The best (and worst?) of both worlds? Combining electronic health record and clinical trial data to understand treatment effect heterogeneity

Elizabeth Stuart

Johns Hopkins Bloomberg School of Public Health

estuart@jhu.edu; @lizstuartdc; www.elizabethstuart.org

April 4, 2024

Acknowledgments

- PhD students: **Carly Lupton Brantner**, Leon di Stefano
- Collaborators: Hwanhee Hong (Duke), Trang Nguyen (Hopkins), Peter Zandi (Hopkins)
- Funders: NIMH R01MH126856; PCORI ME-2020C3-21145
- This presentation is based on research using data from data contributors, Takeda and Lundbeck, that have been 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.

Outline

Outline

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 . . .

Individual studies can only help so much . . .

Randomized trials

- Provide unbiased treatment effect estimates
- Can look at subgroup effects
- But generally powered only for overall main effects
- Rule of thumb is that sample size needs to be 4x as large to look at effect heterogeneity even just across 2 subgroups!

Non-experimental studies

- Potentially large size
- May reflect "real world" use
- May have more representative populations
- But may suffer from confounding . . .

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; will signal extensions (and complications) for bringing in non-experimental studies too
    - Machine learning methods allow for flexible identification of moderators, interactions, etc., with no need to prespecify

Outline

Motivating application: Medication for depression

- Question: Are medications for depression differentially effective?
- Comparison of Duloxetine and Vortioxetine for individuals with major depressive disorder
- Combination of randomized trial data and (eventually) electronic health record data

Context

- 30-40% of people with MDD go into remission after depression therapies; 1/3 respond but have residual symptoms
- First line treatment: SSRI (Prozac, Zoloft, etc.)
- **Duloxetine** (Cymbalta): Serotonin-noradrenaline reuptake inhibitor (SNRI) – increase amount of serotonin and noradrenaline in the brain – by Eli Lilly
- **Vortioxetine** (Trintellix): Direct modulation of receptor activity and inhibition of serotonin transporter – by Takeda/Lundbeck
- Common adverse effects for both: nausea, headache, dry mouth, diarrhea
- RCTs generally showed that both Vortioxetine and Duloxetine had significantly more improvement in symptoms than placebo

More on the trial data

- Four RCTs (n = 575, 436, 418, 418) with participants randomly assigned to treatments for major depression: **Duloxetine** and **Vortioxetine**
- **Eligibility criteria:**
  1. 18-75 years old
  2. Had a Major Depressive Episode (MDE) as a primary diagnosis lasting at least three months
  3. Had a Montgomery-Asberg Depression Rating Scale (MADRS) score of at least 22 (one trial) or 26 (three trials) at both screening and baseline
- **Outcome:** Change in MADRS score from baseline to the last observed follow-up
  - Positive CATE implies Duloxetine more effective than Vortioxetine for decreasing MADRS Score

Characteristics of Trial Participants

	NCT 00635219 (N=575)	NCT 00672620 (N=418)	NCT 01140906 (N=436)	NCT 01153009 (N=418)
Age (Mean)	46.3	43.0	46.3	43.4
Female (%)	67.7	64.4	65.4	74.2
Weight in kg (Mean)	70.6	87.3	74.1	87.5
Has Smoked (%)	32.2	27.3	35.8	25.6
Has Anxiety (%)	3.7	1.9	0.2	3.8
BIn MADRS (Mean)	31.9	29.8	31.4	32.3
BIn HAM-A (Mean)	23.0	18.4	20.8	17.9

Outline

Notation

- $A \in \{0, 1\}$ indicates treatment status
- $\boldsymbol{X}$ are covariates (continuous)
- Y is a continuous outcome
  - $Y(1)$ is the potential outcome under treatment
  - $Y(0)$ is the potential outcome under control
- $S \in \{1, ..., K\}$ is a study indicator
- $\pi_s(\boldsymbol{X})$ is the propensity score (probability of treatment given covariates) in study s

Estimand

The estimand is the study-specific conditional average treatment effect:

$$\tau_s(\boldsymbol{X}) = E(Y(1)|\boldsymbol{X}, S = s) - E(Y(0)|\boldsymbol{X}, S = s)$$

Assumptions

1. **Stable Unit Treatment Value Assumption** (SUTVA) in each study
2. **Unconfoundedness of each RCT:** $\{Y(0), Y(1)\} \perp\!\!\!\perp A|\boldsymbol{X}$ in each study (this satisfied if actually randomized)
3. **Consistency:** $Y = AY(1) + (1 - A)Y(0)$ almost surely in each study
4. **Positivity of treatment assignment:** There exists a constant $c > 0$ such that $c < \pi(\boldsymbol{X}) < 1 - c$ for all $\boldsymbol{X}$ in each study
5. **Positivity of study membership** (*Can be relaxed*): There exists a constant $d > 0$ such that $d < P(S = s|\boldsymbol{X}) < 1 - d$ for all $\boldsymbol{X}$ and s

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 Approaches
     1. Ensemble Tree
     2. Ensemble Forest
     3. Ensemble Lasso
  4. Meta-Analysis

Single-Study Methods

- Several non-parametric approaches exist for estimating the CATE in a single study; we selected three
- Two of the options focus on estimating the conditional mean outcomes first ($\mu(\boldsymbol{X}, A) = E(Y|\boldsymbol{X}, A)$) and then using the difference between those to estimate the CATE
  - S-Learner: estimates model of outcome as a function of covariates and treatment status
  - X-Learner: estimates separate models of outcomes under treatment and under control
- The third option is a forest-based algorithm that partitions the covariates directly based on treatment effect heterogeneity
  - Causal Forest

Aggregation Methods: One-Step Approaches

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

- **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: Three-Step Approaches

Extending Federated Learning Method

1. Build localized models for CATE within each study
2. Apply each of these localized models to each individual across ALL studies to estimate the CATE
   - Ex: For K studies with a total of N individuals in all studies combined, there will be K study-specific CATE models. Then each of these models will be applied to all data points, so every individual will have K different estimates of their CATE. So we will end up with $N * K$ CATE estimates in an "augmented" dataset.
3. Fit model **(tree, forest, lasso)** on the augmented data, where the estimated treatment effect is the outcome, and patient features and study are covariates

Aggregation Methods: Meta-Analysis (One-Step)

Parametric comparison method: **meta-analysis with study random effects**

$$Y = (\alpha_0 + a_s) + \boldsymbol{\alpha}^T \boldsymbol{X} + b_s X_1 + (\zeta + z_s)A + (\theta + t_s)X_1 A + \epsilon.$$

- Fixed components are: α_0, $\boldsymbol{\alpha}$, ζ, and θ
- Random components are: $a_s \sim N(0, \sigma_a^2)$, $b_s \sim N(0, \sigma_b^2)$, $z_s \sim N(0, \sigma_z^2)$, and $t_s \sim N(0, \sigma_t^2)$
- Residual error is: $\epsilon \sim N(0, \sigma^2)$

The CATE is $\tau_s(\boldsymbol{X}) = (\zeta + z_s) + (\theta + t_s)X_1$.

Outline

Data Generation

Simulation setup

- Studies: $K = 10$ with $n_k = 500$ for all k
- $\boldsymbol{X}_i \sim N(0, I_5)$
- $P(A_i = 1) = 0.5$
- Main effect term: $\beta_s \sim N(0, \sigma^2_\beta)$ and interaction effect term: $\delta_s \sim N(0, \sigma^2_\delta)$
  - $(\sigma_\beta, \sigma_\delta) \in \{(0.5, 0), (1, 0), (1, 0.5), (1, 1), (3, 1)\}$
- Scenario: Piecewise linear CATE, non-linear CATE, or variable CATE

Number of Setups

1,000 iterations of 11 parameter combinations

Simulation Results

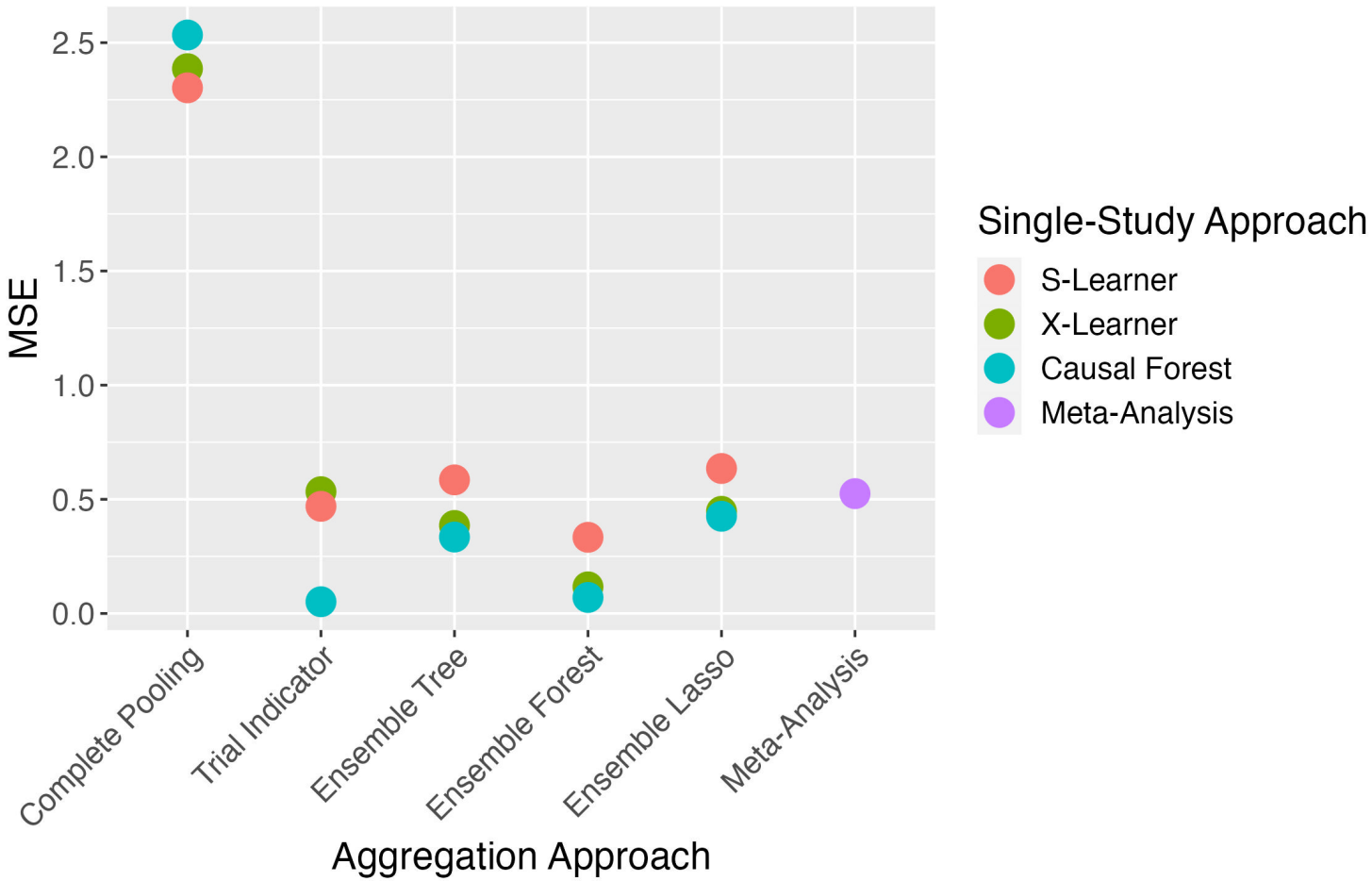

Figure: Average MSE Across All Scenarios and Iterations

Simulation Results continued

Key takeaways:

- As variability of study coefficients increases, MSE increases - this happens much more rapidly for the Complete Pooling approaches and Meta-Analysis
- The Ensemble Lasso performs well for the piecewise linear CATE but poorly for the non-linear CATE
- The S-Learner performs poorly for the piecewise linear CATE and well for the non-linear CATE
- The most consistently effective single-study approach is the Causal Forest, and the most consistently effective aggregation approaches are Pooling with Trial Indicator and the Ensemble Forest

Outline

Application to Depression Treatments

- Applied the methods to the depression treatment data
- Used the causal forest with pooling with trial indicator approach

CATE Estimates

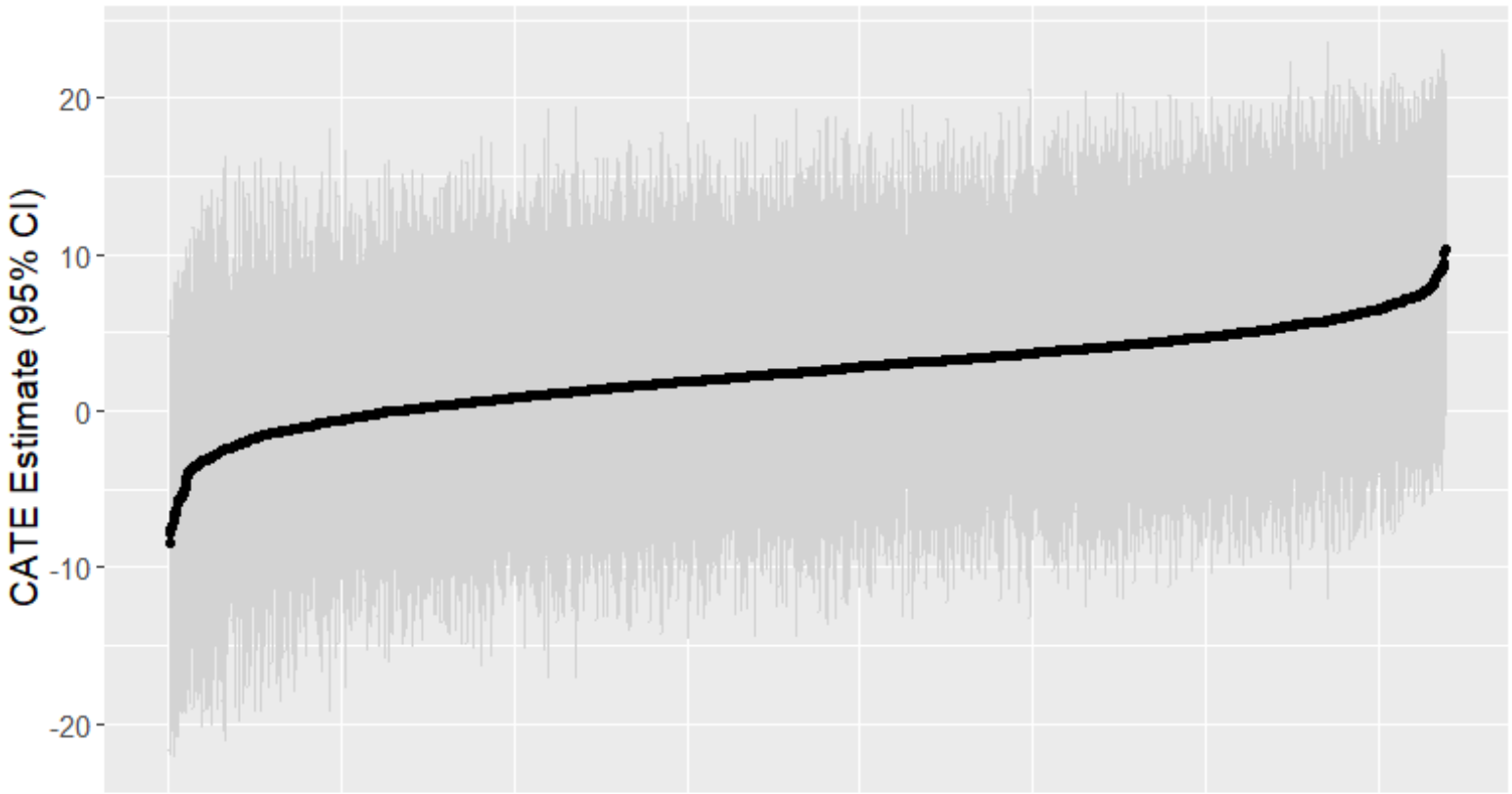

Figure: Distribution of CATEs According to Causal Forest with Pooling with Trial Indicator

Interpretation Tree

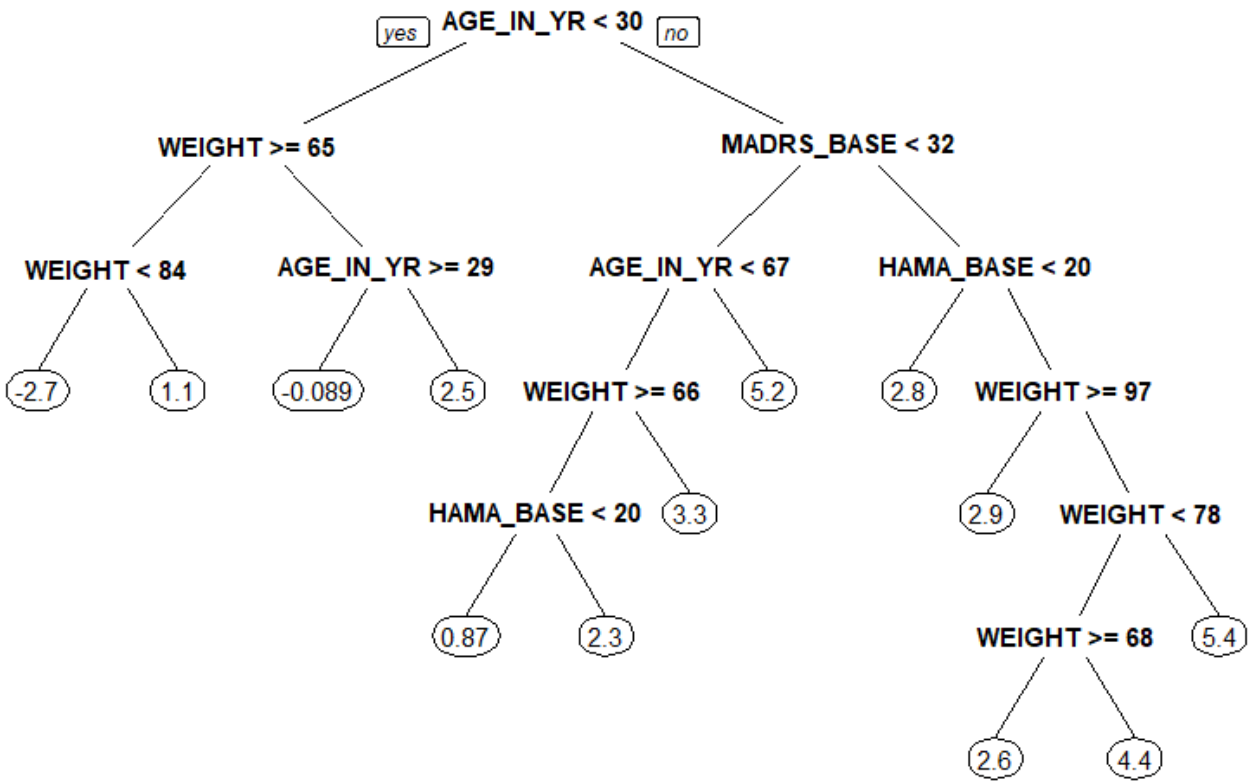

Figure: Interpretation tree for Causal Forest with Pooling with Trial Indicator

CATE by Age

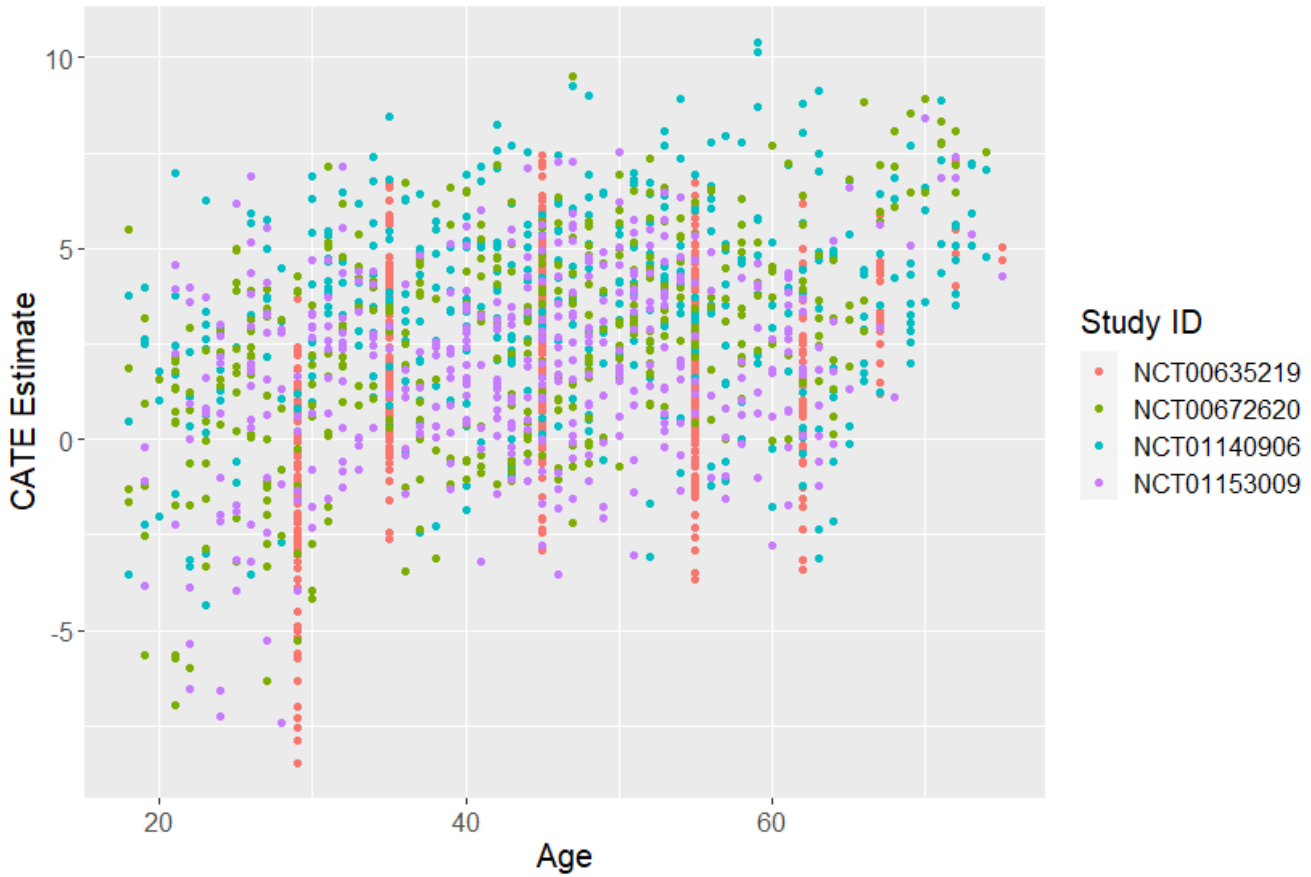

Figure: Scatterplot of CATE Estimate According to Causal Forest with Pooling with Trial Indicator by Age and Trial

More Results from MDD Data

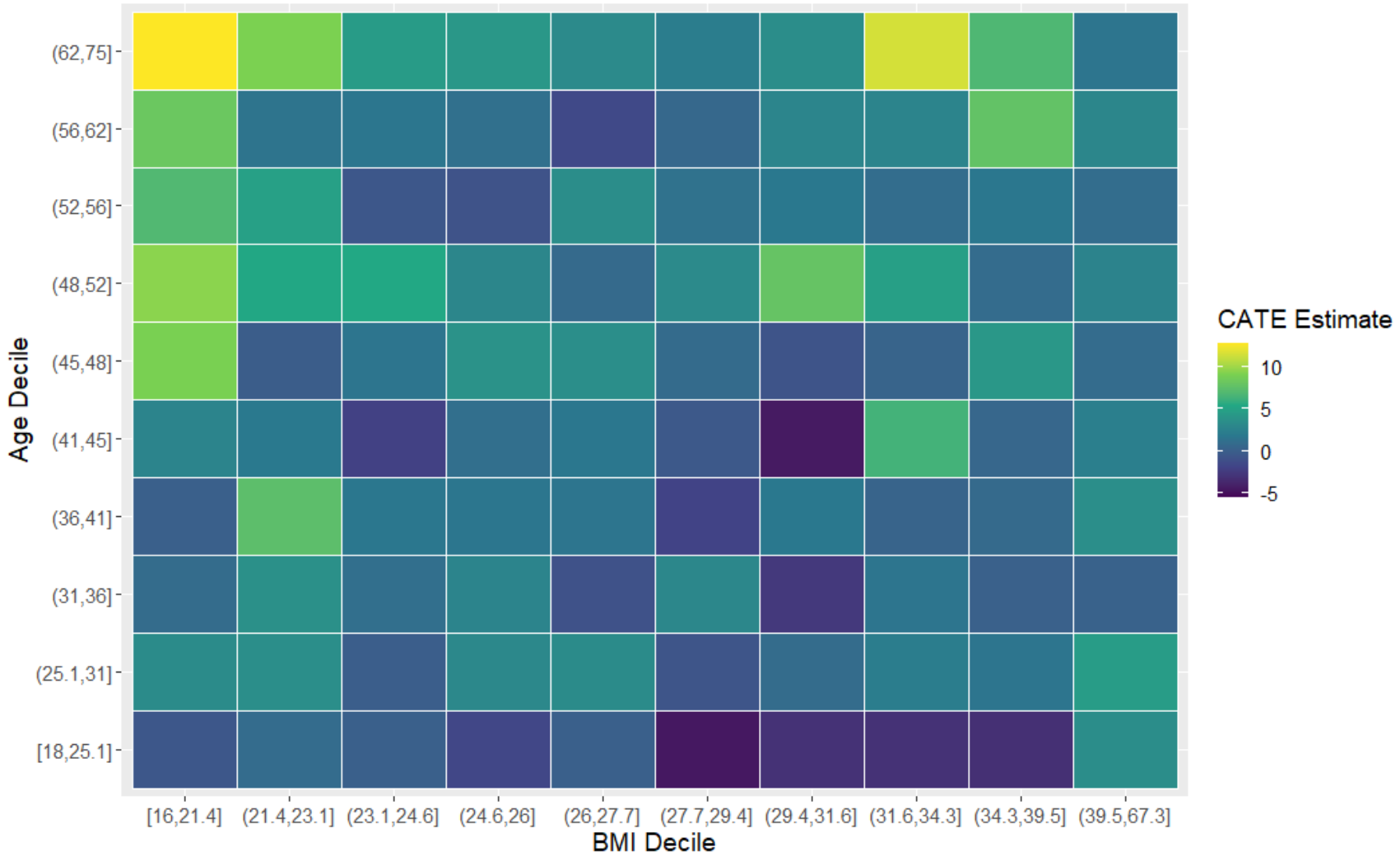

Figure: CATE Estimate According to Causal Forest with Pooling with Trial Indicator by Decile of BMI and Age

Outline

Open questions to have this 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.
- Sto still a work in progress...stay tuned!

Conclusions

- Pooling with Trial Indicator and Ensemble Forests had consistently low mean squared error in all scenarios
  - Especially with the Causal Forest
- Parametric linear approaches struggled with complex CATE functions
- Choice of single-study approach matters and more diagnostics for making this decision will be useful
- Limitations
  - Did not include an exhaustive list of potential approaches or simulation setups
  - Most of the resulting CATE estimates are trial-specific
  - MDD trials were not comparing Duloxetine and Vortioxetine but instead each medication with placebo
  - Lots more work to do for use!

References

- Athey, S., Tibshirani, J., Wager, S. (2019). Generalized random forests. *The Annals of Statistics, 47*(2), 1148-1178.
- Lupton Brantner, C., Chang, T-H., Nguyen, T. Q., Hong, H., Di Stefano, L., Stuart, E. A. (2023). Methods for integrating trials and non-experimental data to investigate treatment effect heterogeneity. Forthcoming in *Statistical Science*. https://arxiv.org/abs/2302.13428.
- Künzel, S. R., Sekhon, J. S., Bickel, P. J., Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences, 116*(10), 4156-4165.
- Brantner, C.L., Nguyen, T.Q., Tang, T., Zhao, C., Hong, H., Stuart, E.A. (2024). Comparison of methods that combine multiple randomized trials to estimate heterogeneous treatment effects. *Statistics in Medicine.* https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9955
- 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. Forthcoming in *Statistical Science.*
- Tan, X., Chang, C. C. H., Tang, L. (2021). A tree-based federated learning approach for personalized treatment effect estimation from heterogeneous data sources. *arXiv preprint arXiv:2103.06261*.

Single-Study Methods: S-Learner

S-Learner

1. Estimate combined function with treatment indicator included: $\mu(\boldsymbol{X}, A) = E(Y|\boldsymbol{X}, A)$ using a random forest
2. Directly calculate the CATE using $\hat{\tau}(\boldsymbol{X}) = \hat{\mu}(\boldsymbol{X}, 1) - \hat{\mu}(\boldsymbol{X}, 0)$

Single-Study Methods: X-Learner

X-Learner

1. Estimate $\mu(\boldsymbol{X}, 1) = E(Y(1)|\boldsymbol{X})$ and $\mu(\boldsymbol{X}, 0) = E(Y(0)|\boldsymbol{X})$ separately using random forests
2. Estimate treatment effects for individuals in each group using the true data and the estimated outcome functions:

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

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

   Regress with $\tilde{D}_i$'s as outcome to get $\hat{\tau}_1(\boldsymbol{X})$ and $\hat{\tau}_0(\boldsymbol{X})$
3. Define CATE ($\hat{\tau}$) as the weighted average of $\hat{\tau}_1$ and $\hat{\tau}_0$:

$$\hat{\tau}(\boldsymbol{X}) = g(\boldsymbol{X})\hat{\tau}_0(\boldsymbol{X}) + (1 - g(\boldsymbol{X}))\hat{\tau}_1(\boldsymbol{X})$$

Single-Study Methods: 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