The promise and pitfalls of integrating data for causal inference, with application to mental health treatment

Elizabeth Stuart

Hurley-Dorrier Professor and Chair

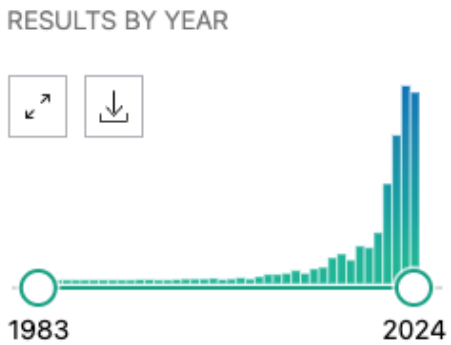
Department of Biostatistics

estuart@jhu.edu; @lizstuartdc

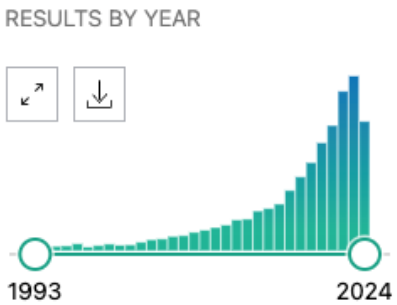
JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Context

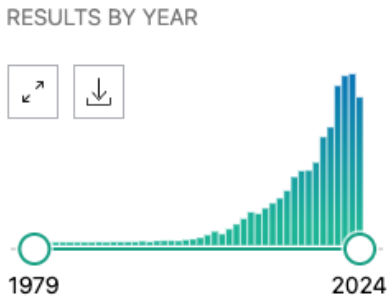
Rising interest in combining data: pubmed trends, National Academies panels



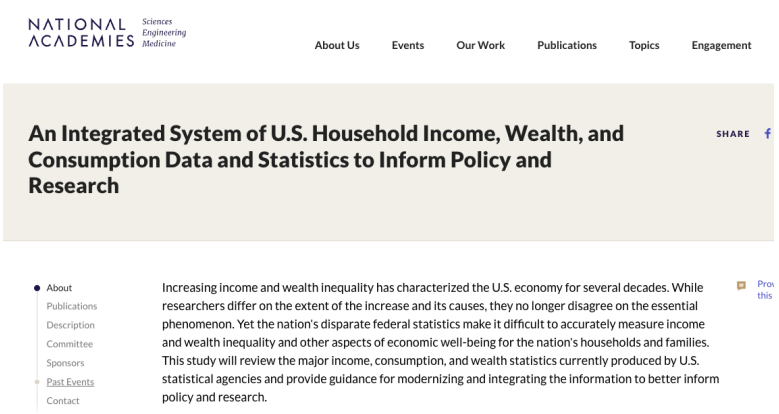
Multimodal data

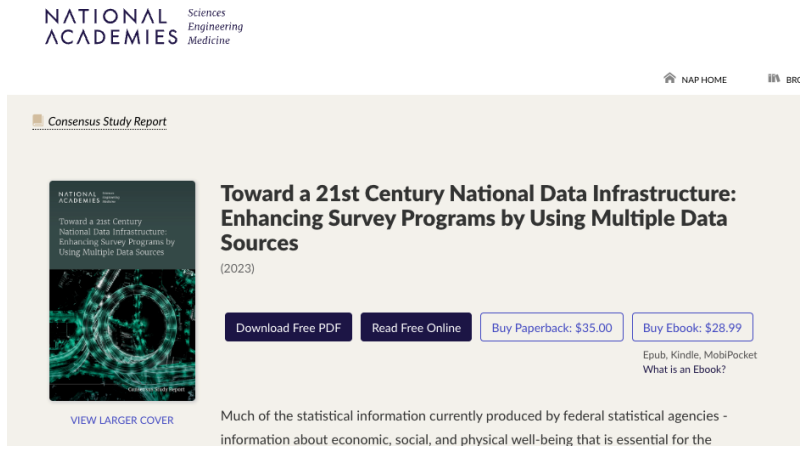


Data fusion

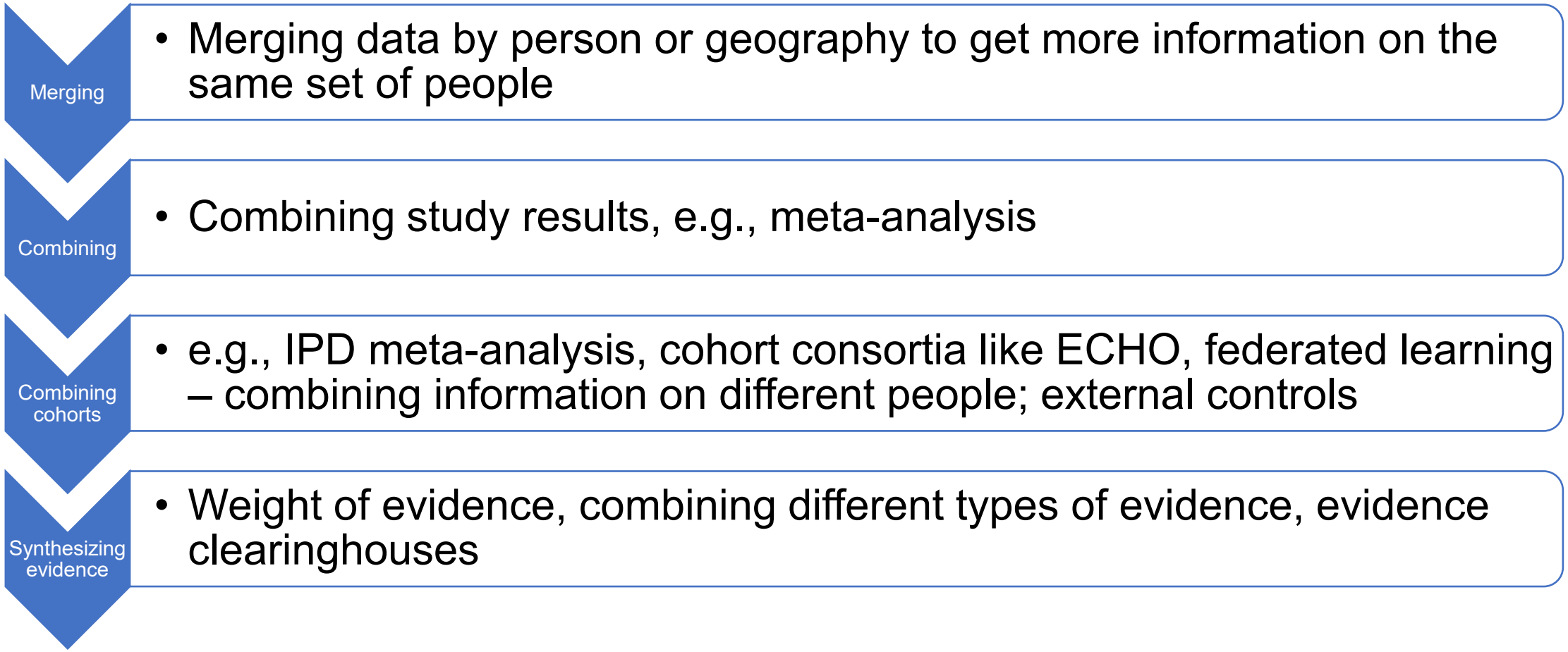


Data integration





Huge spectrum of what is meant by integration



Causal related goals of data integration

- Assessment of RCT generalizability by combining trial and population data
- Identification of a comparison group (e.g., “external controls”)
- Additional control for confounders in non-experimental studies
- Understanding consistency of results across studies or settings
- Increased power by combining data sources
- Cross-design synthesis
  - Combining randomized trials and observational studies for larger samples, more representative results

That sounds great…what's the catch?

Measurement differences

- Large-scale datasets often don't collect the same information -- Stuart and Rhodes (2017) found 6 variables in common between an RCT and large-scale observational data
- RCTs and observational studies (e.g., EHR) often collect different outcomes, even on different time scales

Population differences

- e.g., Cohort consortia such as ECHO are combining very different cohorts, from different geographic locations and different disease categories
- Hard to disentangle true differences from differences due to other factors

Data access

- Very hard to access datasets
- Data often not available publicly
- Federated learning approaches can help but are limited in types of analyses that can be done
- Huge time and effort required to create harmonized datasets

Case study: Combining RCTs and EHR to study treatment effect heterogeneity

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?
- And can we do with reasonable precision?
- 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

But when is it actually helpful?? My own journey….

There are so many methods out there…

Review of Data Integration Methods for estimating the CATE

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

Let’s start with a simpler scenario…

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
- [Note: Also examining Bayesian approaches, led by Hwanhee Hong at Duke; not discussing those today]

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

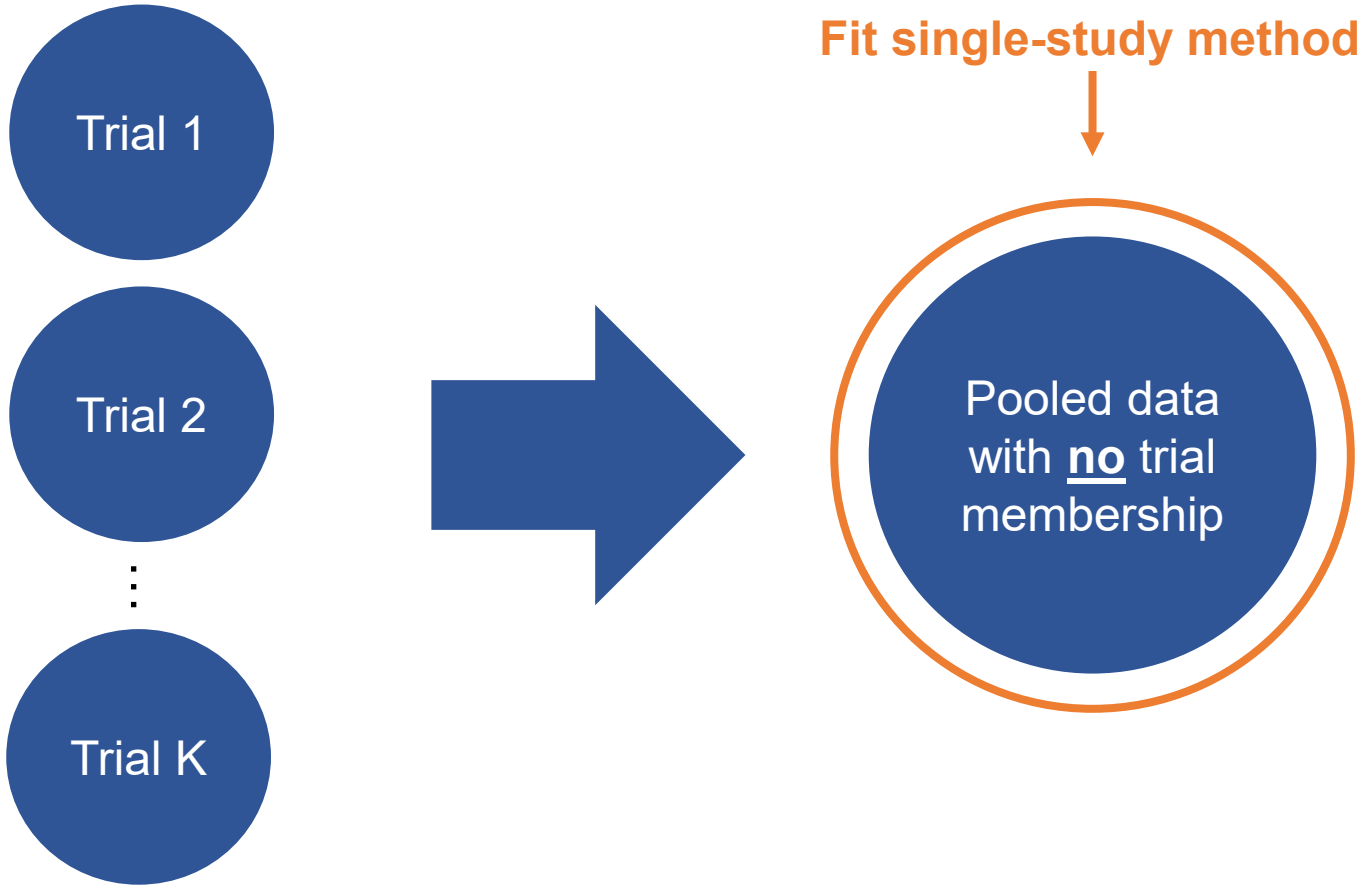
Single-Study Methods

- 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]; fit one combined model with covariates and treatment
  - **X-Learner** [Kunzel et al., 2019]; fit separate models in treatment and control groups

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

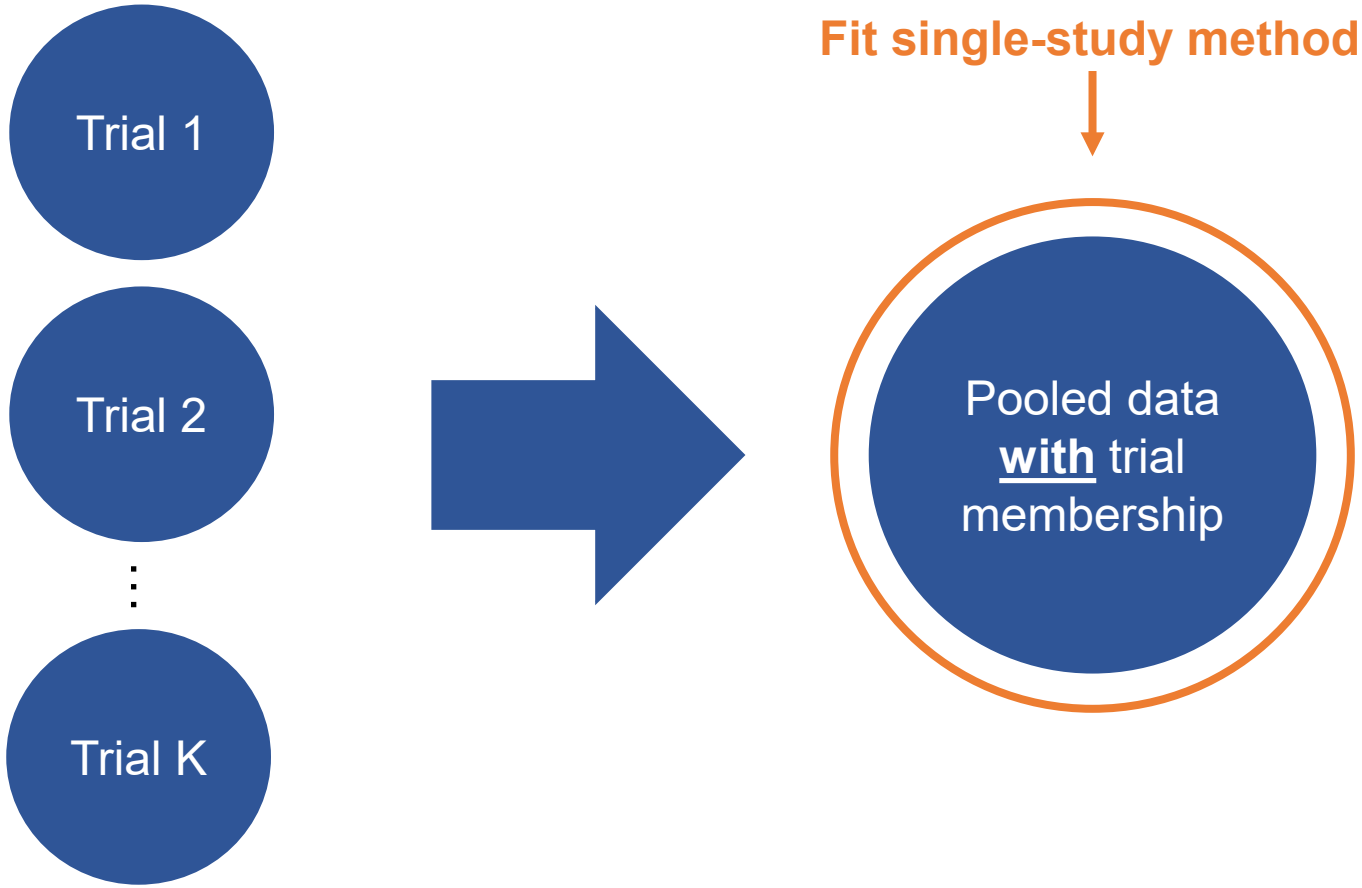
Aggregation Methods

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



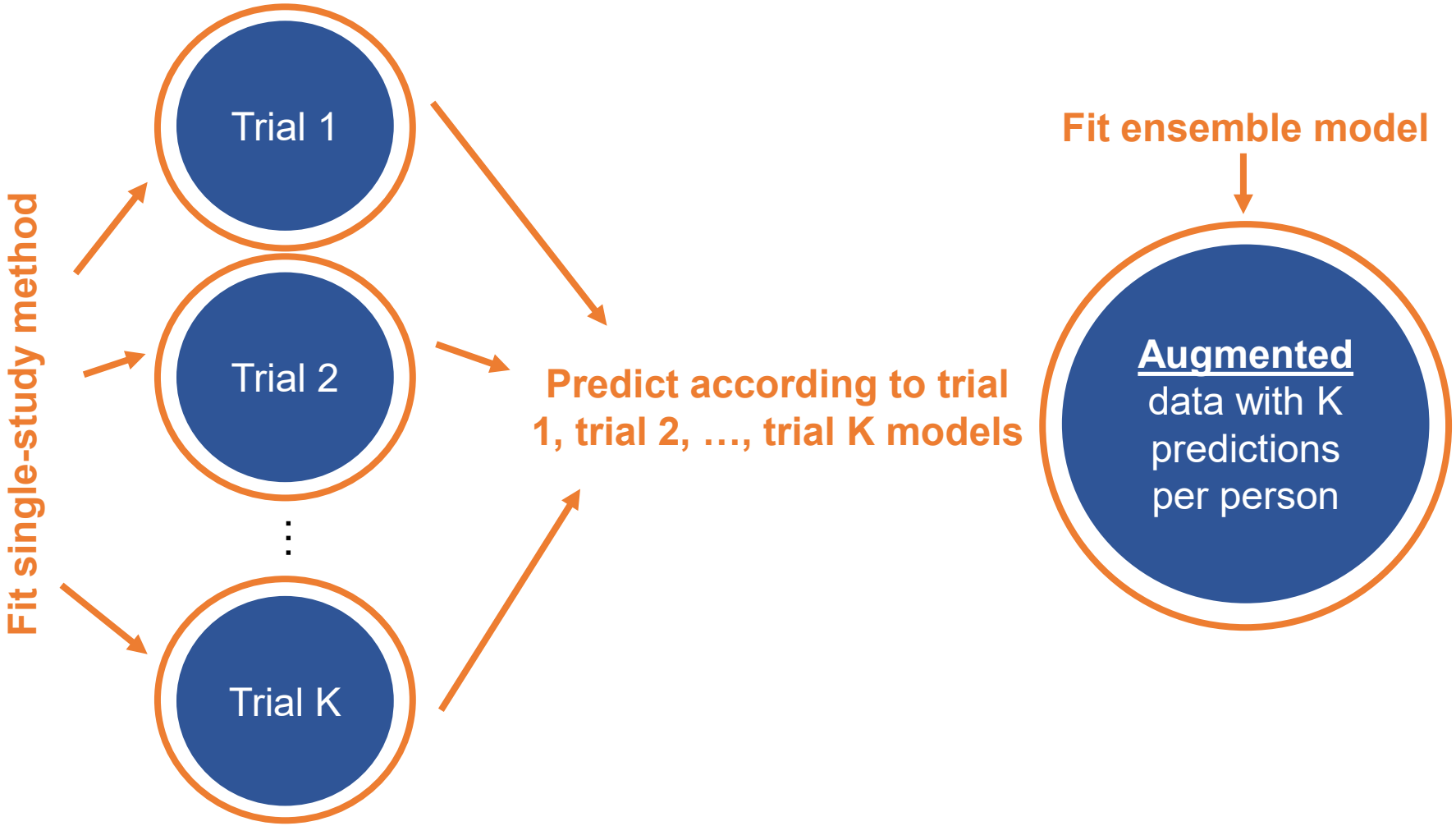
Aggregation Methods

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



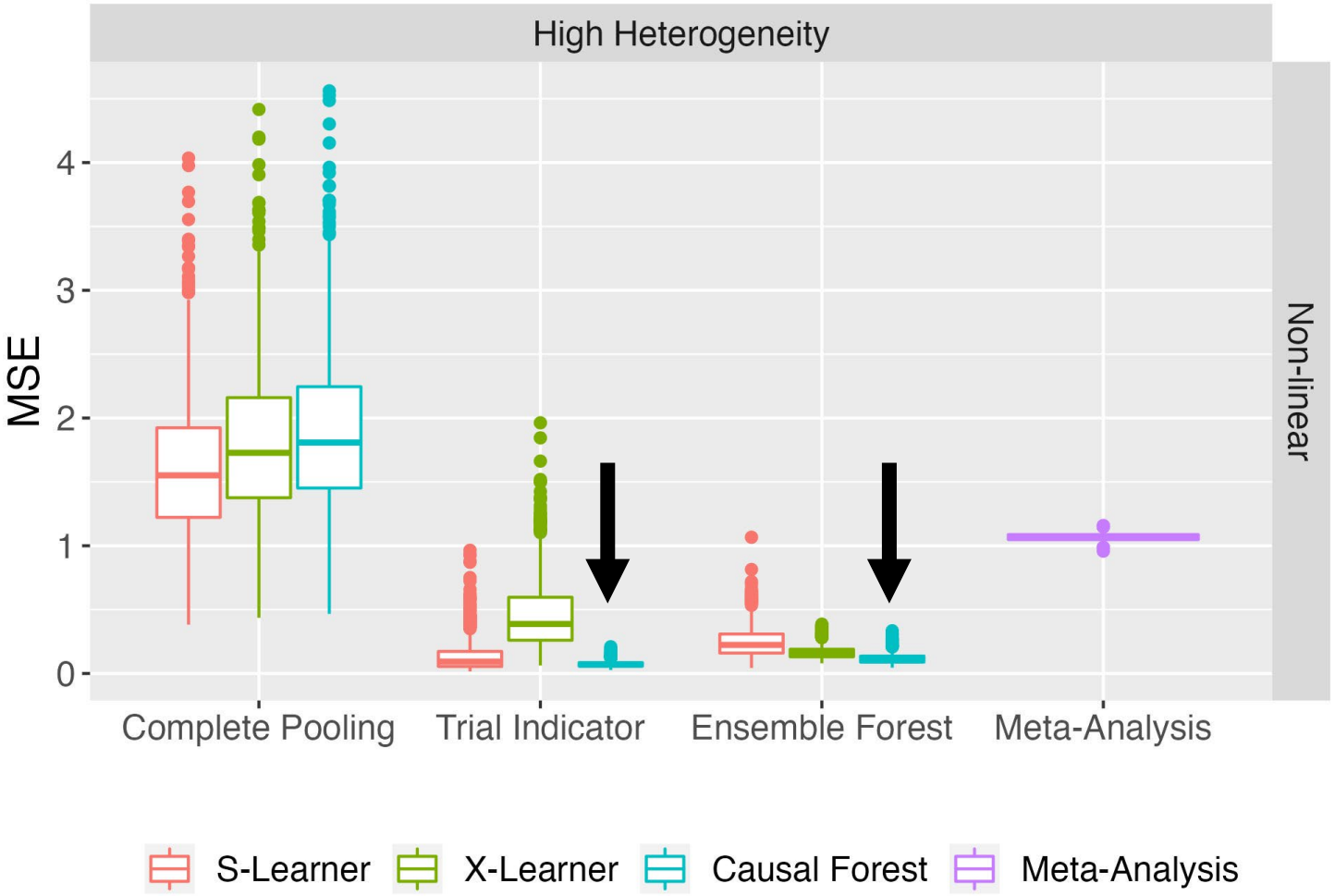
Aggregation Methods

Meta-Analysis with fixed effects and random effects

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

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

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

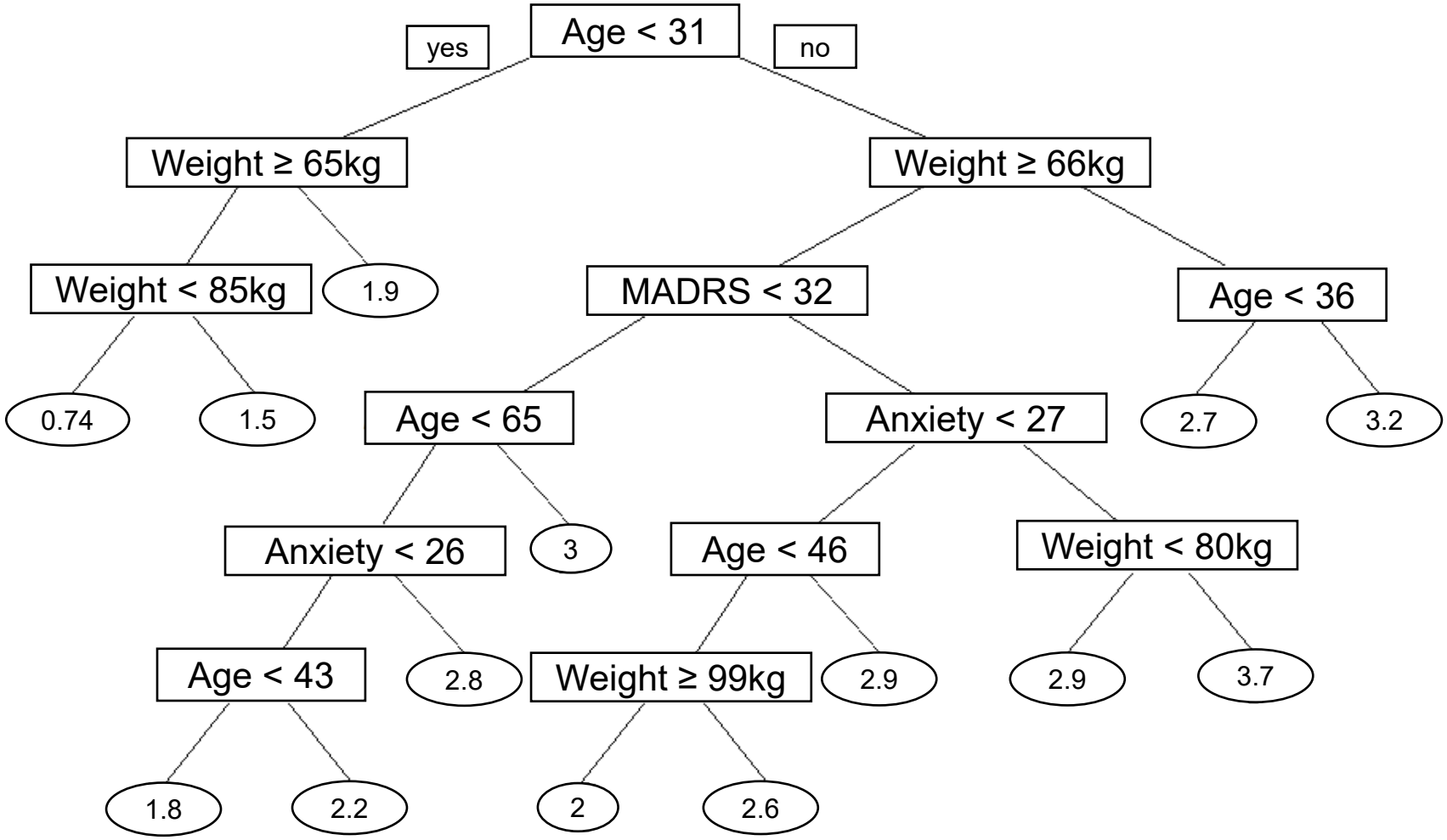
Okay so we can combine the data and fit complex models…

But how do we make sense of the results?

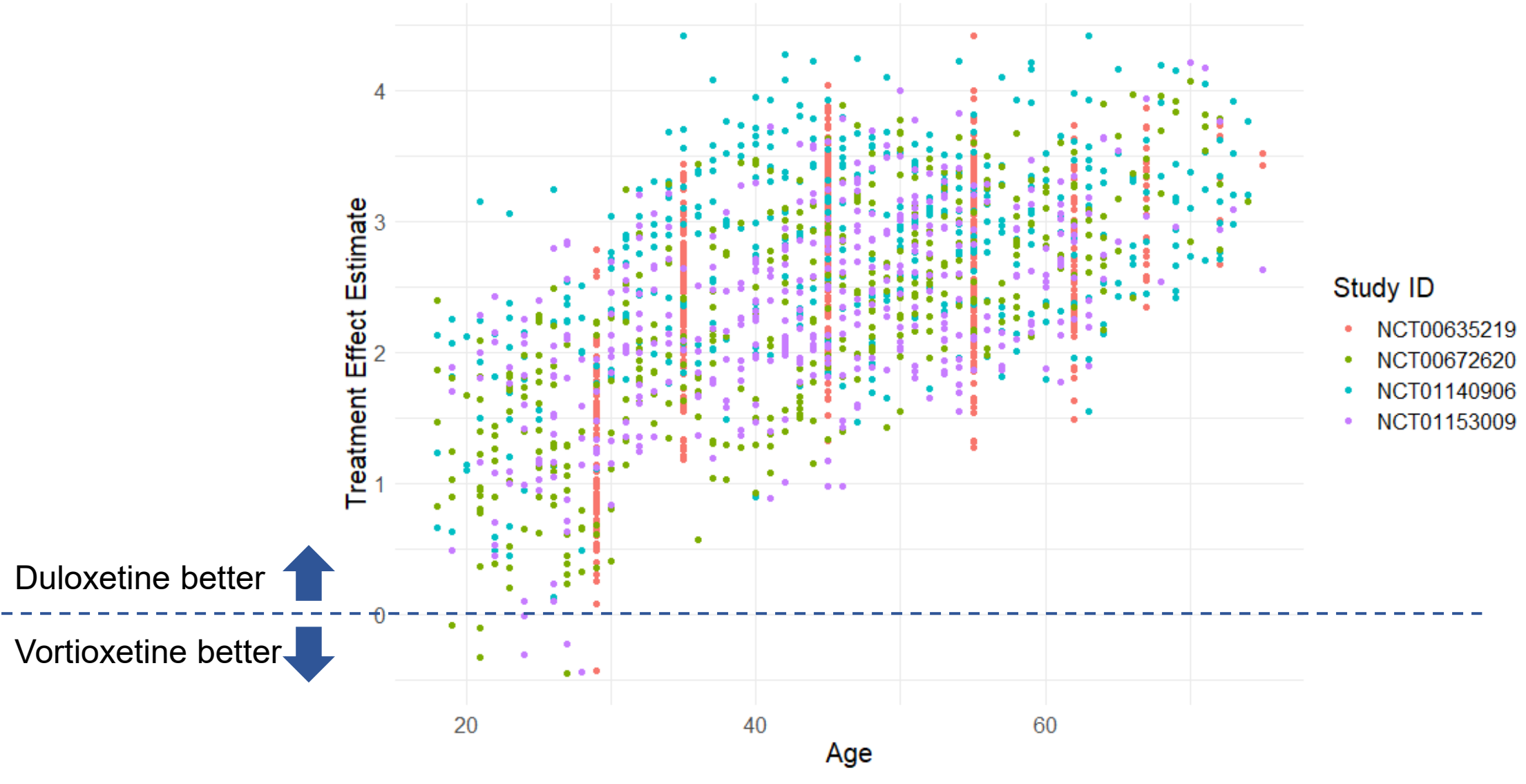
Motivating Application: Depression Treatments

- 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?

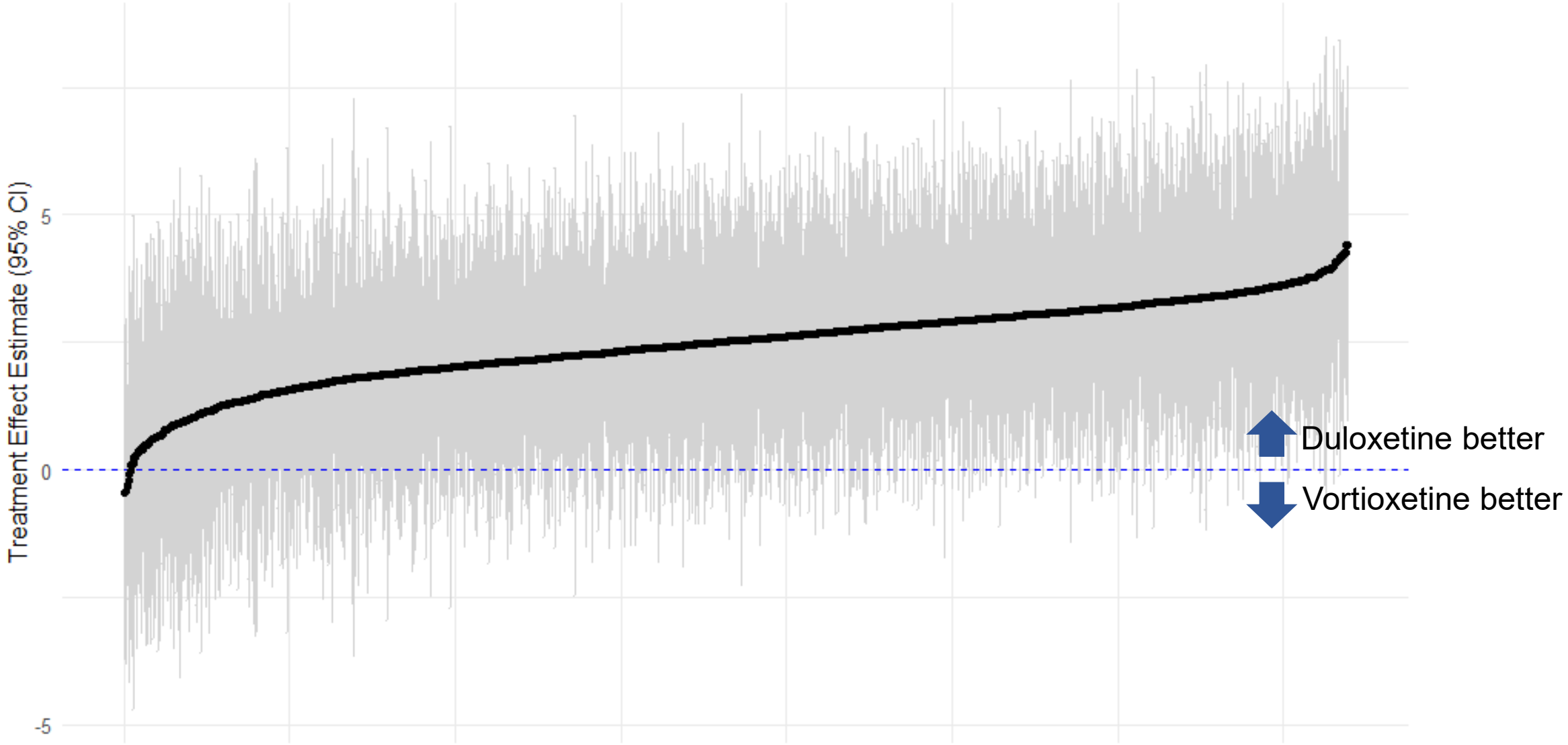
Interpretation Tree



Scatterplot of Treatment Effect by Age



Uncertainty of CATE



But…we don’t really care about the people in the trials…

…what about predicting CATEs for future patients (who aren’t in a trial)?

(And we’ll start to bring in the EHR data)

Question

How can we apply our models to individuals who come from outside of the training sample?

We need to account for study, but future patients aren't in a study!

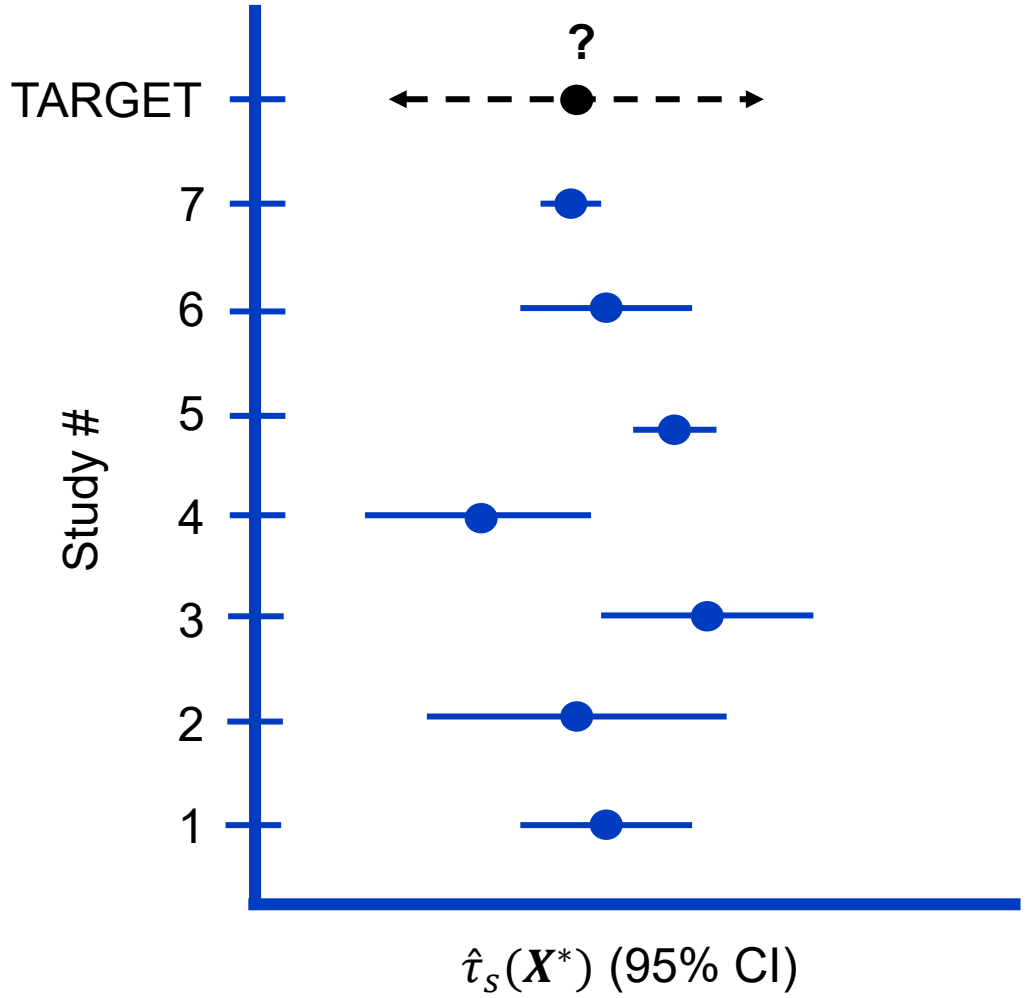
Example: Target group of patients in a hospital system at baseline

Main idea: Marginalize over trial in the predicted CATEs

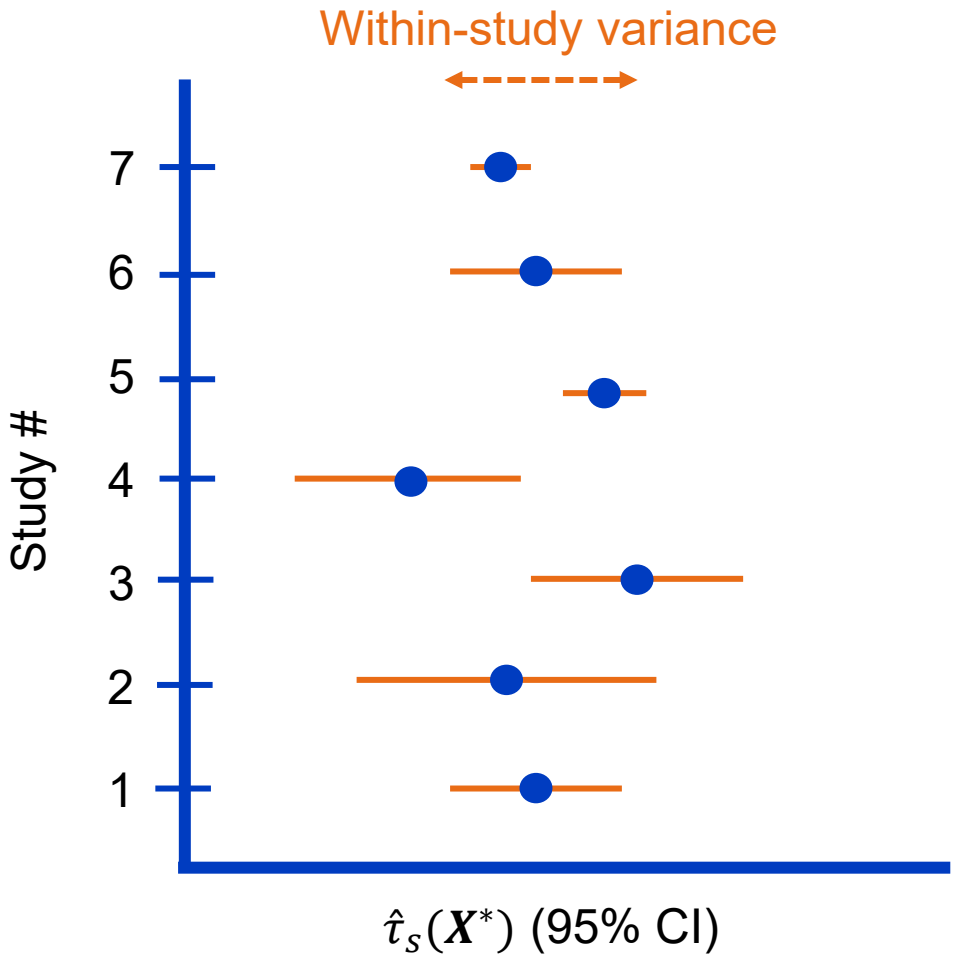
Partitioning Variance

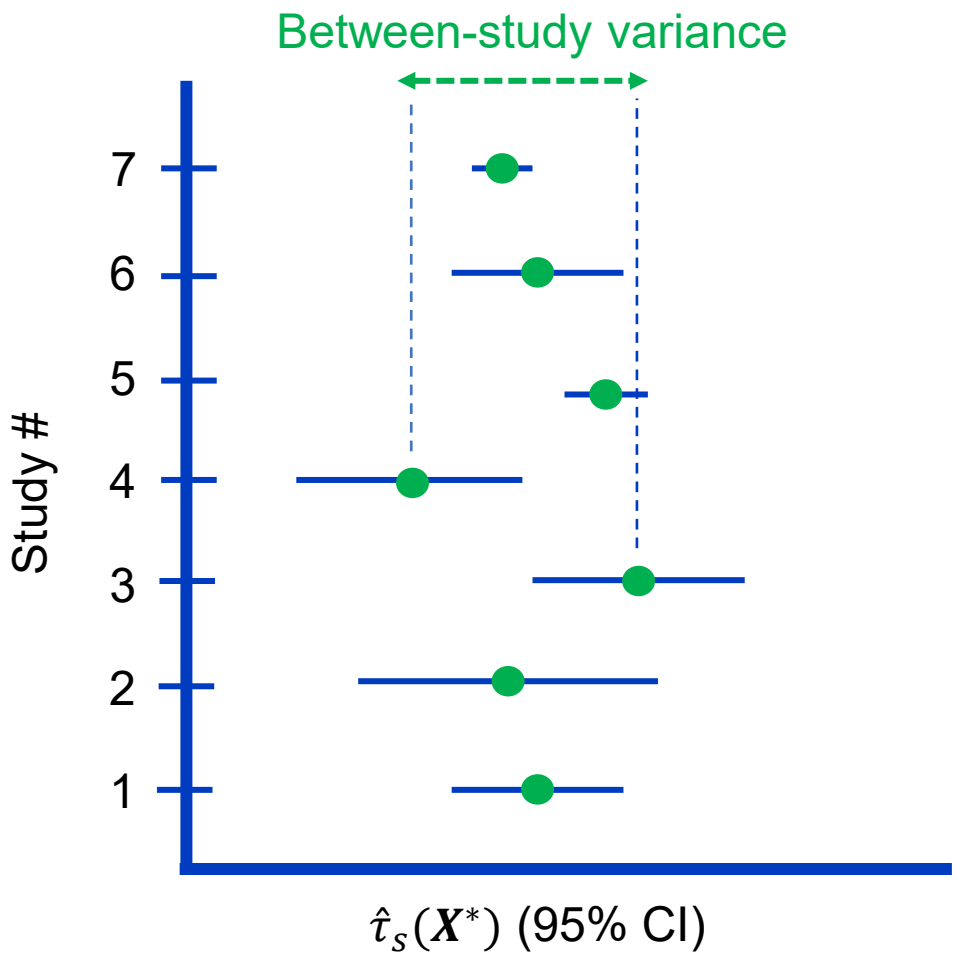
Consider a covariate profile $\boldsymbol{X}^*$.

How do we account for both the within-study uncertainty for $\tau_{K+1}(\boldsymbol{X}^*)$ and the between-study differences?



Partitioning Variance





Motivating Application: Depression Treatments

Target setting: Patient profiles with depression at Duke

Who in our target setting should take *Vortioxetine*, and who should take *Duloxetine*?

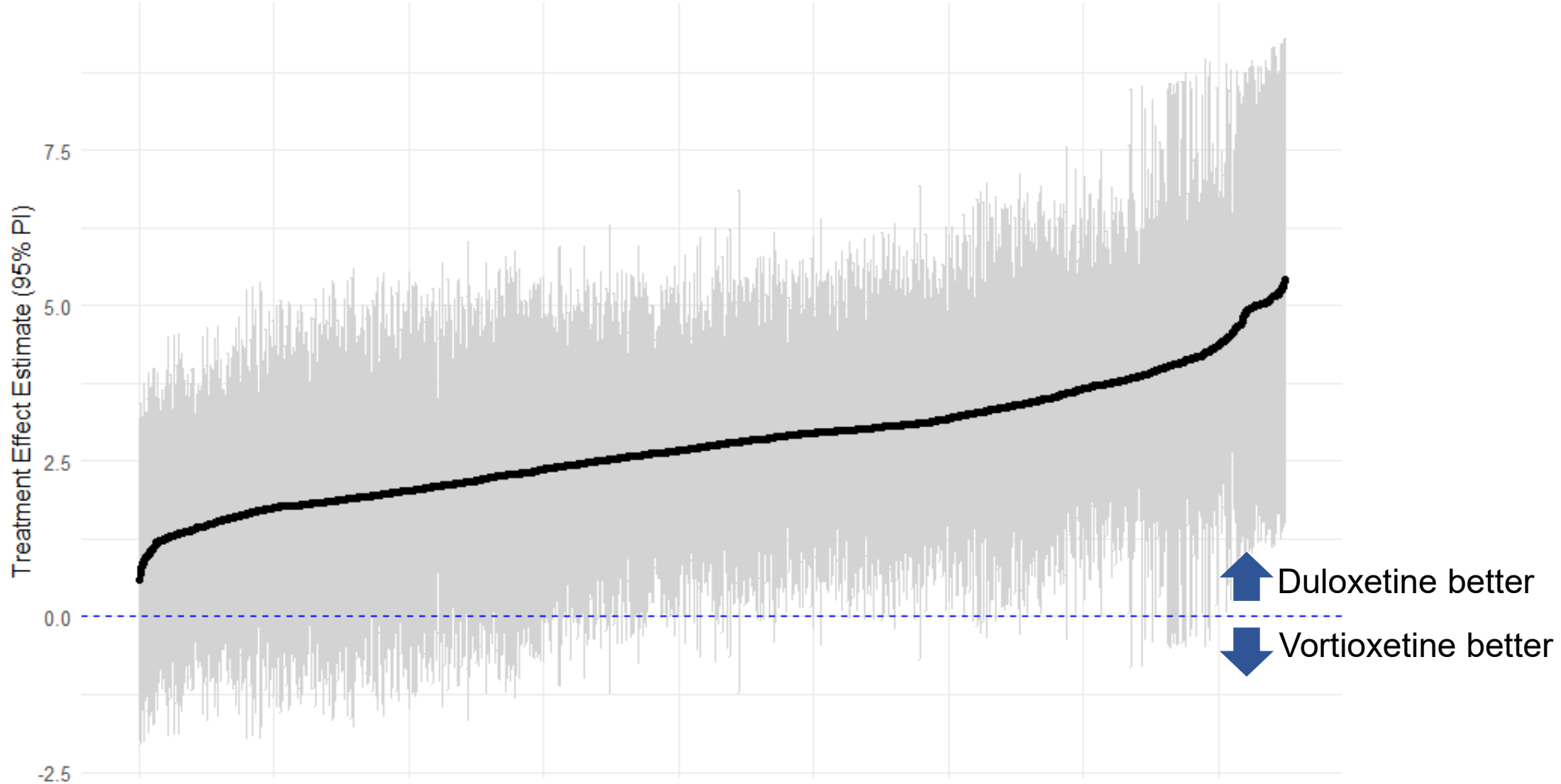
1. Apply model to estimate CATE in RCTs
2. Estimate treatment effect prediction intervals in Duke patient profiles

Sample Characteristics

	Trial Participants (N = 1,272)	Duke Patients (N = 2,123)
Age	44.3 (14.3)	44.7 (12.7)
BMI	29.1 (7.4)	31.5 (8.8)
Female	67.9%	75.7%
Severe Baseline Depression (vs. Moderate)*	19.7%	27.3%
Has Anxiety	2.0%	61.3%
Has Diabetes	2.8%	20.2%

*Trials measured depression using MADRS, and Duke measured depression using PHQ-9. Duke patients with less than moderate depression were excluded.

Prediction Intervals in Duke Patient Profiles



Patient Profiles by Treatment Effect Significance

	Duloxetine Significantly Better (N = 1,380)	Prediction Interval Crosses 0 (N = 743)
Age	49.2 (10.7)	36.3 (11.7)
BMI	30.3 (8.2)	33.5 (9.2)
Female	73.0%	80.8%
Severe Baseline Depression (vs. Moderate)	29.7%	22.7%

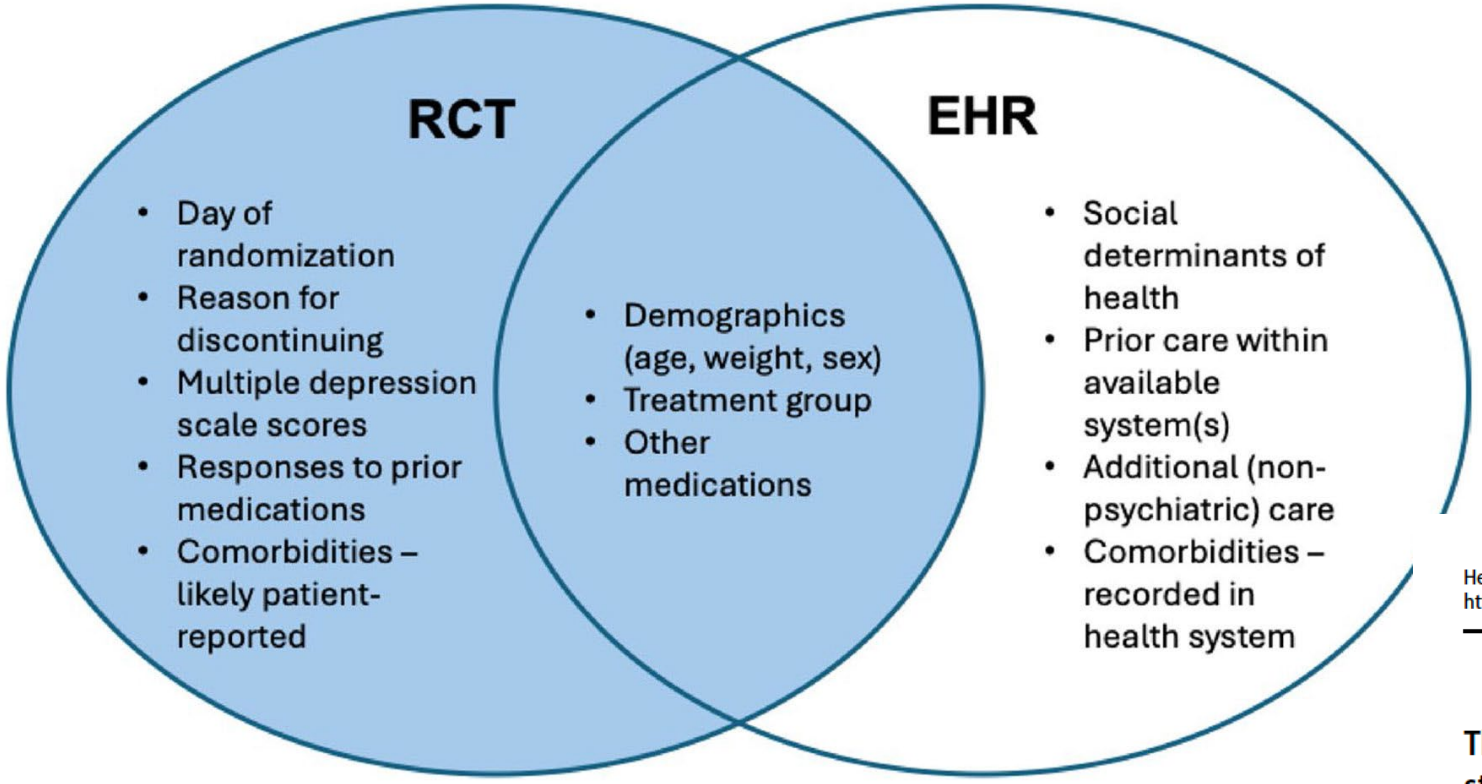
What about using the RCT + EHR data to estimate CATEs?

Wasn't that the original goal?

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.

Brantner et al. Figure 1: Disparate covariates



Health Services and Outcomes Research Methodology
https://doi.org/10.1007/s10742-025-00363-8

Check for updates

The challenges of integrating diverse data sources: A case study in major depression

Carly L. Brantner[1,2] · Wenshan Yu[1] · Congwen Zhao[1] · Kyungeun Jeon[3] · Grace V. Ringlein[3] · Qiao Wang[4] · Elaona Lemoto[1] · Trang Quynh Nguyen[3,5] · Jane P. Gagliardi[6] · Peter P. Zandi[7] · Benjamin A. Goldstein[1,2] · Elizabeth A. Stuart[3,5,8] · Hwanhee Hong[1,2]

Received: 21 March 2025 / Revised: 23 September 2025 / Accepted: 3 October 2025

And different outcomes…

What if two datasets measure outcome differently:

Primary

unit-id	X_1	…	X_p	T	Y	W
1	x_{11}	…	x_{1p}	t_1	y_1	?
2	x_{21}	…	x_{2p}	t_2	y_2	?
…	…	…	…	…	…	?
n_0	…	…	…	…	…	?

Auxiliary

unit-id	X_1	…	X_p	T	Y	W
n_0+1	$X_{n0+1,1}$	…	$x_{n0+1,p}$	t_{n0+1}	?	w_{n0+1}
…	…	…	…	…	?	…
$n_0 + n_1$	…	…	…	…	?	…

Linking Primary and Auxiliary Outcomes

$$\mathbb{E}[Y \mid X] = \alpha(X)\mathbb{E}[W \mid X] + \beta(X)$$

Assumptions (from strongest to weakest)

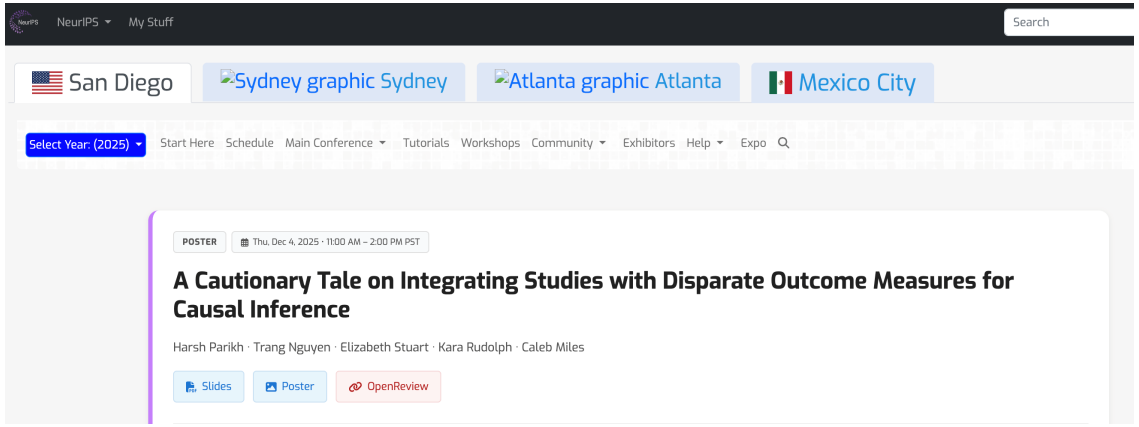
A.a. α and β are a priori known

A.b. only β is a priori is known; α unknown

A.c. both α and β unknown

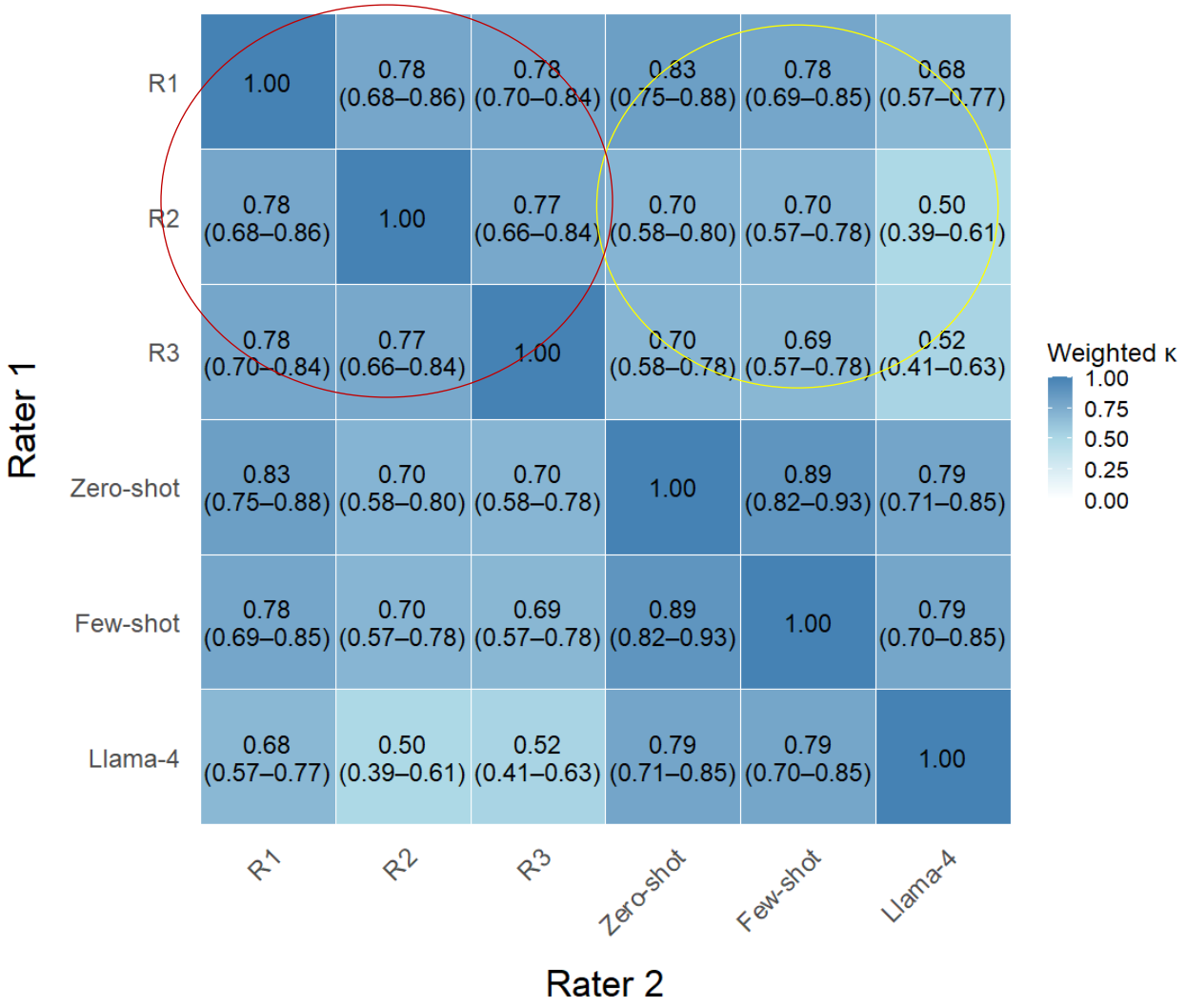
Efficiency Gain *only* under strongest assumption

$$Var(\hat{\theta}_0) = Var(\hat{\theta}_c) \approx Var(\hat{\theta}_b) > Var(\hat{\theta}_a)$$



Can we make them more similar?

- Developing LLM to extract clinical global impression scale from clinical notes
  - Hope is that this will be more similar to the symptom scores available in the RCTs
- But we don't have gold standard measures either!
- Documenting whether the LLM agrees with humans as frequently as humans agree with each other



Conclusions

Open questions for this to be useful in practice

- Can we actually combine the RCT + EHR data?
- How to interpret these results and findings?
- 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?
- Is this a lot of work and fancy methods when in reality there often isn't really any effect heterogeneity?

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

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*.

Brantner, C.L., Chang, T-Y., Hong, H., Di Stefano, L., Nguyen, T.Q., and Stuart, E.A. (2024). Methods for Integrating Trials and Non-Experimental Data to Examine Treatment Effect Heterogeneity. *Statistical Science*.

Brantner, C.L., Nguyen, T.Q., Parikh, H., Zhao, C., Hong, H., and Stuart, E.A. (2025). Precision Mental Health: Predicting Heterogeneous Treatment Effects for Depression through Data Integration. https://arxiv.org/abs/2509.04604

Brantner, C.L., Yu, W., Zhao, C. et al. (2025). The challenges of integrating diverse data sources: A case study in major depression. Health Serv Outcomes Res Methods. https://doi.org/10.1007/s10742-025-00363-8

Parikh, H., Nguyen, T.Q., Stuart, E.A., Rudolph, K.E., and Miles, C.H. (2025). A Cautionary Tale on Integrating Studies with Disparate Outcome Measures for Causal Inference. NeurIPS.

Funding
PCORI Award ME-2020C3-21145 *(PI: Stuart);* NIMH R01MH126856 *(PI: Stuart)*

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

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 $d < P(S \in \{1, \ldots, K\}|\boldsymbol{X} = \boldsymbol{x}) < 1 - d$ for all $\boldsymbol{x}$ in the target setting

S-Learner

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

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

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 $\mu(\boldsymbol{X_i}, 0) = E(Y_i(0)|\boldsymbol{X_i})$

Aggregation Methods

Meta-Analysis with fixed effects and random effects

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

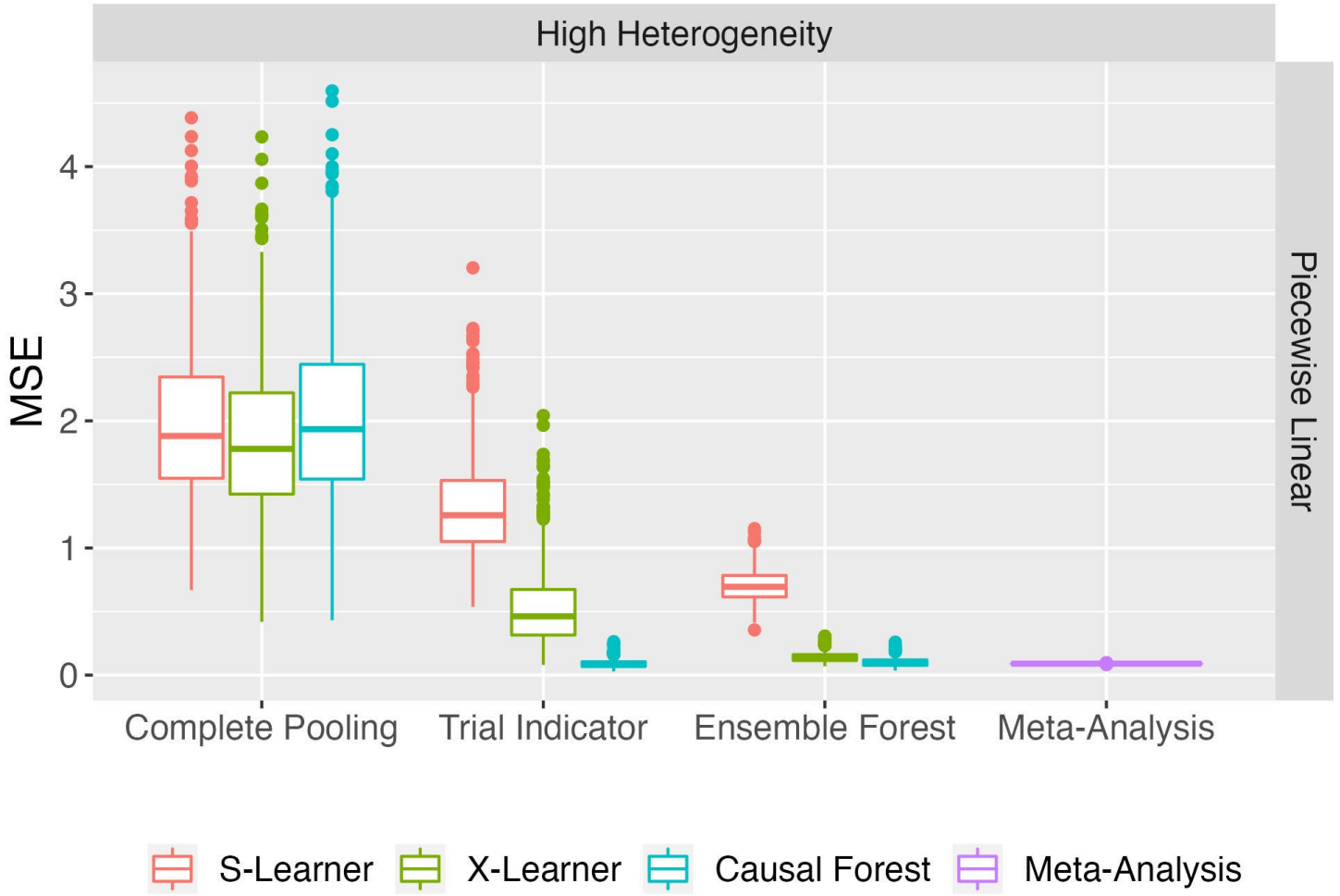
Simulation Setup

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

Parameters varied:

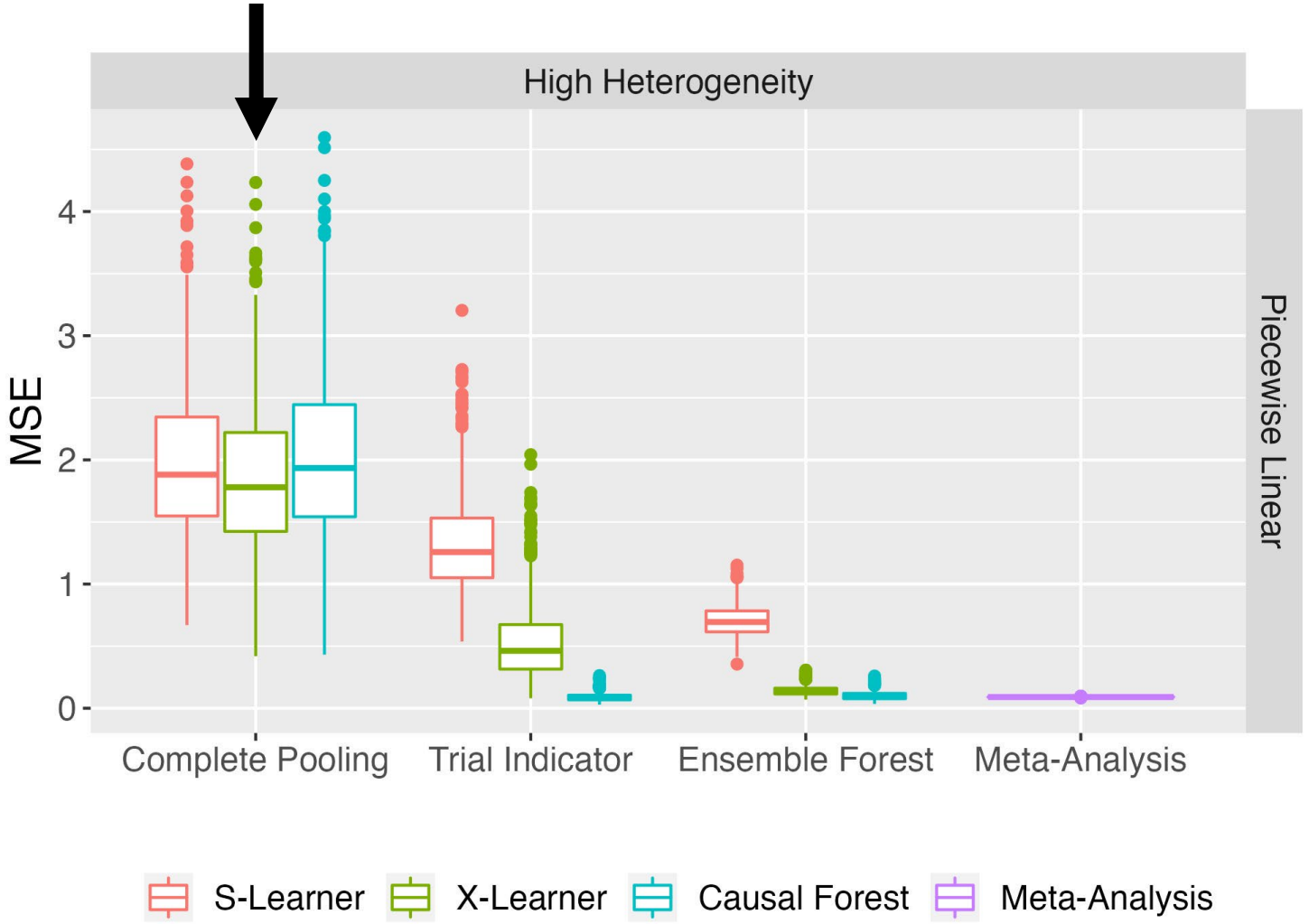
- Heterogeneity of effect across studies (low, medium, **high**)
- Form of CATE (**piecewise linear** or **non-linear**)
- Trial sample sizes (**all 500**, one large, half and half)
- Number of trials (**10** or 30)

Simulation Results



Key Takeaways

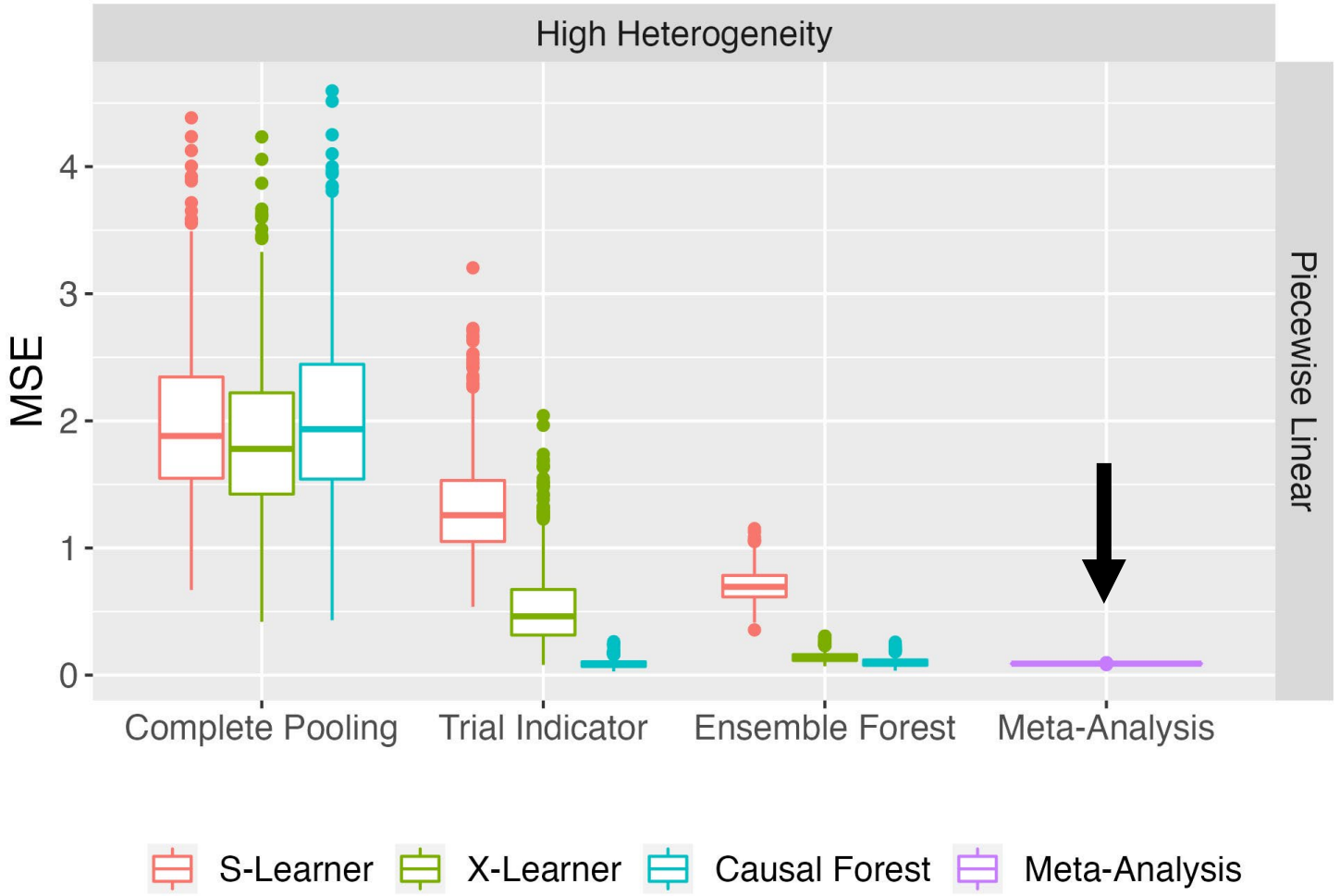
Simulation Results



Key Takeaways

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

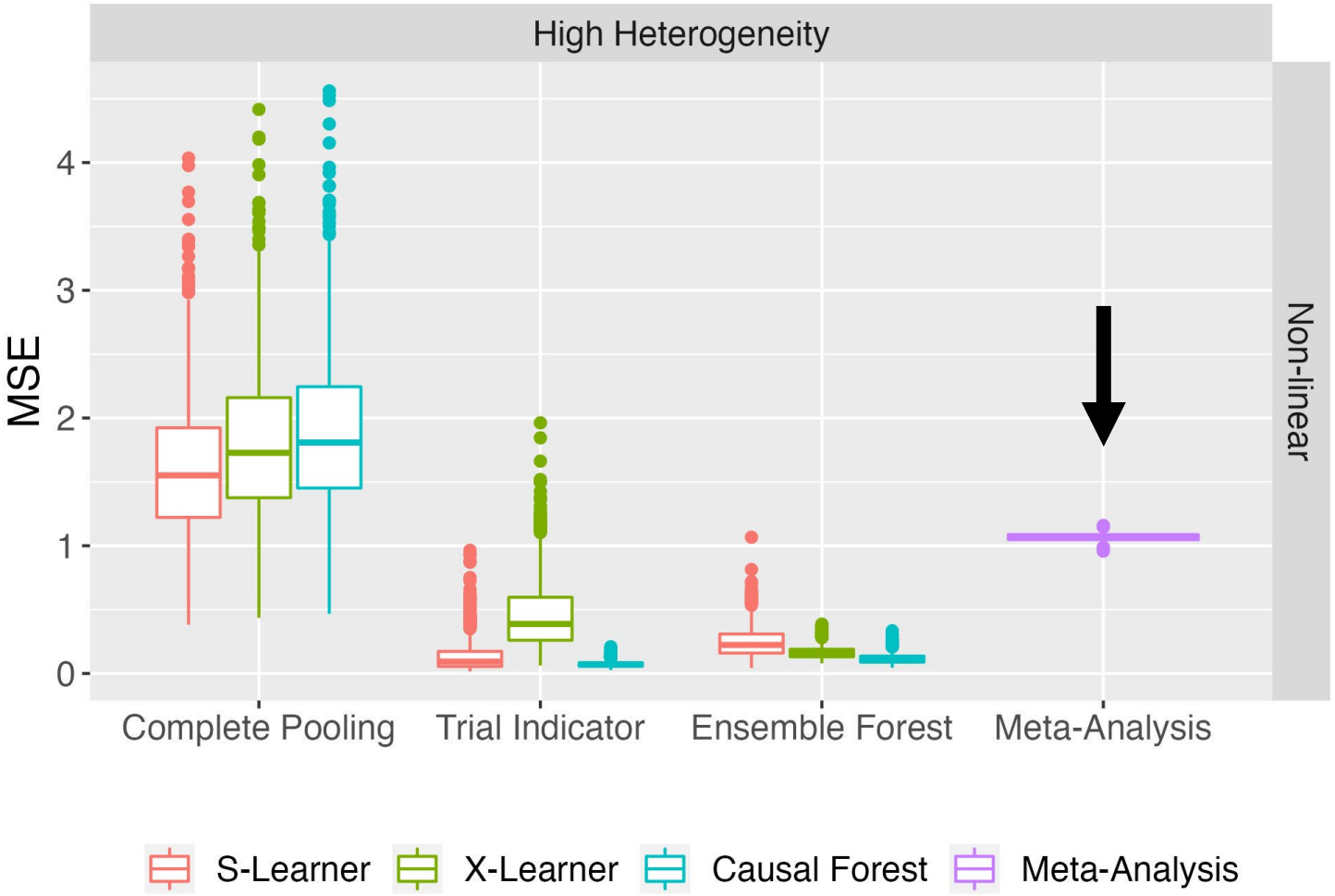
Simulation Results



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)