Data Integration Approaches to Estimate Heterogeneous Treatment Effects

Elizabeth Stuart

Hurley-Dorrier Professor and Chair

Department of Biostatistics

estuart@jhu.edu; @lizstuartdc

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Funding/Disclaimer and Acknowledgments

Funding

- PCORI Award ME-2020C3-21145 *(PI: Stuart)*
- NIMH R01MH126856 *(PI: Stuart)*
- NIA Epidemiology & Biostatistics of Aging T32 AG000247 *(PI: Bandeen-Roche)*

Disclaimer

This presentation is based on research using data from data contributors, Takeda and Lundbeck, made available through Vivli, Inc. Vivli has not contributed to or approved, and is not in any way responsible for, the contents of this presentation.

Thanks to collaborators

JHU: Carly Lupton Brantner, Leon di Stefano, Peter Zandi, Trang Nguyen, Harsh Parikh

Duke: Hwanhee Hong, Wen Zhao, Qiao Wang

Introduction & Background

Causal Inference

Question: What is the effect of a treatment (W) on an outcome (Y)?

Key concept: Potential outcomes

$Y_i(1)$

Outcome that *would have been observed* if unit *i* received the **treatment**

$Y_i(0)$

Outcome that *would have been observed* if unit *i* received the **control**

Fundamental problem of causal inference: we can only observe one of the potential outcomes for a unit at a given time

Motivation…

- The holy grail: determining “what works for whom”
- Treatment effect heterogeneity / modification / moderation
- Do treatment (causal) effects vary across individuals?
- Can we use this to inform treatment decisions for individuals?
- That would be great . . .

Causal Estimands

Causal effect: Difference in potential outcomes under treatment versus control

Average treatment effect (ATE)

$$\delta = E(Y(1) - Y(0))$$

What if the effect of the treatment depends on covariates, $\boldsymbol{X}$?

Conditional average treatment effect (CATE)

$$\tau(\boldsymbol{X}) = E(Y(1) - Y(0)|\boldsymbol{X})$$

Challenge

Randomized controlled trials:

- Unconfounded treatment assignment (random)
- Often powered to detect main effects (ATE) rather than effect *heterogeneity* (CATE)

Observational data:

- Confounded treatment assignment
- Larger and often more representative of target population

Potential solution: Data integration

Combining Data Sources

- Can we get the best of both worlds?
- Combine the unbiasedness of trials with the large size and representativeness of non-experimental studies?
- LOTS of methods work in this area right now, known sometimes as data fusion, data integration, hybrid designs, individual patient data meta-analysis, . . .
- So far we have mostly been adapting machine learning and Bayesian methods to combine multiple randomized trials; eventually want to bring in electronic health record data too
- Machine learning methods allow for flexible identification of moderators, interactions, etc., with no need to prespecify

Review of Data Integration Methods

Causal Assumptions

1. **Stable Unit Treatment Value Assumption (SUTVA)** in each study

2. **Unconfoundedness**: $\{Y(0), Y(1)\} \perp W|\boldsymbol{X}$ in each study

3. **Consistency**: $Y = WY(1) + (1 - W)Y(0)$ almost surely in each study

4. **Positivity of treatment assignment**: There exists a constant $b > 0$ such that $b < P(W = 1|\boldsymbol{X} = \boldsymbol{x}) < 1 - b$ for all $\boldsymbol{x}$ in each study

5. **Positivity of study membership** [Combining trials]: There exists a constant $c > 0$ such that $c < P(S = s|\boldsymbol{X} = \boldsymbol{x}) < 1 - c$ for all $\boldsymbol{x}$ and s

6. **Positivity of study membership** [Extending to new setting]: There exists a constant $d > 0$ such that $d < P(S \in \{1, \dots, K\}|\boldsymbol{X} = \boldsymbol{x}) < 1 - d$ for all $\boldsymbol{x}$ in the target setting

Reviewed Approaches

TABLE 1
Comparison of approaches to estimate CATE using multiple studies

Approach	Data level	Data types	Model	Estimand	Motivation
Meta-Analysis of Interactions	AD	RCTs	Parametric	Pooled	Pool treatment-covariate interactions
Meta-Regression	AD	RCTs	Parametric	Pooled	Model group-level treatment-covariate interactions
Meta-Analysis of Local Models	FL	RCTs	Parametric	Pooled	Pool treatment-covariate interactions
Tan, Chang and Tang (2021)	FL	RCTs	Nonparametric	Study-specific	Borrow information from other studies to improve model
One-Stage Meta-Analysis	IPD	RCTs	Parametric	Pooled	Model individual-level treatment-covariate interactions
Meta-Analysis of IPD and AD	IPD/AD	RCTs	Parametric	Pooled	Adaptively incorporate AD as auxiliary data
Rosenman et al. (2022)	IPD	RCT and OD	Parametric	Pooled	Weight combination of CATE estimators based on OD bias
Rosenman et al. (2020)	IPD	RCT and OD	Parametric	Pooled	Weight combination of CATE estimators based on OD bias
Cheng and Cai (2021)	IPD	RCT and OD	Nonparametric	Study-specific	Weight combination of CATE estimators based on OD bias
Yang, Zeng and Wang (2020)	IPD	RCT and OD	Parametric	Pooled	Weight combination of CATE estimators based on OD bias
Kallus, Puli and Shalit (2018)	IPD	RCT and OD	Nonparametric	Pooled	Estimate confounding function
Yang, Zeng and Wang (2022)	IPD	RCT and OD	Parametric	Pooled	Estimate confounding function
Wu and Yang (2021)	IPD	RCT and OD	Nonparametric	Pooled	Estimate confounding function
Hatt et al. (2022)	IPD	RCT and OD	Nonparametric	Pooled	Estimate confounding function

AD = aggregate-level data, FL = federated learning, IPD = individual participant-level data, RCT = randomized controlled trial, OD = observational data

Key Consideration: Data Level

	Overview	*Benefits*	*Challenges*
Aggregate-Level Data (AD)	Summary-level data available	• Draw conclusions about average effects • Easily accessible	• Aggregation bias • Limited power to detect effect moderation
Federated Learning (FL)	IPD accessible within studies and only AD sharable across studies	• Maintain data privacy • More control over analysis methods	• Less flexible than IPD • Studies can only learn from each other on aggregate
Individual Participant-Level Data (IPD)	Individual data available and shareable across all studies	• Highest modeling flexibility • High power	• Unknown causes of study-level heterogeneity

Key Consideration: Data Level

	Overview	*Benefits*	*Challenges*
Aggregate-Level Data (AD)	Summary-level data available	• Draw conclusions about average effects • Easily accessible	• Aggregation bias • Limited power to detect effect moderation
Federated Learning (FL)	IPD accessible within studies and only AD sharable across studies	• Maintain data privacy • More control over analysis methods	• Less flexible than IPD • Studies can only learn from each other on aggregate
Individual Participant-Level Data (IPD)	Individual data available and shareable across all studies	• Highest modeling flexibility • High power	• Unknown causes of study-level heterogeneity

Key Consideration: Data Level

	Overview	*Benefits*	*Challenges*
Aggregate-Level Data (AD)	Summary-level data available	• Draw conclusions about average effects • Easily accessible	• Aggregation bias • Limited power to detect effect moderation
Federated Learning (FL)	IPD accessible within studies and only AD sharable across studies	• Maintain data privacy • More control over analysis methods	• Less flexible than IPD • Studies can only learn from each other on aggregate
Individual Participant-Level Data (IPD)	Individual data available and shareable across all studies	• Highest modeling flexibility • High power	• Unknown causes of study-level heterogeneity

Key Consideration: Modeling Approach

Parametric

- Require distributional assumptions and pre-specification of hypothesized relationships
- Typically highly interpretable
- Might miss complex interactions or nonlinearities

Non-Parametric

- Do not require distributional assumptions or pre-specification
- Challenging to interpret
- More flexible

Combining RCTs to Estimate Heterogeneous Treatment Effects

Approach

- Combine IPD from multiple randomized controlled trials (RCTs)
- **Common approach:** Meta-analysis
- **Our approach:** Extend single-study non-parametric (machine learning) approaches to estimate the CATE in multiple trials

Motivating Application: Depression Treatments

Question: Are medications for depression differentially effective?

- Comparison of Duloxetine and Vortioxetine for individuals with major depressive disorder

Duloxetine	Vortioxetine
(Cymbalta) SNRI; increases serotonin and noradrenaline Used in practice at time of trials	(Trintellix) Modulates receptor and inhibits serotonin transporter New at time of trials
Better than placebo	Better than placebo

Trial Data

- Four RCTs* (n = 575, 436, 418, 418) with participants randomly assigned to Duloxetine or Vortioxetine
- **Eligibility criteria:**

18-75 years old	Had a major depressive episode lasting ≥ 3mo	Had MADRS score ≥ 22 or 26 at screening & baseline

- **Outcome:** Change in MADRS score from baseline to the last observed follow-up

*Baldwin et al., 2021; Boulenger et al., 2014; Mahableshwarkar et al., 2013; Mahableshwarkar et al., 2015

Methods: Overview

Single-study methods

1. S-Learner
2. X-Learner
3. Causal Forest

+

Aggregation methods

1. Complete Pooling
2. Pooling with Trial Indicator
3. Ensemble Forest
4. Meta-Analysis

Single-Study Methods

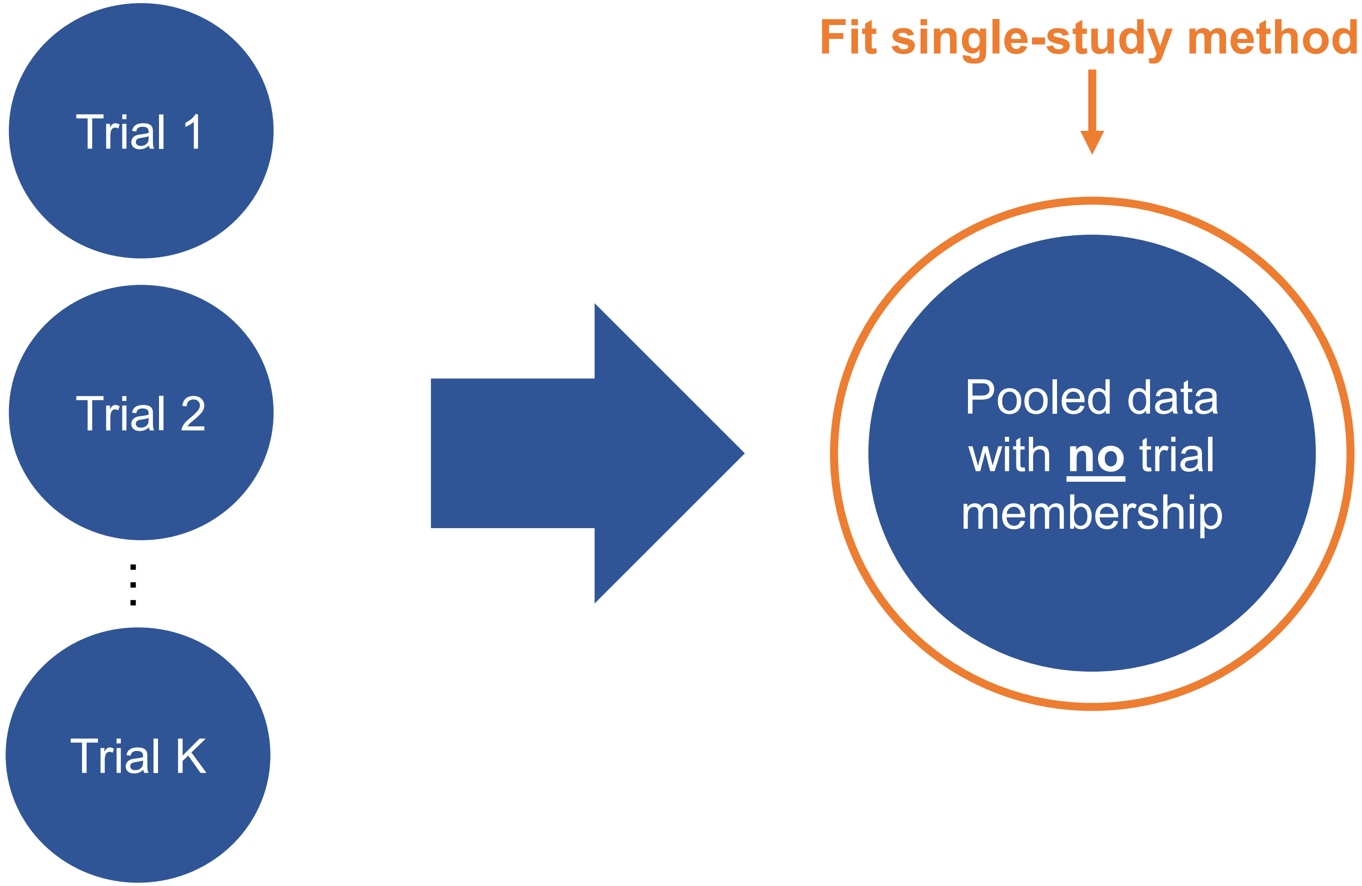
- Estimate conditional mean outcomes: $\mu(\boldsymbol{X_i}, W_i) = E(Y_i|\boldsymbol{X_i}, W_i)$ and then calculate the difference: $\mu(\boldsymbol{X_i}, 1) - \mu(\boldsymbol{X_i}, 0)$
  - **S-Learner** [Kunzel et al., 2019]
  - **X-Learner** [Kunzel et al., 2019]

  *Base learner = random forest in this chapter; Bayesian additive regression trees in next chapter

- Forest-based algorithm: partition the covariates based on treatment effect heterogeneity
  - **Causal Forest** [Athey et al., 2019]

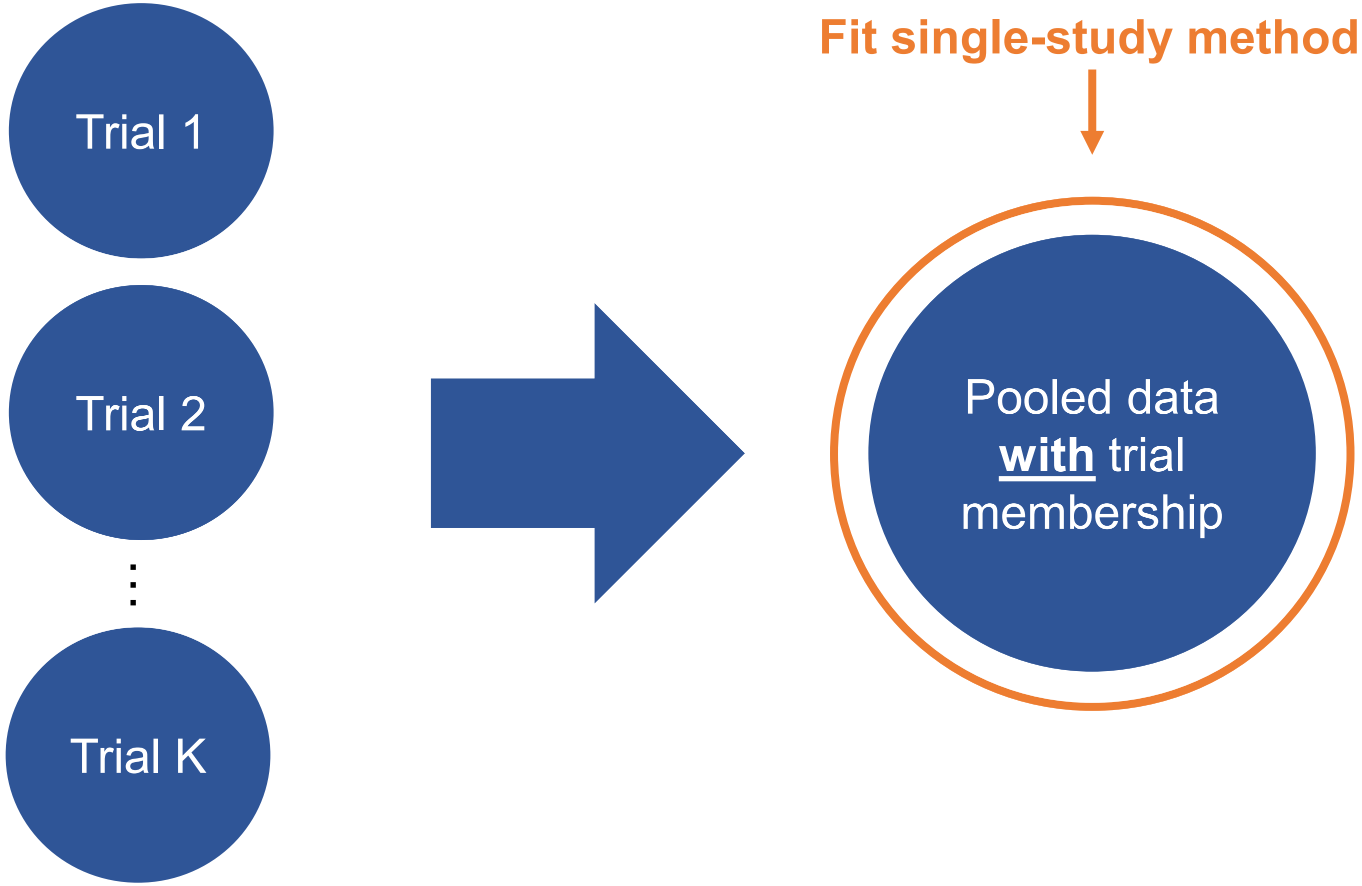
Aggregation Methods

Complete Pooling: Treat all data as if it were from a *single study* – pool together and then apply one of the single-study approaches



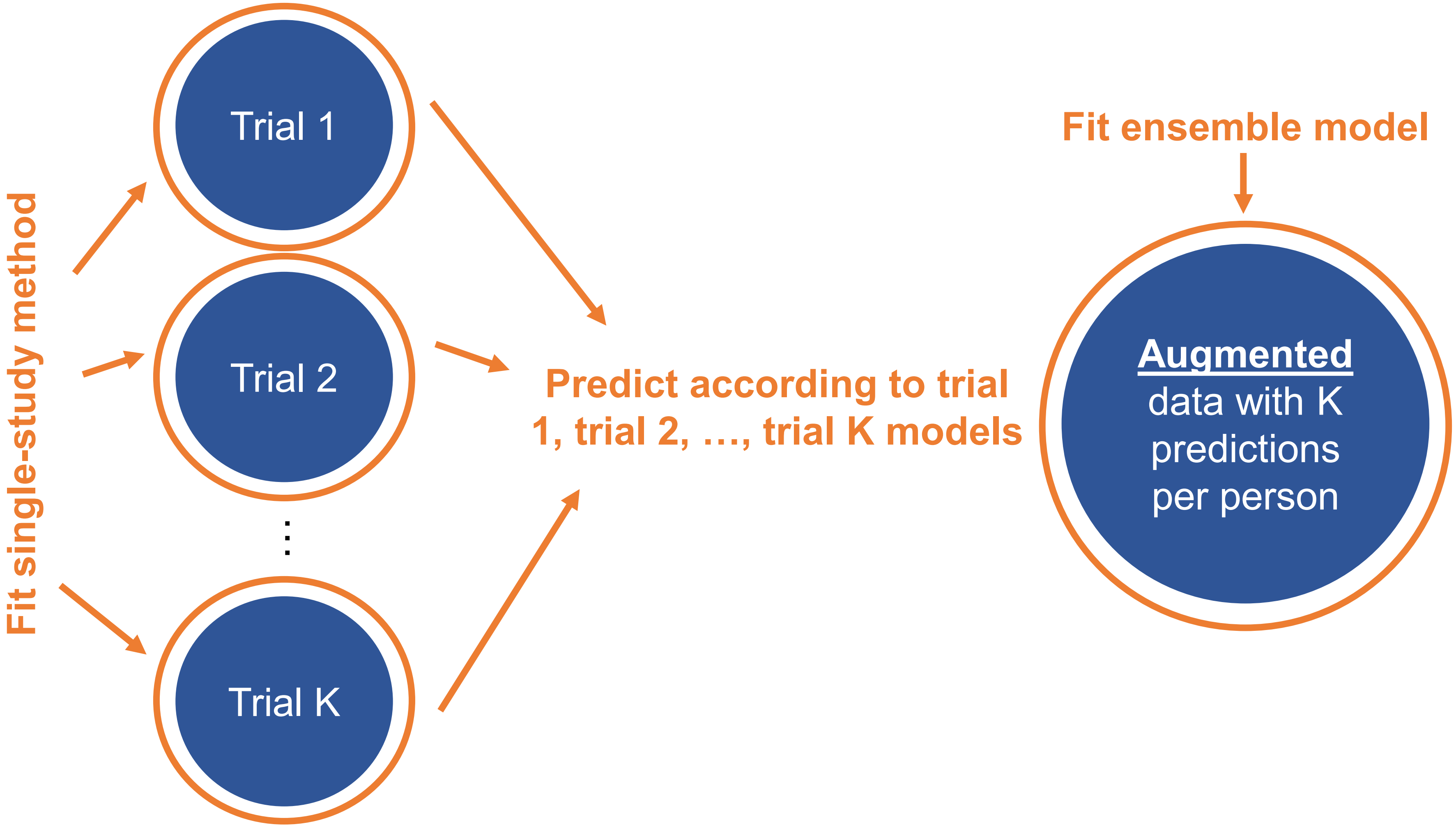
Aggregation Methods

Pooling with Trial Indicator: Pool all data together but keep *study as an indicator* and include that as a covariate in the single-study approaches



Aggregation Methods

Ensemble Forest: Fit model within each study, apply each model to all individuals, and then fit an ensemble random forest to the augmented data



Aggregation Methods

Meta-Analysis with fixed effects and random effects

$$E[Y_{is}] = (\alpha_0 + a_s) + \boldsymbol{\alpha}^T \boldsymbol{X}_{is} + b_s X_{1is} + (\delta + c_s) W_{is} + (\theta + d_s) X_{1is} W_{is}$$

Aggregation Methods

Meta-Analysis with fixed effects and random effects

$$E[Y_{is}] = (\alpha_0 + a_s) + \boldsymbol{\alpha}^T \boldsymbol{X}_{is} + b_s X_{1is} + (\delta + c_s) W_{is} + (\theta + d_s) X_{1is} W_{is}$$

The CATE is: $\tau(\boldsymbol{X}_{is}) = (\delta + c_s) + (\theta + d_s) X_{1is}$

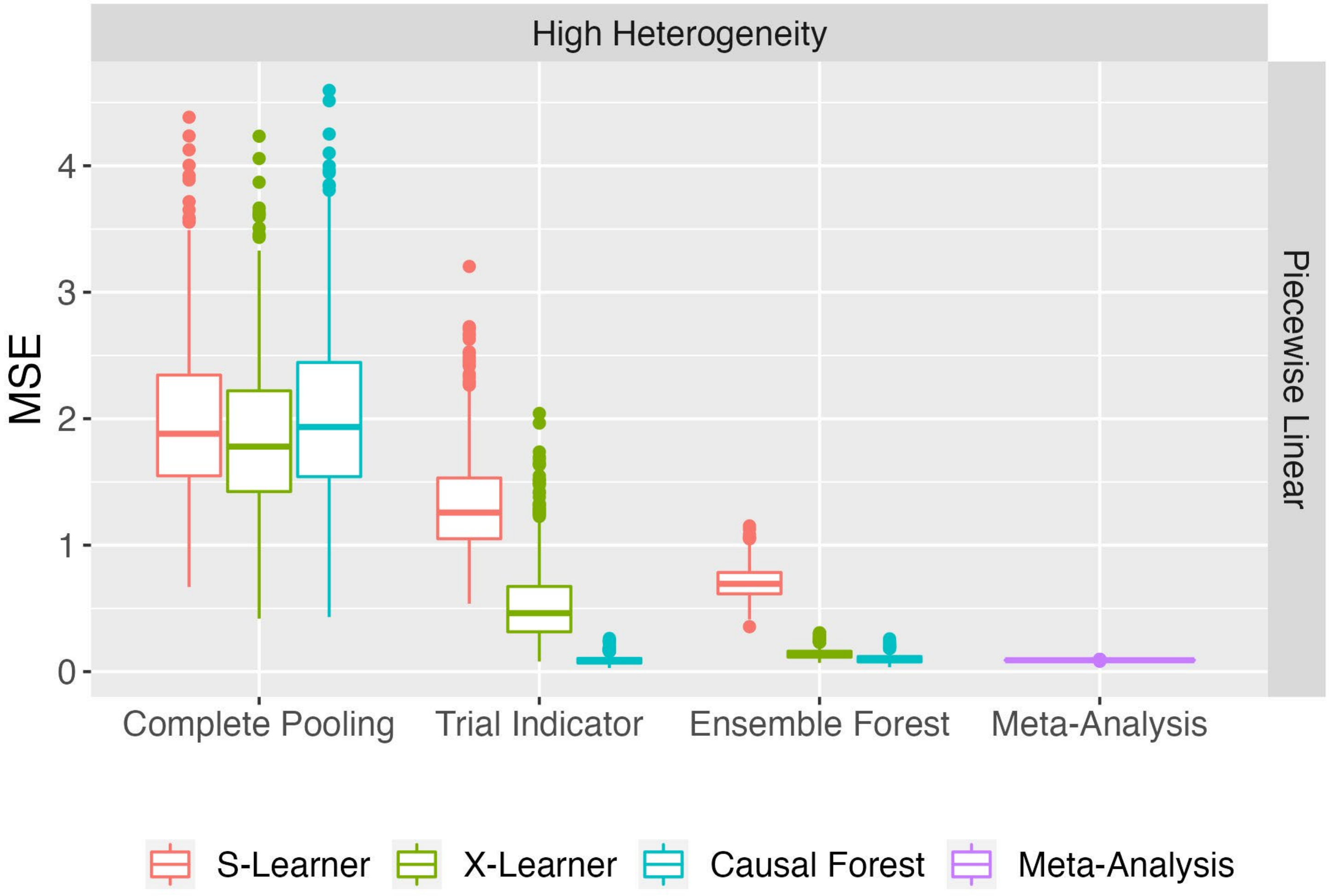
Simulation Setup

- 5 continuous covariates with low correlation
- Probability of treatment is 0.5

Parameters varied:

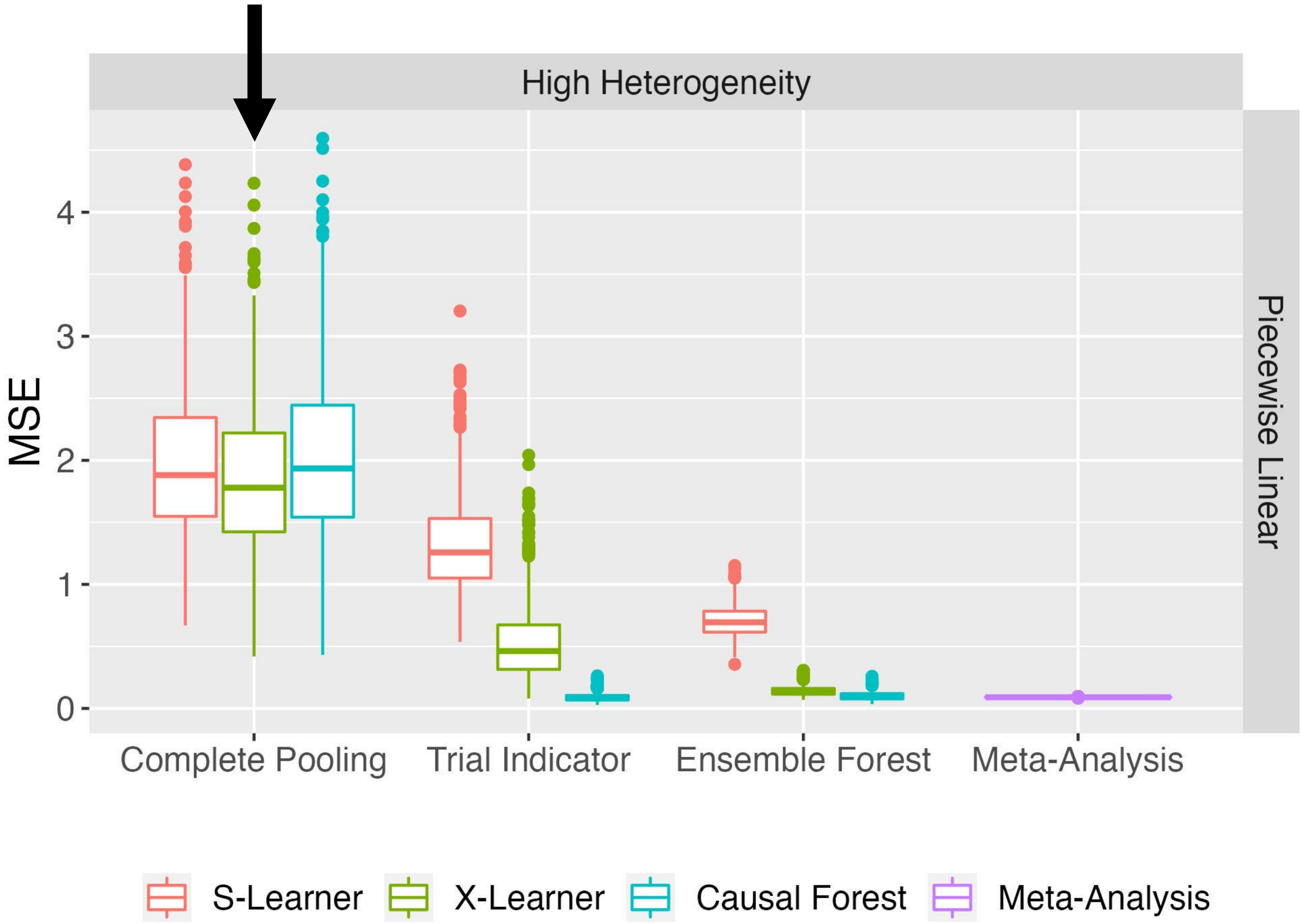
- Heterogeneity of effect across studies (low, medium, **high**)
- Form of CATE (**piecewise linear** or **non-linear**)
- Trial sample sizes (**all 500**, one large, half and half)
- Number of trials (**10** or 30)

Simulation Results



Key Takeaways

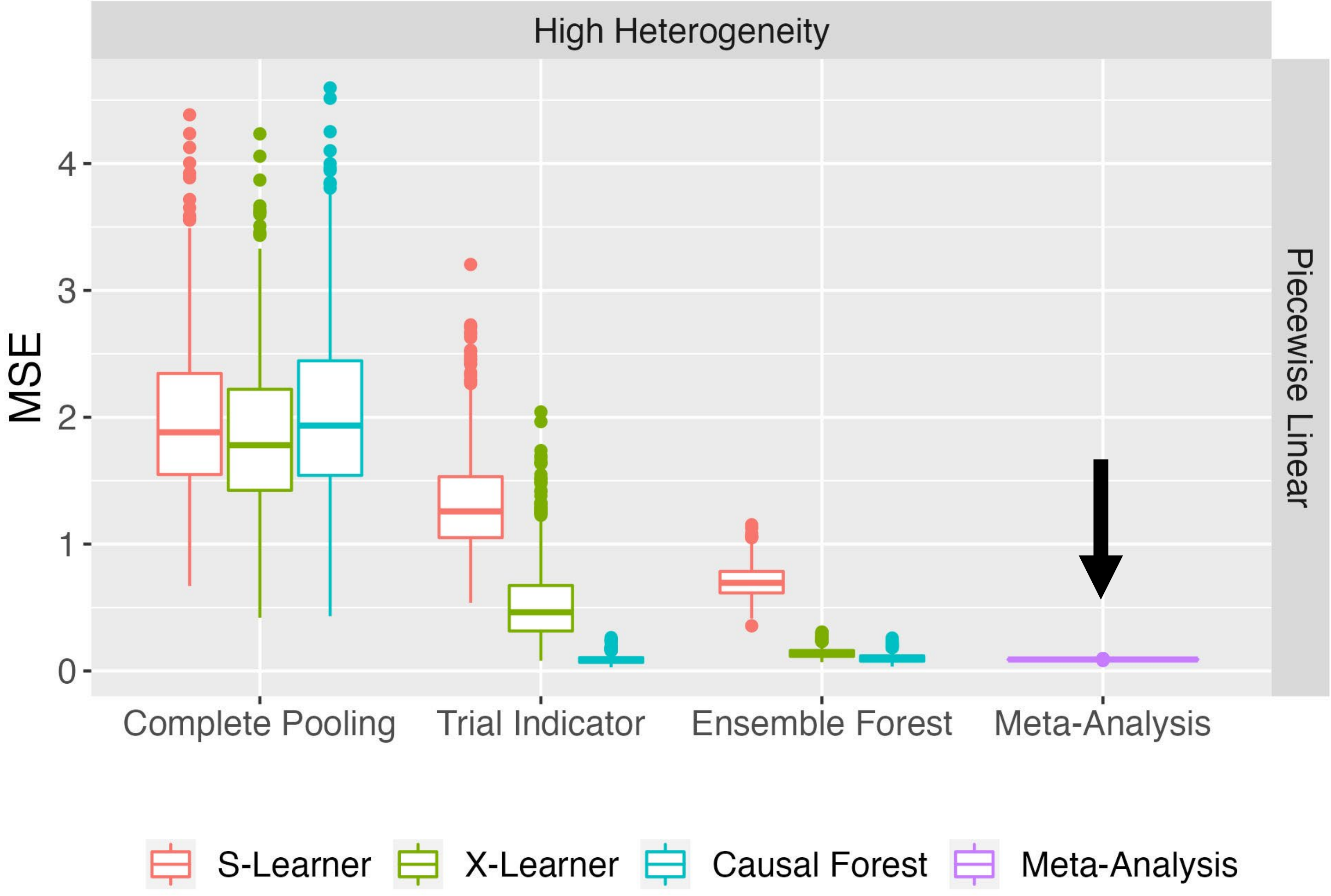
Simulation Results



Key Takeaways

- Complete pooling performs poorly in the presence of heterogeneity of the CATE across trials

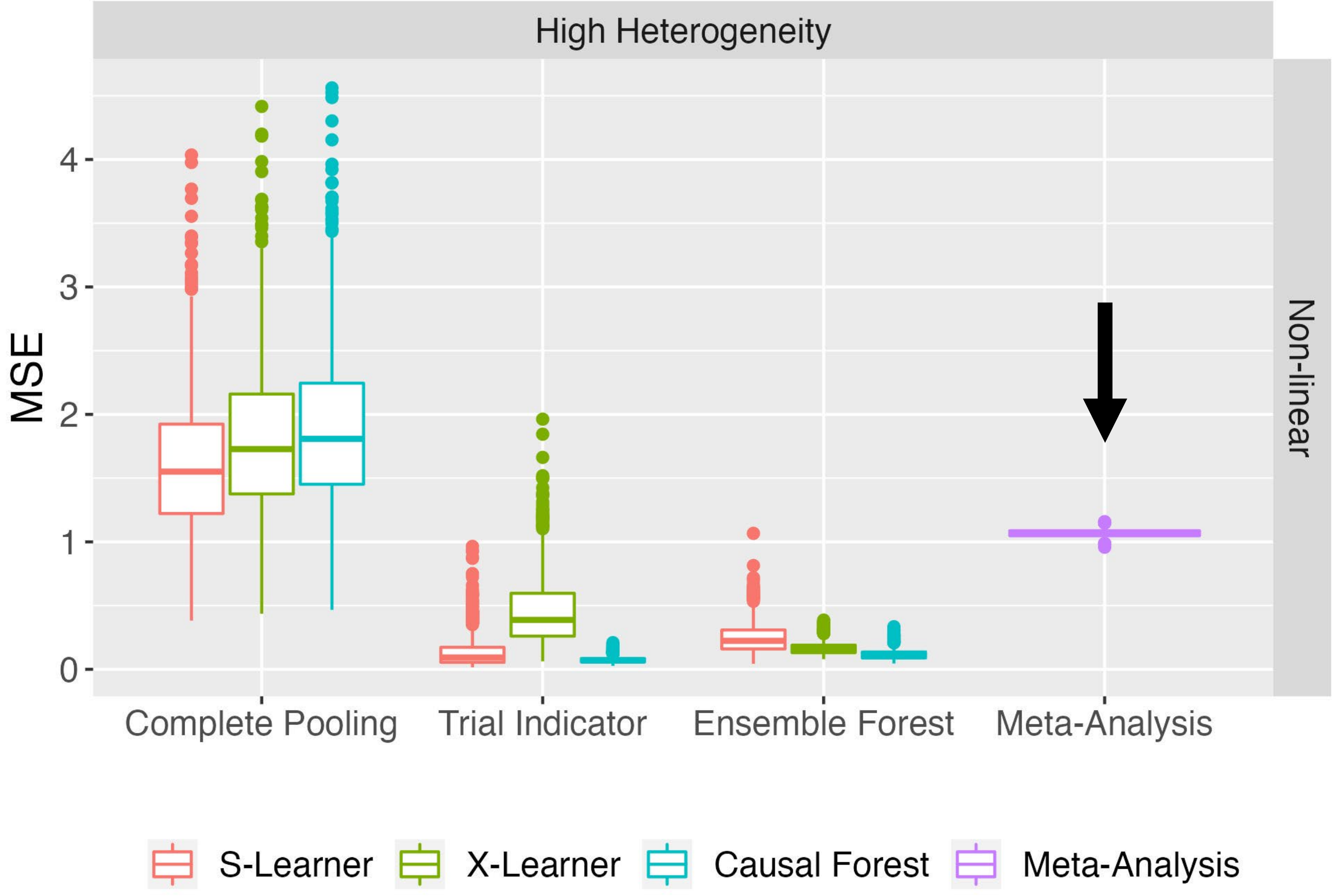
Simulation Results



Key Takeaways

- Complete pooling performs poorly in the presence of heterogeneity of the CATE across trials
- Meta-analysis performs well when correctly specified and poorly when incorrect (non-linear CATE)

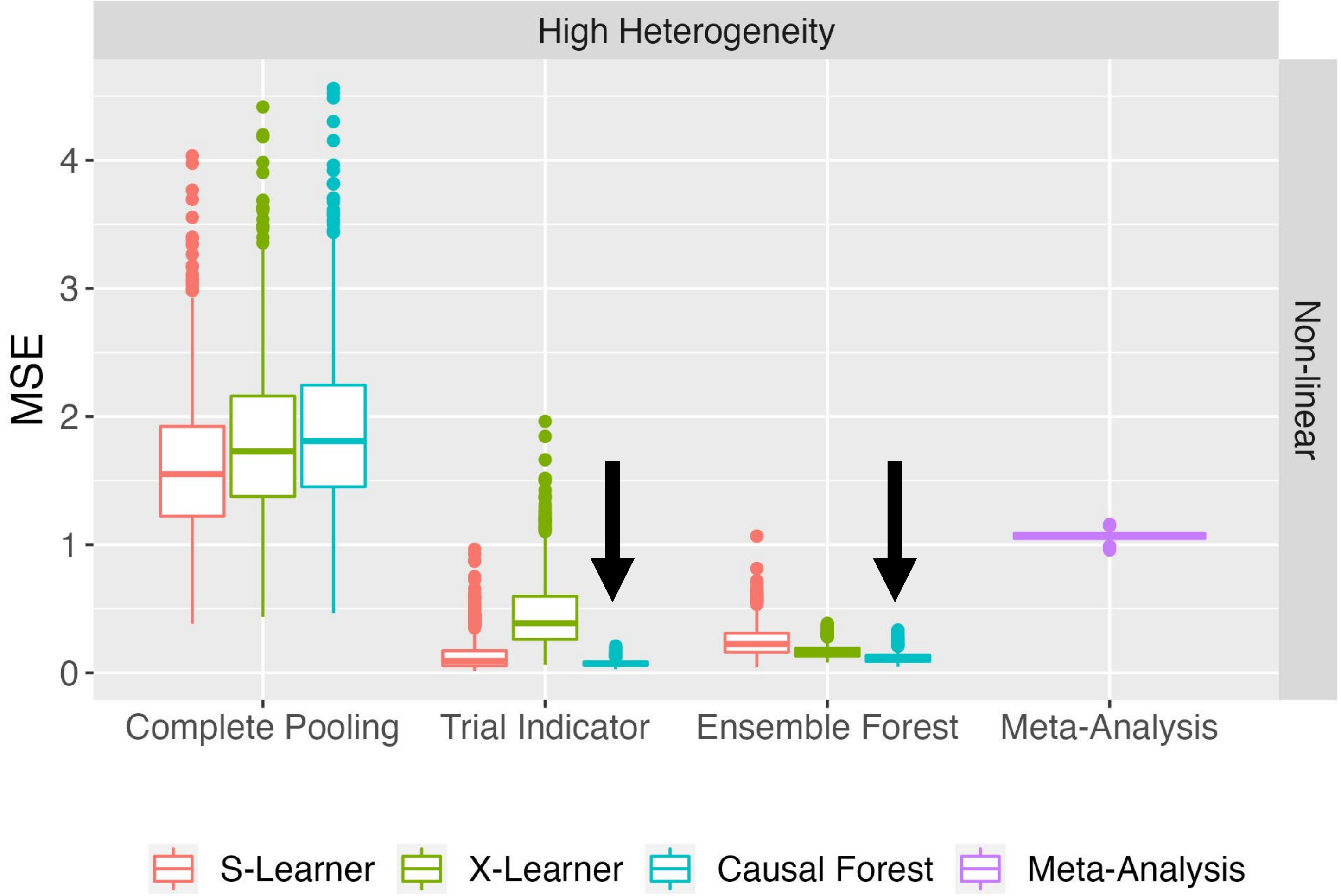
Simulation Results



Key Takeaways

- Complete pooling performs poorly in the presence of heterogeneity of the CATE across trials
- Meta-analysis performs well when correctly specified and poorly when incorrect (non-linear CATE)

Simulation Results



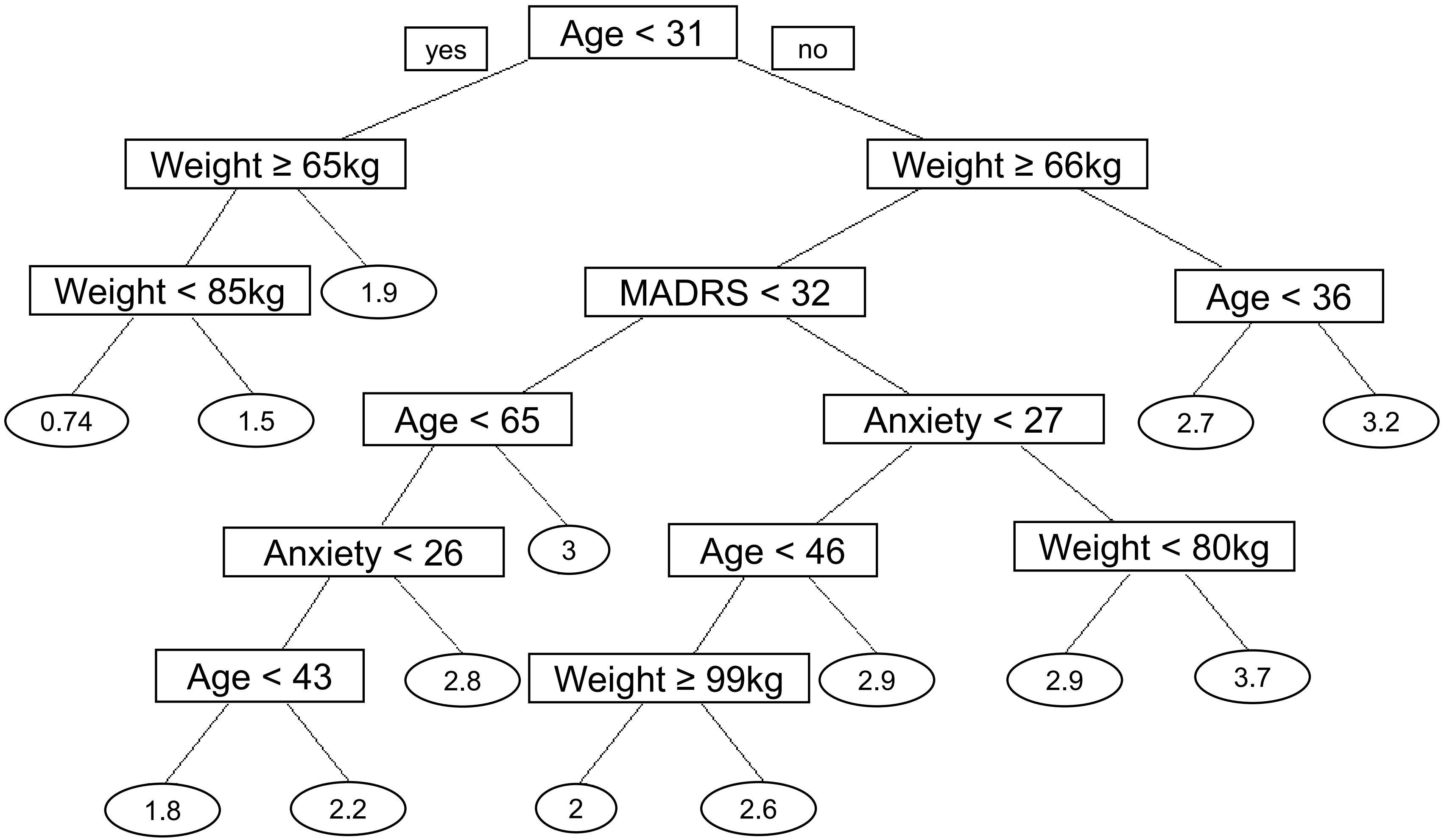
Key Takeaways

- Complete pooling performs poorly in the presence of heterogeneity of the CATE across trials
- Meta-analysis performs well when correctly specified and poorly when incorrect (non-linear CATE)
- Causal forest performs consistently best with pooling with trial indicator or ensemble forest

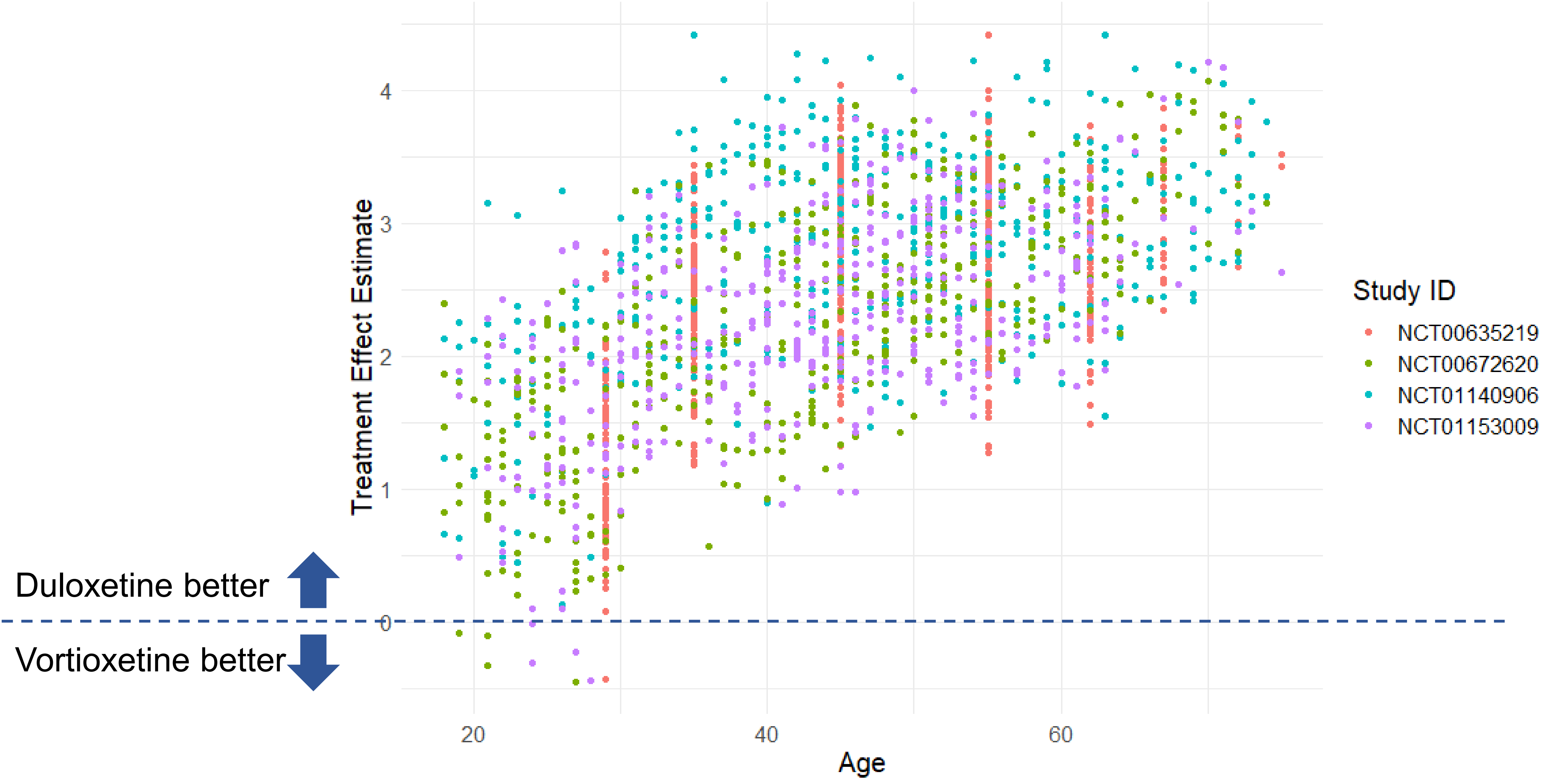
Motivating Application: Depression Treatments

- Applied methods explored in simulations to four RCTs comparing Vortioxetine (“treatment”) and Duloxetine (“control”)
- Focus on *causal forest with pooling with trial indicator* results
- **Key question:** How to interpret the non-parametric CATE estimates?

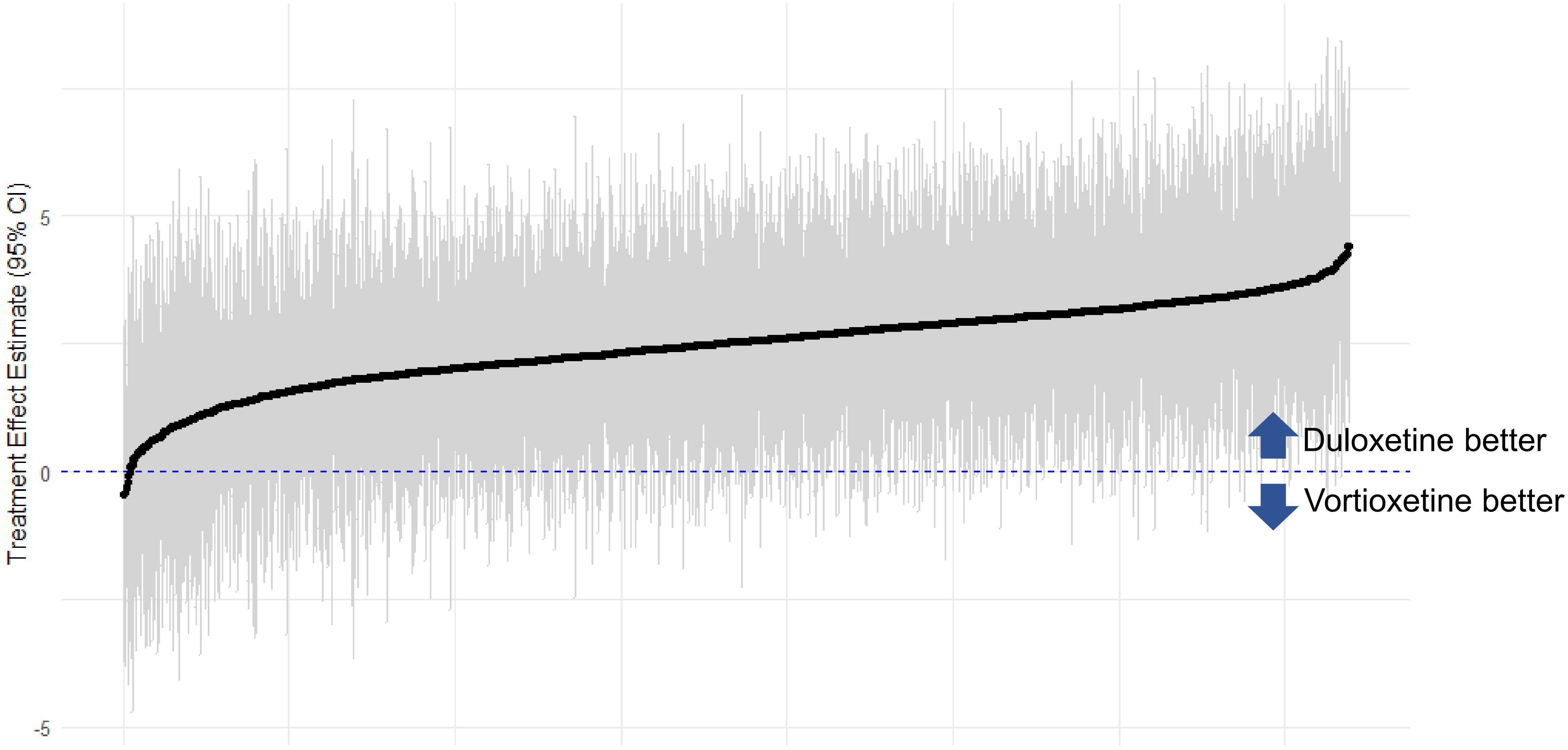
Interpretation Tree



Scatterplot of Treatment Effect by Age



Uncertainty of CATE



Conclusions

Open questions for this to be useful in practice

- How to interpret these results and findings?
- How to best summarize and illustrate them?
- What is the use of the fancy CATE models if in the end we probably go back to simple examination of individual moderators? Exploratory vs. confirmatory?
- How to fully account for uncertainty in the CATE estimates?
- How to predict effects for future individuals, not from an individual study?
- Is this a lot of work and fancy methods when in reality there often isn’t really any effect heterogeneity?

And what about the EHR data?

- Big methods questions about how to combine trial and non-experimental data
- Different populations, confounding in the EHR data
- BUT also fundamental data comparison challenges: different covariates, different outcomes (service utilization vs. symptoms), etc.
- Unclear if there is much to be gained if the outcomes are different (without having to make lots of assumptions)

- So still a work in progress...stay tuned!

General lessons

- Need to be realistic about what we can learn about heterogeneous treatment effects, even when combining data sources

- Fancy methods can only get us so far: Need high quality data, comparable measures, etc.

- Look for methods that are transparent, replicable, and with diagnostics

- Remember the fundamental problem of causal inference!

Thank You!

Email: estuart@jhu.edu
Website: www.elizabethstuart.org
X: @lizstuartdc
LinkedIn: @estuartdc

References

Brantner, C. L., Nguyen, T. Q., Tang, T., Zhao, C., Hong, H., and Stuart, E. A. (2024). Comparison of methods that combine multiple randomized trials to estimate heterogeneous treatment effects. *Statistics in Medicine.*

Lupton Brantner, C., Chang, T-Y., Hong, H., Di Stefano, L., Nguyen, T.Q., and Stuart, E.A. (in press). Methods for Integrating Trials and Non-Experimental Data to Examine Treatment Effect Heterogeneity. *Statistical Science*.

S-Learner

1. Estimate single conditional mean outcome function using random forest:

$$\mu(\boldsymbol{X_i}, W_{\boldsymbol{i}}) = E(Y_i | \boldsymbol{X_i}, W_{\boldsymbol{i}})$$

2. Directly calculate the CATE: $\hat{\tau}(\boldsymbol{X}_i) = \hat{\mu}(\boldsymbol{X_i}, 1) - \hat{\mu}(\boldsymbol{X_i}, 0)$

X-Learner

1. Estimate two conditional mean outcome functions using random forests: $\mu(\boldsymbol{X_i}, 1) = E(Y_i(1)|\boldsymbol{X_i})$ and $\mu(\boldsymbol{X_i}, 0) = E(Y_i(0)|\boldsymbol{X_i})$

2. Estimate treatment effects for individuals in each group using the true data and the estimated outcome functions:

$$\widetilde{D}_{i:\,A_i=1} = Y_{i:\,A_i=1} - \hat{\mu}\left(\boldsymbol{X}_{i:\,A_i=1}, 0\right)$$
$$\widetilde{D}_{i:\,A_i=0} = \hat{\mu}\left(\boldsymbol{X}_{i:\,A_i=0}, 1\right) - Y_{i:\,A_i=0}$$

3. Regress with $\widetilde{D}_i$ as outcomes to get $\hat{\tau}_1(\boldsymbol{X_i})$ and $\hat{\tau}_0(\boldsymbol{X_i})$

4. Define CATE as weighted average of $\hat{\tau}_1$ and $\hat{\tau}_0$

Causal Forest

- Causal tree involves recursive partitioning of the covariates to best split based on treatment effect heterogeneity (difference in average outcomes between treatment and control groups within leaves)
- Causal forest is an aggregation of causal trees using weights [Athey et al., 2019]
- Orthogonalization: before running the forest, two regression forests are trained to estimate propensity scores and marginal outcomes
  - Then compute residuals W-e(X) and Y-m(X) and train a causal forest on those (R-learner) [Nie and Wager, 2021]

Bayesian Additive Regression Trees (BART)

- Sum-of-trees model
- Uses regularization prior to restrict the amount of relationships that each tree can explain
  - (1) Prior prefers trees with few bottom nodes; (2) Shrinks terminal means towards 0; (3) Suggests standard deviation is less than least squares estimate
- Estimates the outcome and provides posterior draws to produce credible intervals

- Two options:
  - S-Learner: $\mu(\boldsymbol{X_i}, W_i) = E(Y_i | \boldsymbol{X_i}, W_i)$
  - T-Learner: $\mu(\boldsymbol{X_i}, 1) = E(Y_i(1) | \boldsymbol{X_i})$ and $\mu(\boldsymbol{X_i}, 0) = E(Y_i(0) | \boldsymbol{X_i})$