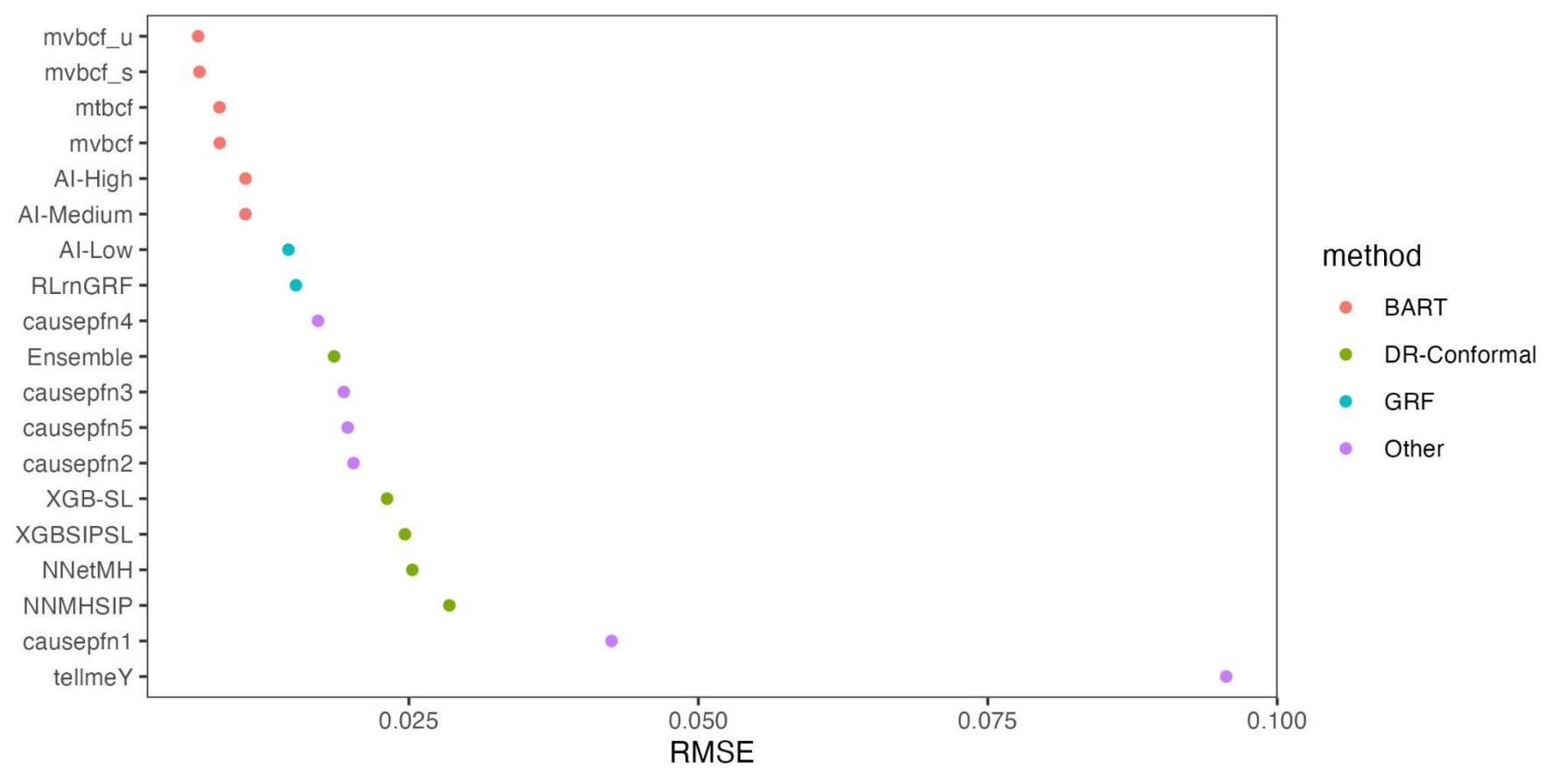
sCATE - family wise error rate



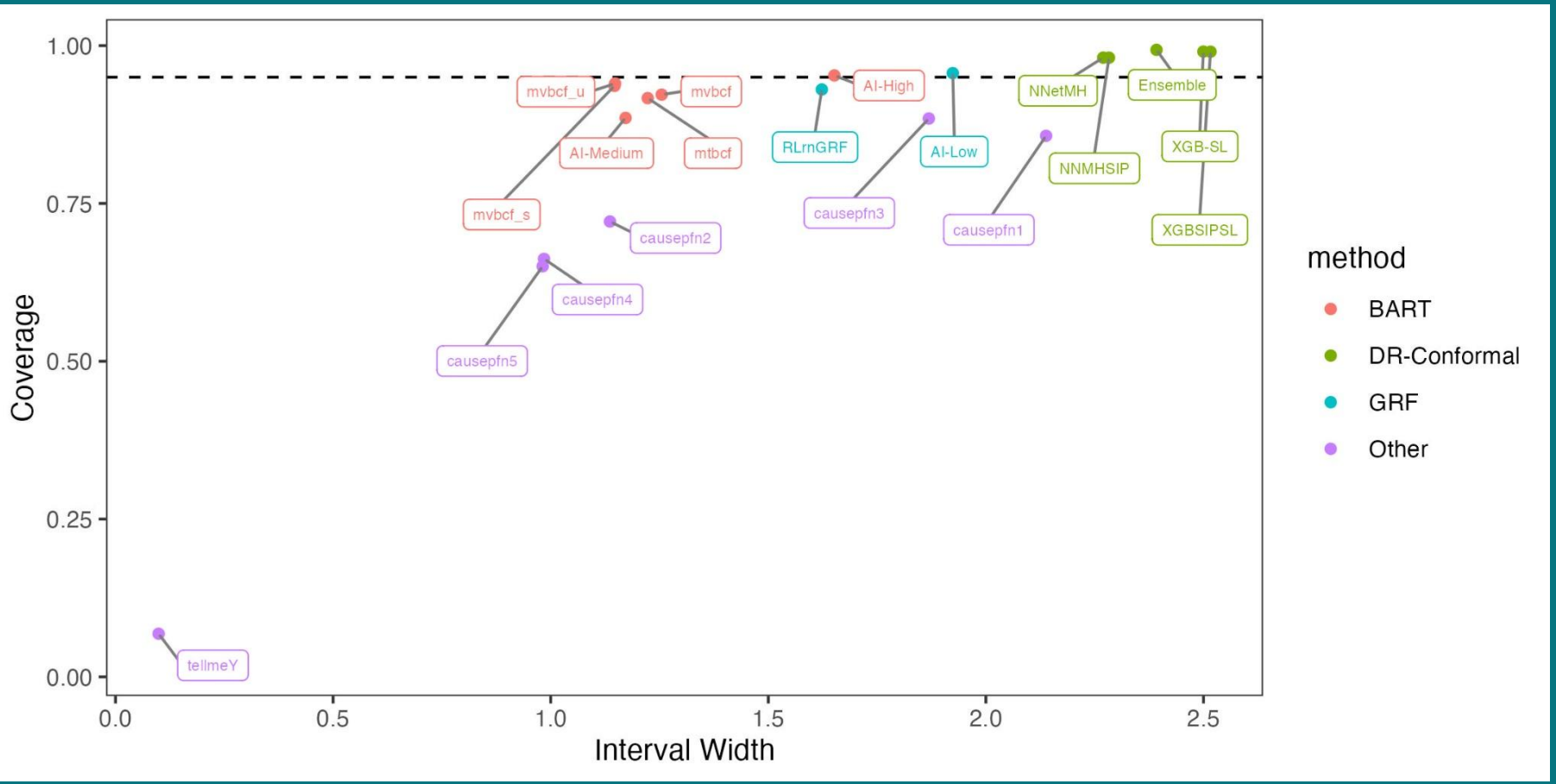
sCATE - coverage and interval width



subCATE - RMSE



subCATE - coverage and interval width



take-aways

What have we learned so far?

- Disclaimer: can only see broad trends right now. "Rankings" could flip as we disaggregate by setting.
- The action was in the uncertainty quantification. Standard trade-offs between FWER, coverage, and interval width
- Regularizing the treatment effect didn't make as big a difference this year -- likely because of the randomized experiment setting.
- Amortized methods underperformed because they weren't trained on datasets that were closely related to our problem -- they appear to be sensitive to the training data
- Conformal approaches were quite conservative (not surprisingly)

what next?

Analyses to come

We have just started our exploration. Additionally we plan to look at the following:

- Curation track analysis and announcement
- Differences by simulation setting (the 18 scenarios created by combinations of the 3 by 6 simulation knobs)
- Differences by type of AI assist
- Predictive power of features of the methodological approach
- Predictive power of more fine-grained features of the simulations

What should we test next time?

And who wants to be in charge?

Let us know.....!

Thank you!!!