Bayes Bayesian Biostatistics Rockville, MD 2024

PHASE V

Robust CATE Estimation Using Model Ensembles

Overview

- We want to understand **what works**, and **for whom**
- Several available approaches, **each can fall short** in certain scenarios
- Since the scenario (DGP) is unknown in a real setting, we **look for methods that are robust to the scenario**
- **Ensembles improve robustness** of estimation

Background – Treatment effect

The treatment effect for an individual can be thought of as the contrast between their two potential outcomes – $e_i = y_i^{T=1} - y_i^{T=0}$

This individual effect is unobservable!

Hence, a common focal point is the **Average Treatment Effect**:

$$ATE = \mathbb{E}(y^{T=1} - y^{T=0}) = \mathbb{E}(y^{T=1}) - \mathbb{E}(y^{T=0})$$

In an RCT $\mathbb{E}\left(y^{T=i}\right) = \mathbb{E}(y \mid T = i)$. Therefore:

$$ATE = \mathbb{E}(y \mid T = 1) - \mathbb{E}(y \mid T = 0)$$

There's more to it than just the average

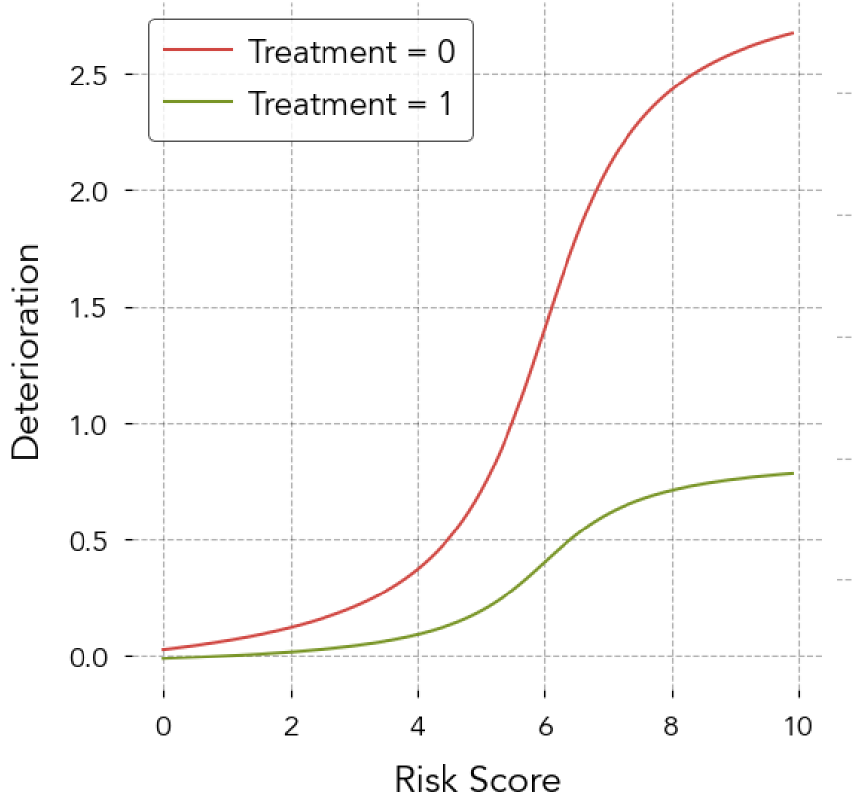
However, the ATE is not always enough.

When effect heterogeneity is plausible, focus may shift to the **Conditional ATE** (CATE):

$$CATE(x) = \mathbb{E}(y^{T=1} - y^{T=0} \mid X = x)$$

However, for CATE (even in an RCT) averaging by treatment is not a practical approach:

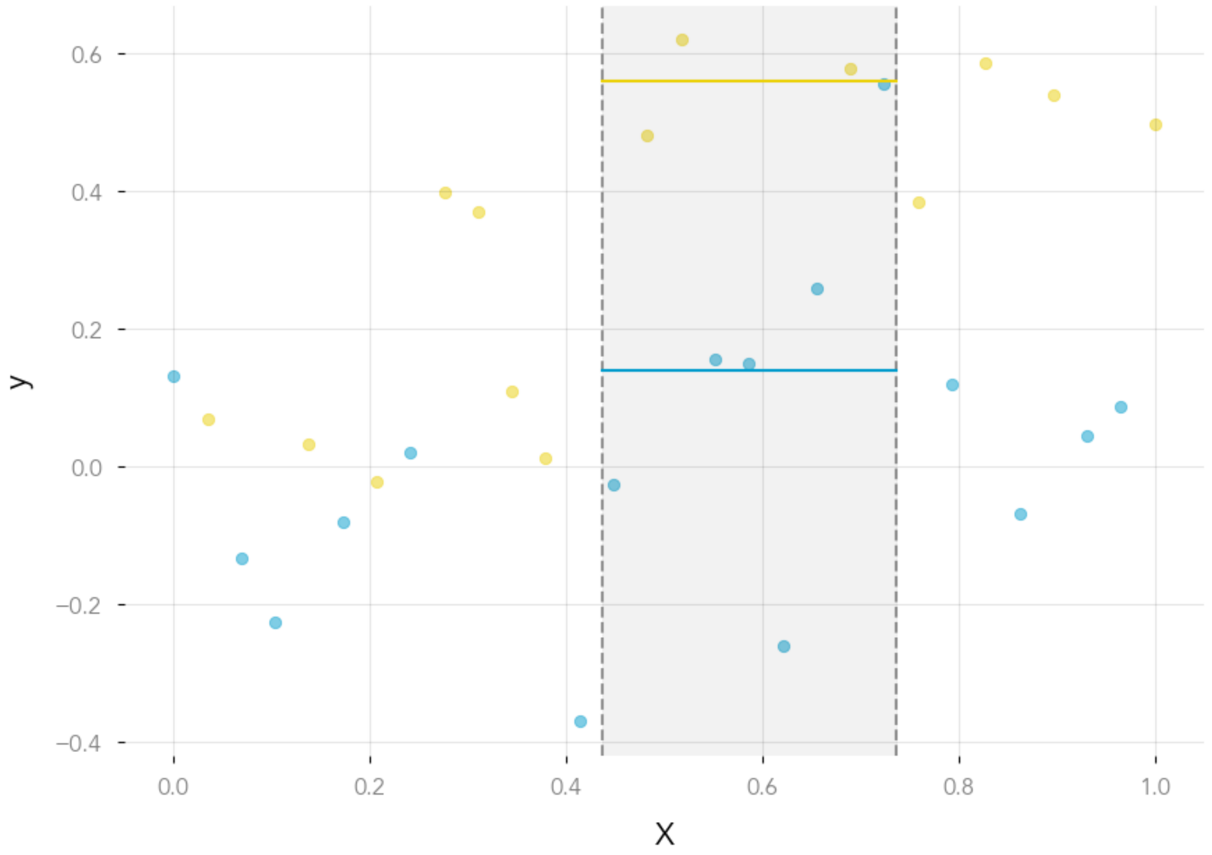
$$CATE(x) = \mathbb{E}(y|T = 1, X = x) - \mathbb{E}(y|T = 0, X = x)$$



So What Can We Do?

- **Causal Forest:**

If averaging is infeasible at a single point level, how about averaging in “areas”?



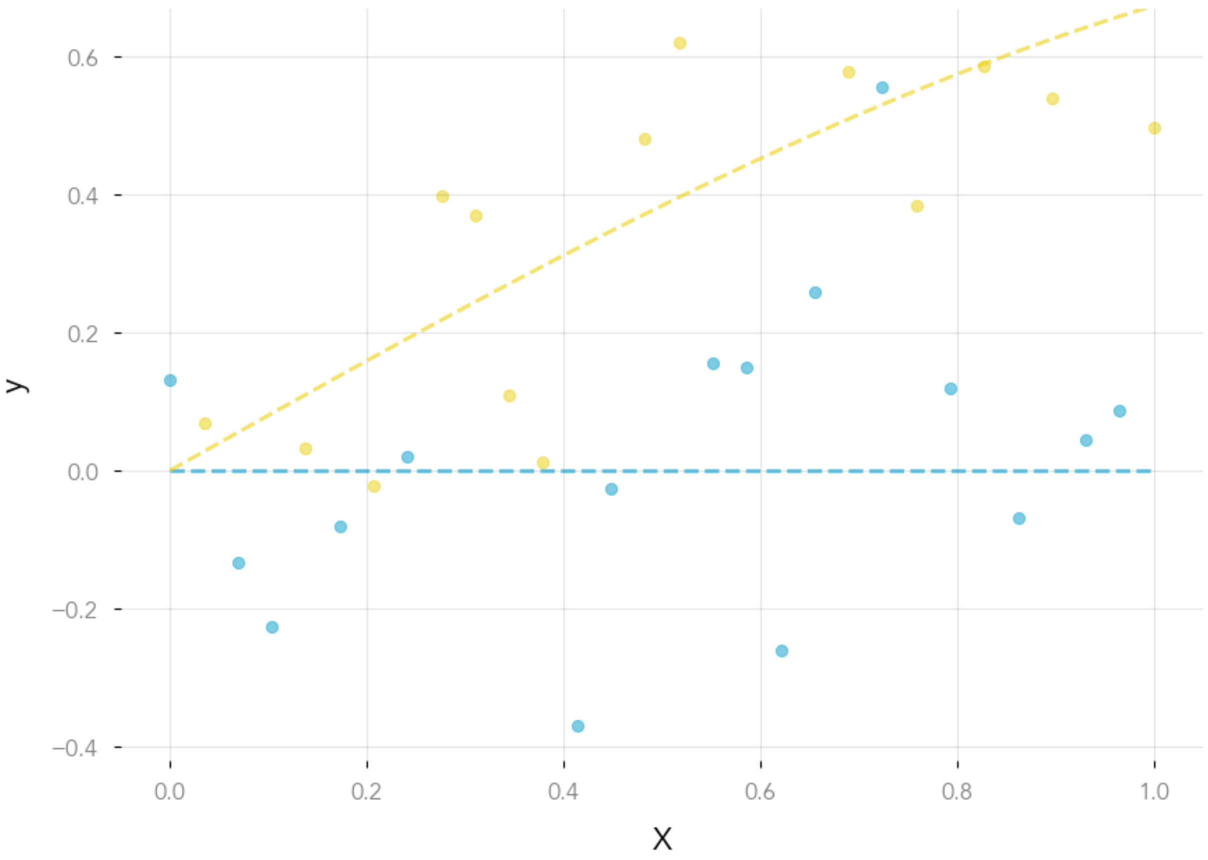
So What Can We Do?

- **Causal Forest:**

If averaging is infeasible at a single point level, how about averaging in "areas"?

- **Meta-Learners:**

Use global models to estimate the conditional outcomes (and other "nuisance" functions).



Meta Learners

PHASE V

- S (Single)

Train an outcome model using both X and T:

$$\mu(x,t) = \widehat{\mathbb{E}}[Y \mid X = x, T = t]$$

Estimate CATE using the difference:

$$\widehat{CATE}(x) = \mu(x,\mathbf{1}) - \mu(x,\mathbf{0})$$

Meta Learners

PHASEV

- S (Single)
- T (Two)

$$\mu_0(x) = \hat{\mathbb{E}}\,[Y|X = x, T = 0]$$

$$\mu_1(x) = \hat{\mathbb{E}}\,[Y|X = x, T = 1]$$

$$\widehat{CATE}(x) = \mu_1(x) - \mu_0(x)$$

Meta Learners

- S (Single)
- T (Two)
- X (Cross)

$$\mu_0(x) = \hat{\mathbb{E}}\,[Y|X = x, T = 0]$$

$$\mu_1(x) = \hat{\mathbb{E}}\,[Y|X = x, T = 1]$$

$$D_0 = \{\mu_1(x_i) - y_i : T_i = 0\}$$

$$D_1 = \{y_i - \mu_0(x_i) : T_i = 1\}$$

$$\delta_0(x) = \hat{\mathbb{E}}[D_0 \mid X = x]$$

$$\delta_1(x) = \hat{\mathbb{E}}[D_1 \mid X = x]$$

$$\pi(x) = \hat{P}(T = 1 \mid X = x)$$

$$\widehat{CATE}(x) = \pi(x)\delta_0(x) + \big(1 - \pi(x)\big)\delta_1(x)$$

Meta Learners

PHASEV

- S (Single)
- T (Two)
- X (Cross)
- R (Residualized)

Meta Learners

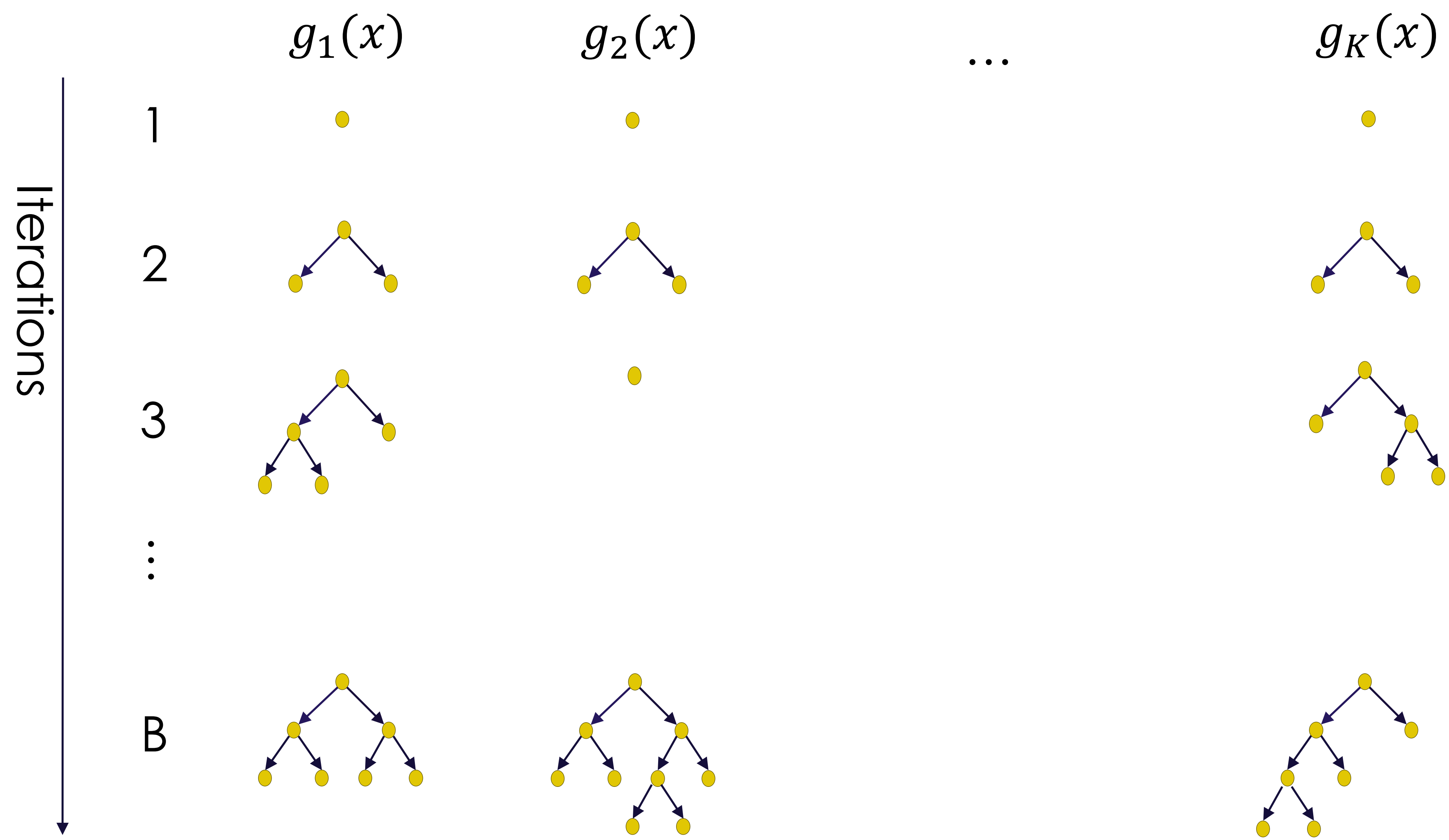
- S (Single)
- T (Two)
- X (Cross)
- R (Residualized)
- DR (Doubly Robust)

Meta Learners

- S (Single)
- T (Two)
- X (Cross)
- R (Residualized)
- DR (Doubly Robust)

- Can utilize any "base" model for learning the "nuisance" functions:
  - GLMs
  - Random Forests
  - Boosting
  - NN
  - BART

BART (Bayesian Additive Regression Trees)

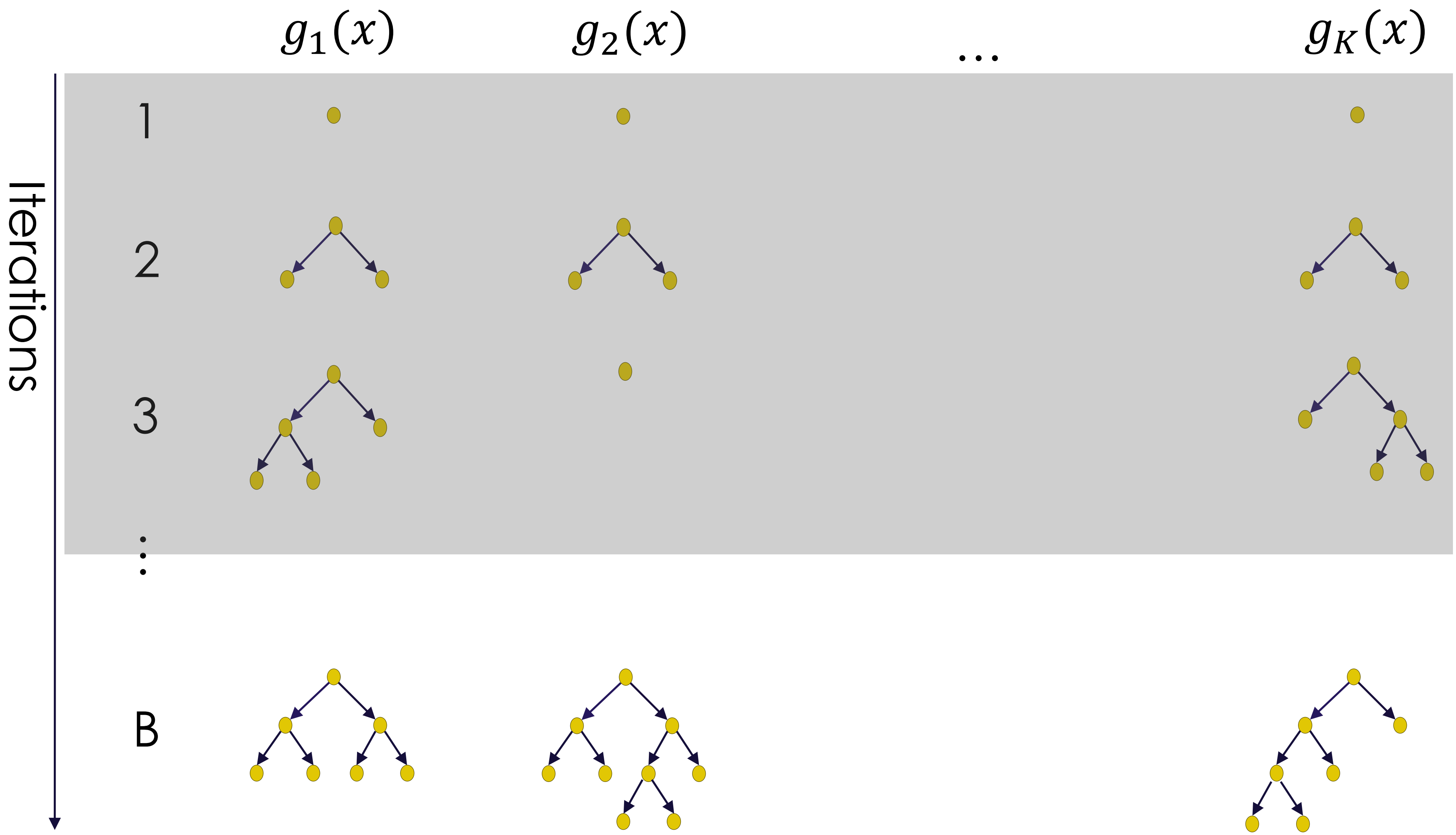


init: $g_i(x) = \frac{\bar{y}}{K}$

Iteratively fit $g_k(x)$ to: $y - \sum_{i \neq k} g_i(x)$

Fit is restricted by a regularizing prior on tree structure and terminal predictions

BART (Bayesian Additive Regression Trees)



init: $g_i(x) = \frac{\bar{y}}{K}$

Iteratively fit $g_k(x)$ to: $y - \sum_{i \neq k} g_i(x)$

Fit is restricted by a regularizing prior on tree structure and terminal predictions

Excluding a burn-in, the chain of iterations provides a posterior sample for $f(x)$: $\{\sum g_i^b(x)\}_{b=b_0}^{B}$

Bayesian Causal Forest (BCF)

PHASEV

A BART-tailored meta-learner with a disciplined approach for controlling the regularization of CATE explicitly:

$$\mu(x_i) = BART(x_i, \pi(x_i))$$

$$CATE(x_i) = BART(x_i)$$ ← more heavily regularised

$$y_i = \mu(x_i) + CATE(x_i) * T_i + \epsilon_i$$

Fitted using a Gibbs sampler that iteratively sets one of $\mu(x_i)$; $CATE(x_i)$ constant, and updates the other.

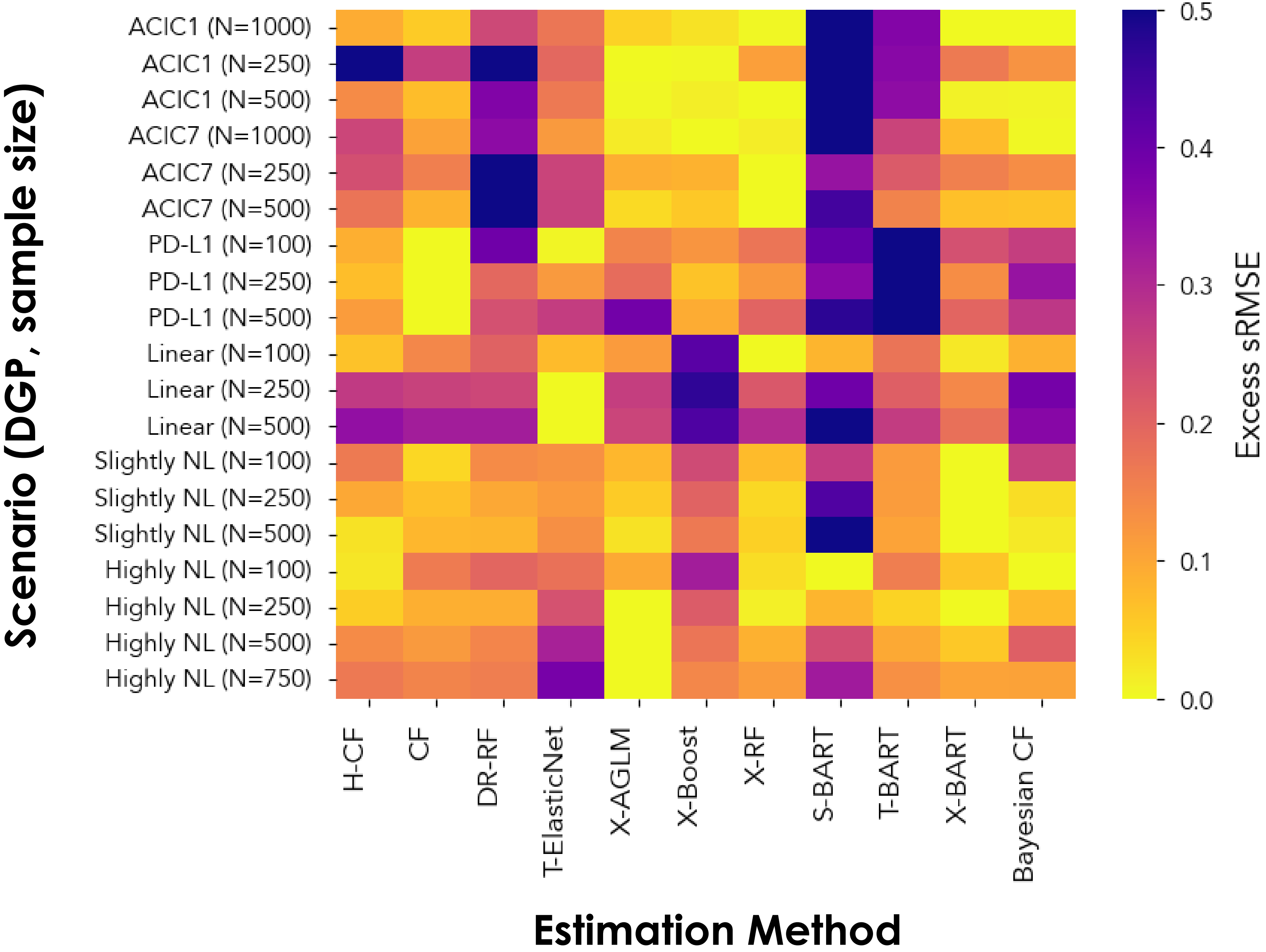
Simulation Study

PHASEV

- Scenarios (DGP):
  - ACIC – well known and used benchmark dataset
  - PDL1 – A Mechanistic model of PDL1 pathophysiology in oncology
  - Multivariate linear additive model (prognostic + predictive)
  - Multivariate non-linear models (various kinds)
- Sample sizes: 100 – 1000, to represent clinical data
- **Key performance measure**: standardised RMSE * (RMSE / s.d.(CATE))

* Aka PEHE in this context

No Single Dominant Model



Which method is "best"?

In each DGP, different methods perform better/worse.

+

In reality the DGP is unknown.

+

Ability to validate is limited:

- Individual effects are unobserved
- In clinical datasets – samples are relatively small

=

We want methods that are robust to the scenario (DGP)

Stacking (in classical prediction setting)

PHASEV

We want to combine models $\hat{y}^1 \dots \hat{y}^K$.

We do so by regressing them on the true outcome (in a test sample)

$$\hat{y} = \sum_{k=1}^{K} \omega^k \hat{y}_i^k \quad : \quad \omega = argmin_{\omega} \left\{ \sum_{i=1}^{N} \left[y_i - \sum_{k=1}^{K} \omega^k \hat{y}_i^k \right]^2 : \omega \geq 0 \right\}$$

In the causal setting:

The "label" is not y_i ,but $\mathrm{e}_i = y_i^{T=1} - y_i^{T=0}$, which is unobserved.

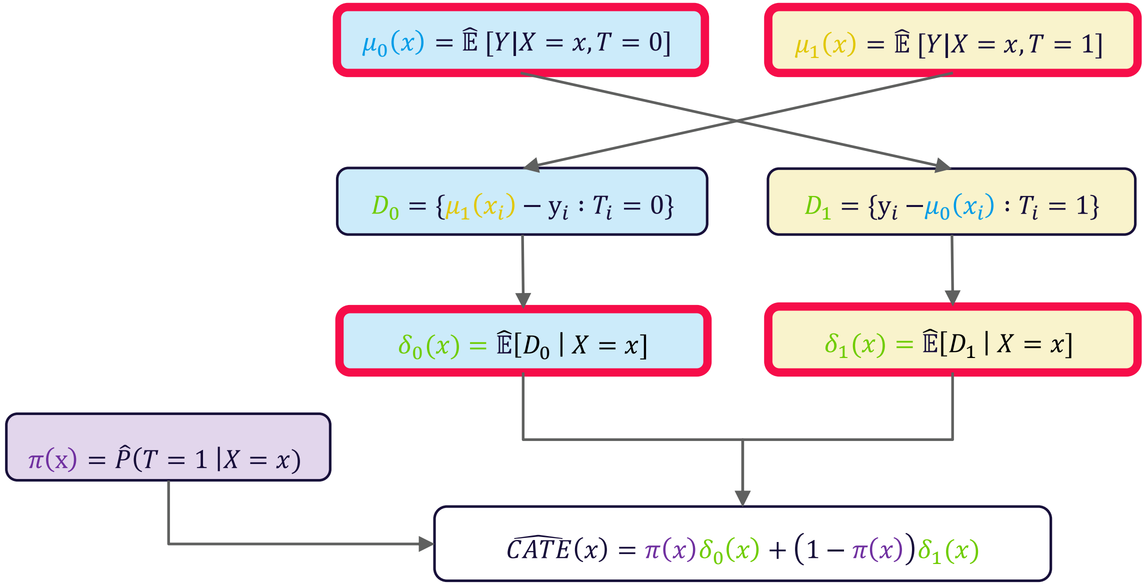
Several workarounds were suggested to substitute the missing label.

Nuisance Model Stacking

PHASEV

While we cannot directly stack on the unobserved effect e_i, we can benefit from stacking models for the outcome y_i ($\mu_0(x)$, $\mu_1(x)$).

In an X-Learner, we can also apply in the "pseudo-outcomes" D_i ($\delta_0(x)$, $\delta_1(x)$).



Bayesian Stacking

PHASEV

Train “base” models $f^k(x)$.
Also train a “null” model $f^0(x) = \bar{y}_{train}$.

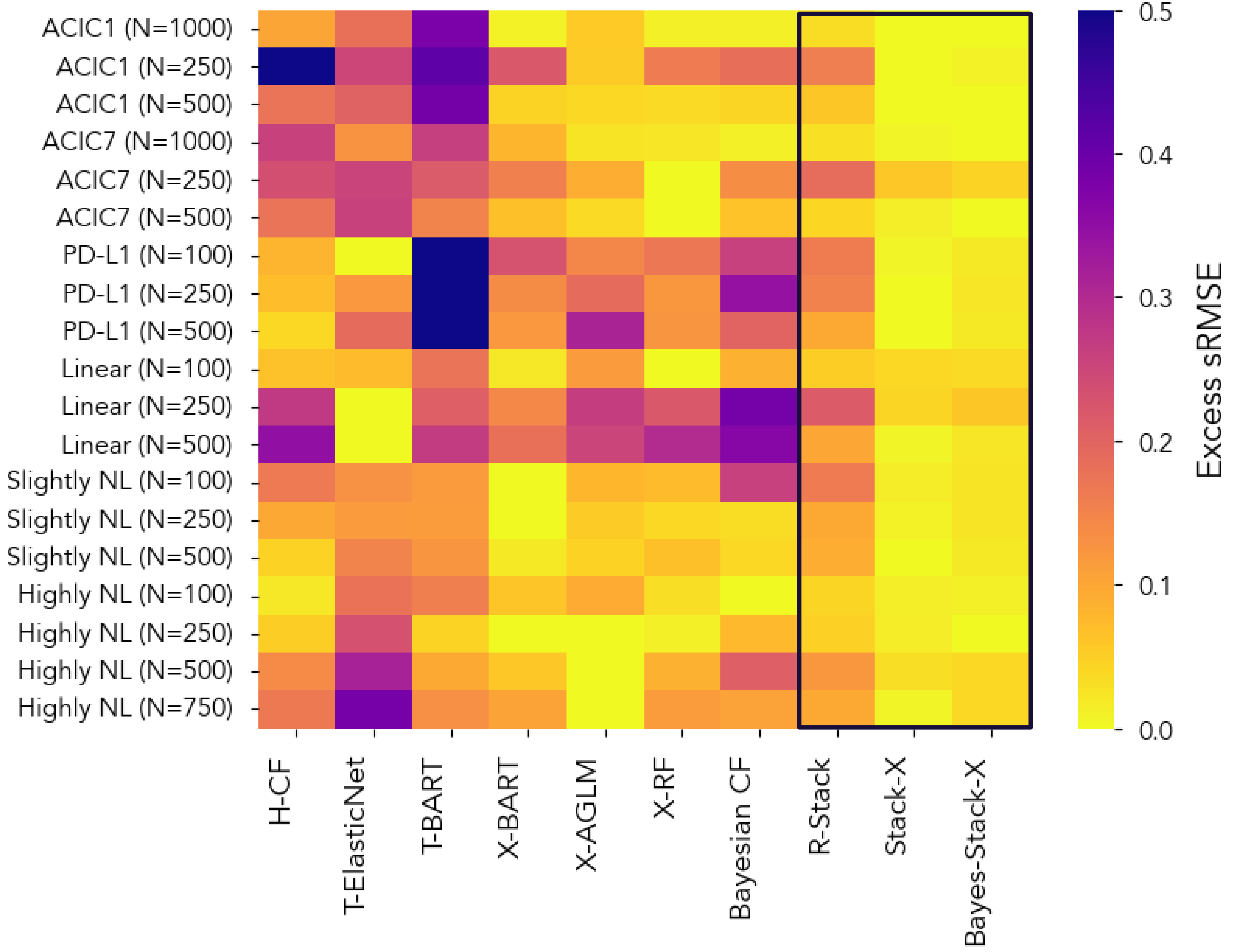
$$y_i = \omega^0 f^0(x_i) + \sum_{k=1}^{K} \omega^k f^k(x_i) + \varepsilon$$

$$\omega^0, \omega^1, \omega^2 \ldots \omega^K \sim Dirichlet\left(1, \frac{1}{10}, \frac{1}{10} \ldots \frac{1}{10}\right)$$

$$\varepsilon \sim N(0, \sigma^2)$$

$$\sigma \sim HN\left(0, \frac{\sqrt{var(y_{train})}}{3}\right)$$

Results


ACIC1 (N=1000)
ACIC1 (N=250)
ACIC1 (N=500)
ACIC7 (N=1000)
ACIC7 (N=250)
ACIC7 (N=500)
PD-L1 (N=100)
PD-L1 (N=250)
PD-L1 (N=500)
Linear (N=100)
Linear (N=250)
Linear (N=500)
Slightly NL (N=100)
Slightly NL (N=250)
Slightly NL (N=500)
Highly NL (N=100)
Highly NL (N=250)
Highly NL (N=500)
Highly NL (N=750)
H-CF
T-ElasticNet
T-BART
X-BART
X-AGLM
X-RF
Bayesian CF
R-Stack
Stack-X
Bayes-Stack-X
0.5
0.4
0.3
0.2
0.1
0.0
Excess sRMSE

References

PHASEV

Causal Forest doi.org/10.1073/pnas.1510489113
Recursive partitioning for heterogeneous causal effects, S. Athey G. Imbens, 2016

Meta Learners doi.org/10.1073/pnas.1804597116
Metalearners for estimating heterogeneous treatment effects using machine learning, S. Kunzel et al, 2019

DR Learner www.aeaweb.org/articles?id=10.1257/aer.p20171038
Double/Debiased/Neyman Machine Learning of Treatment Effects, V. Chernozhukov et al, 2017

BART doi.org/10.1214/09-AOAS285
BART: Bayesian additive regression trees, H. Chipman, E. George, R. McCulloch, 2010

BCF 10.1214/19-BA1195
Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects, R. Hahn et al, 2020

Stacking https://hastie.su.domains/ElemStatLearn/
The Elements of Statistical Learning, T. Hastie, R. Tibshirani, J. Friedman, 2008, Chapter 8.8

Bayesian Stacking 10.1214/17-BA1091
Using Stacking to Average Bayesian Predictive Distributions, Y. Yao et al , 2018