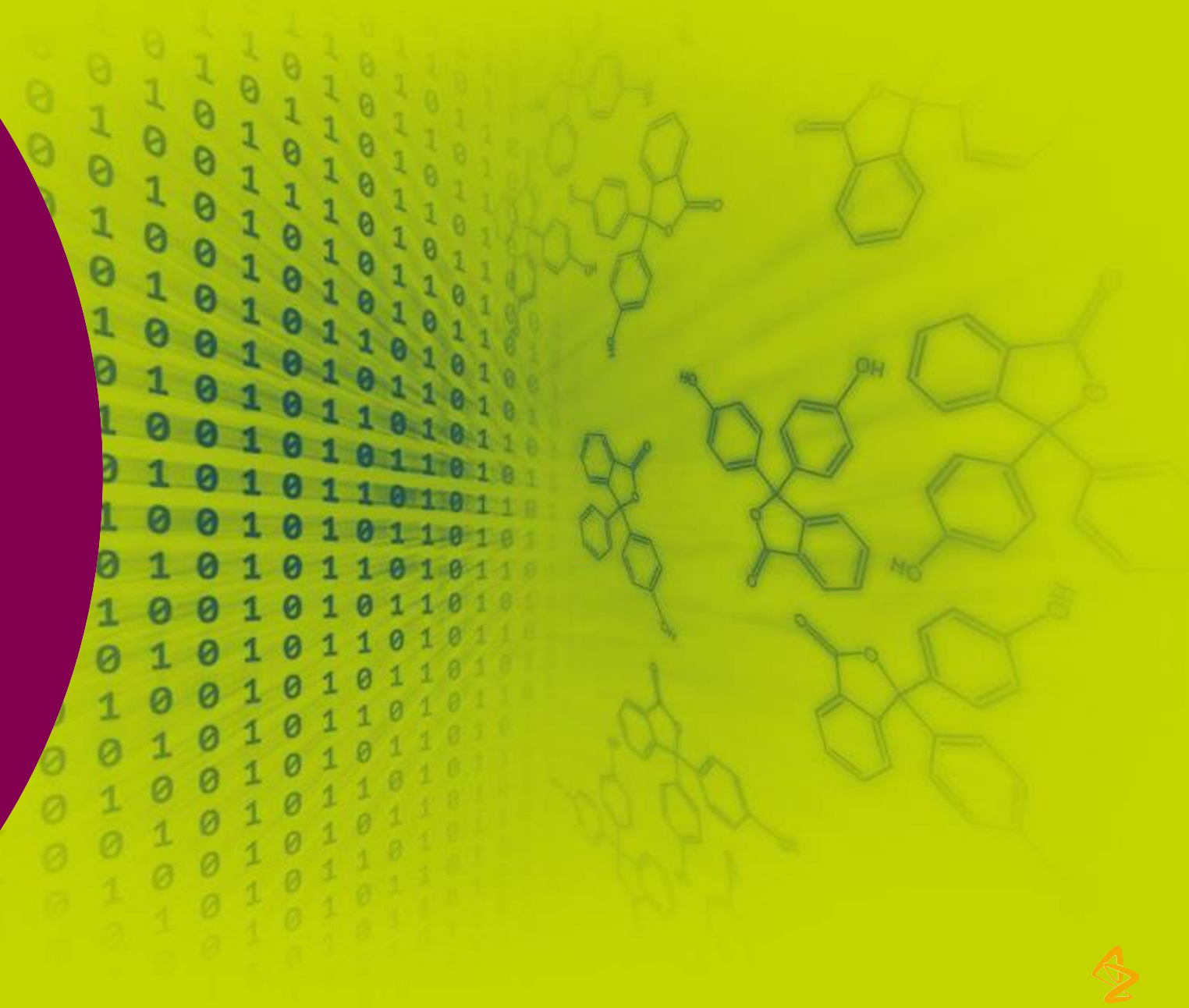
PSI Amsterdam 2024

Comparison of modern approaches for subgroup identification from clinical and observational data

David Svensson Statistical Innovation, AstraZeneca R&D

Joint work with Ilya Lipkovich[2], Bohdana Ratitch[3], Alex Dmitrienko[4]

(2) Eli Lilly and Company (3) Bayer (4) Mediana



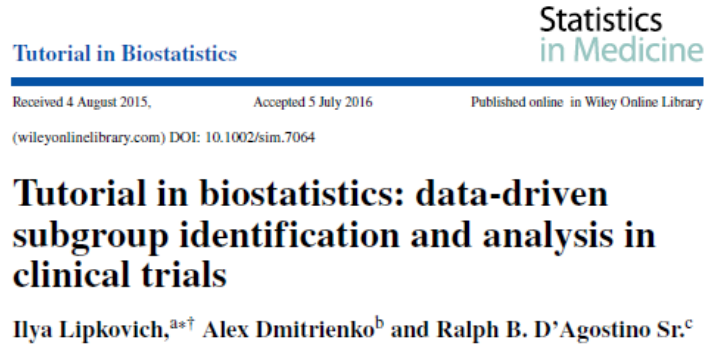
Content

Lead of cross-industry EFSPI Subgroup Special Interest Group (SIG) since 2018.

The SIG activities generally has resulted in various collaborative research efforts.

E.g., Recently, two papers by I. Lipkovich, B. Ratitch, A. Dmitrienko & D. Svensson

- Updates of a seminal subgroup detection paper [1] 2016:



- New papers 2023/2024: [2],[3] overview of developments since that time (+ some benchmarking)

https://arxiv.org/abs/2311.14889

Today: some shapshots of this work, (selected aspects only)

Some keywords: Causal Inference, Machine Learning [ML], Individual Treatment Effects

A short detour: AI

Strong and broad focus on **AI** across many domains. **Expect proposals to use it everywhere**! What about Subgroup Detection // Individual Treatment Effects?

- *Can a machine learn automatically who responds better to active treatment (by examples)?*

AI (often vast neural nets) excels when

(1) *data is cheap,* [chess! Images! Text on the web! ...]

(2) the *ground truth available in training data* [cats!? won/lost games?, words, ...]

WE DON'T HAVE THIS in RCT/RWE data, especially not (2) for fundamental reasons.

- But can we **'almost do AI'** for finding novel subgroups? Let's have a look...

Rubin's Potential Outcome framework:

Each patient has two Potential Outcomes of Y, i.e., $Y^{(0)}$ and $Y^{(1)}$ corresponding to Trt=0, 1

- Only one of them is observed in a trial (parallel design)
- I.e., ITE = $Y^{(1)} - Y^{(0)}$ is *fundamentally* unobservable ("no ground truth in the training data")

  - patient gets either active or control!

Target becomes $\Delta(\mathbf{x}) := \mathbb{E}[Y^{(1)} - Y^{(0)} \mid \mathbf{X}=\mathbf{x}]$, where $\mathbf{x}=(x_1,\ldots,x_p)$ is baseline biomarkers.

This is **CATE** (**C**onditional **A**verage **T**reatment **E**ffect), ... target in many recent papers ...

- Assumptions required for RWE data (when propensity scores often enters)

Stressing

CATE:

... for a patient represented by these covariates

$$\Delta(\mathbf{x}) := \mathbb{E}[Y^{(1)} - Y^{(0)} \mid \mathbf{X}=\mathbf{x}]$$ as a (multivariate) function of $\mathbf{x}=(x_1,\ldots,x_p)$

Expected (individual) trt. Effect ...

Representing an agnostic look at the data "AI style" (Let The Data Speak)

- Do (at least) some types of patients benefit? If so, can we figure out what is typical about them?

From CATE estimates to Subgroup: $\hat{S}=\{\hat{\Delta}(\mathbf{x})>0\}$ (='the patients benefitting more from active treatment').

Interestingly, **other industries** look at such problems [7] .... (based on Machine Learning).

- 'Who is more likely to respond to a personalized ad, new policy in society, etc"

We benchmarked some approaches …

CATE Estimator	ML type/Base Learner	Outcome model?
T-Learning	XGboost	Yes
S-Learning	Xgboost	Yes
X-Learning	Xgboost	Yes
R-Learning	Xgboost	Hybrid
Causal Forest	Causal trees	No
Bayesian Forest	BART	No
A-Learning	Xgboost	No
A-Learning Augmented	Xgboost	Hybrid
W-Learning	Xgboost	No
W-Learning Augmented	Xgboost	Hybrid

Our tutorial paper also covered many other aspects (but excluded here).

E.g., ITR, post-selection subgroup inference, global tests, case studies, interconnection between the methodologies, ...

Modelling School no. 1: '*Indirect approach*'

CATE Estimator
T-Learning
S-Learning
X-Learning
R-Learning
Causal Forest
Bayesian Forest
A-Learning
A-Learning Augme
W-Learning
W-Learning Augme

"Predictions first, in 'data science style'"

$\Delta(\mathbf{x}) := \mathbb{E}[Y^{(1)} - Y^{(0)} | \mathbf{X}=\mathbf{x}] =$

$= \mathbb{E}[Y^{(1)} | \mathbf{X}=\mathbf{x}] - \mathbb{E}[Y^{(0)} | \mathbf{X}=\mathbf{x}]$ (for trivial reason)

$= \mathbb{E}[Y | \mathbf{X}=\mathbf{x}, Trt=1] - \mathbb{E}[Y | \mathbf{X}=\mathbf{x}, Trt=0]$ (standard assumptions)

E.g., $\hat{\Delta}(\mathbf{x}) = \hat{m}_1(\boldsymbol{x}) - \hat{m}_0(\boldsymbol{x})$ two regression models

I.e., first outcome modelling (using off-the-shelf ML), only then derive CATE

Modelling School no. 2: '*Direct approach*'

CATE Estimator		
T-Learning		
S-Learning		
X-Learning		
R-Learning		
Causal Forest		
Bayesian Forest		
A-Learning	Xgboost	
A-Learning Augmented	Xgboost	
W-Learning	Xgboost	No
W-Learning Augmented	Xgboost	Hybrid

"N*ot interested in predicting Y, just give us the contrasts*"

Set up a suitable loss function **L** expressed in terms of Y, Trt and **x** and a candidate **f**(**x**)

$\hat{\boldsymbol{f}} = argmin_{\{f \in C\}}(\boldsymbol{L}(Y, Trt, \boldsymbol{x}; f))$ renders **Δ(x)**

$\hat{f}$ can be constructed using off-the-shelf ML

$$L_A(f) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{M}(Y_i, \{(T_i+1)/2 - \pi(\mathbf{x}_i)\} \times f(\mathbf{x}_i))$$

$$L_W(f) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{M}(Y_i, T_i \times f(\mathbf{x}_i))}{T_i\pi(\mathbf{x}_i) + (1-T_i)/2}$$

(Formulas above borrowed from Huling [4].)

Modelling School no. 2b: '*Direct approach but [...]*'

CATE Estimator	
T-Learning	
S-Learning	
X-Learning	
R-Learning	
Causal Forest	
Bayesian Forest	
A-Learning	
A-Learning Augmented	
W-Learning	Xgboost
W-Learning Augmented	Xgboost

"oops high variance, let's help it a bit...":

$\hat{f} = argmin_{\{f \in C\}}(\boldsymbol{L}(Y, Trt, \boldsymbol{x}; f))$ renders **Δ(x)** but now ***Hybrid***

(=sneaking in **outcome** modelling as a **nuisance** parameter, cross-fitting, etc)

Still using off-the-shelf ML.

Nie Wager 2020 [4]: '*decomposition, **R1**esiduals on **R2**esiduals*'

$$\hat{\Delta}(\mathbf{x}) = argmin_{\{f \in C\}} \left(\frac{1}{n} \Sigma_i \overbrace{\left(Y_i - \hat{m}^{\{-i\}}\right)}^{R1} - \overbrace{\left(T_i - \hat{\pi}^{\{-i\}}\right)}^{R2} f(\boldsymbol{x_i}) \right)^2$$

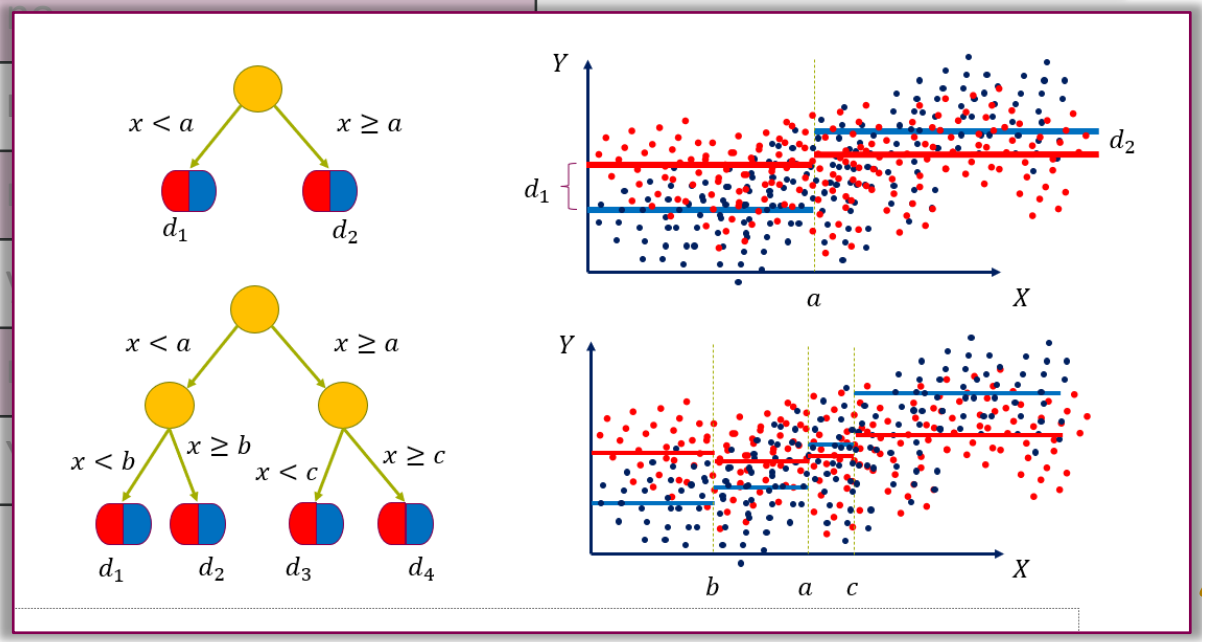
R1= Residuals: (Outcome – Outcome.model) ("$\hat{m}^{\{-i\}}$" = cross-fitted prognostic model) R2= Residuals: (Treatment – Treatment.propensity.model) ("$\hat{\pi}^{\{-i\}}$" = cross-fitted prop.scores)

Modelling School no. 3: '*Tailormade for CATE*'

Modified versions of standard machine learning (e.g., such as RandomForest) to targeting **Δ(x)** instead of Y (i.e., no off-the-shelf ML)

E.g., **Causal Forest** = popular approach,
Biomarker splits trying to capture differential effects...

CATE Estimator	
T-Learning	
S-Learning	
X-Learning	
R-Learning	
Causal Forest	Causal trees
Bayesian Forest	BART
A-Learning	Xgboost
A-Learning Augmented	Xgboost
W-Learning	Xgboost



Sometimes stated 'honest' (unbiased) due to separation of data (for biomarker splits, estimation).

Graph shared by Stefan Franzén (AstraZeneca)
Treatment effect heterogeneity – a practical example [PSI2022]

Simulation Landscape: S1-S4 ("making it difficult")

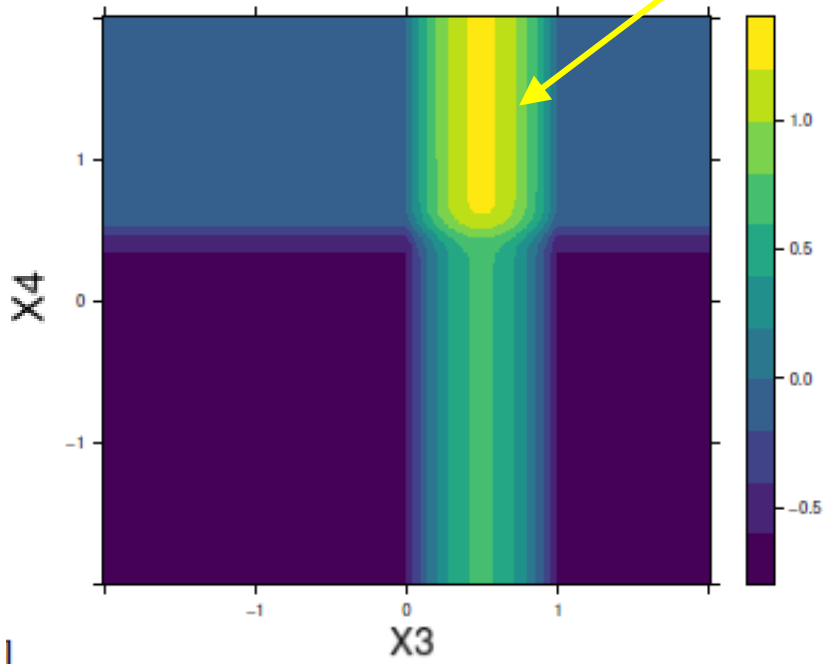
Simulation	No. Prognostic x	Trial Type	TRT assignment	Predictive x
S1	few	**RCT**	3:1 rand (more active)	x3, x4
S2	**many**	**RCT**	3:1 rand (more active)	x3, x4
S3	**many**	Observational	Prognostic assignment (≈1:3)	x3, x4
S4	**many**	Observational	Predictive assignment (≈1:3)	x3, x4

S3: mimicking a physician who assigns patients to Active if their SOC prognosis is poor, i.e., true propensities are driven by the prognostic part of the model for Y , and for S4 the predictive part drivs.

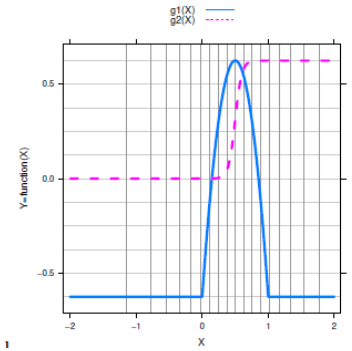
Y=continuous. 19 candidate baseline x

True Treatment Effects?

- non-linear, non-monotone



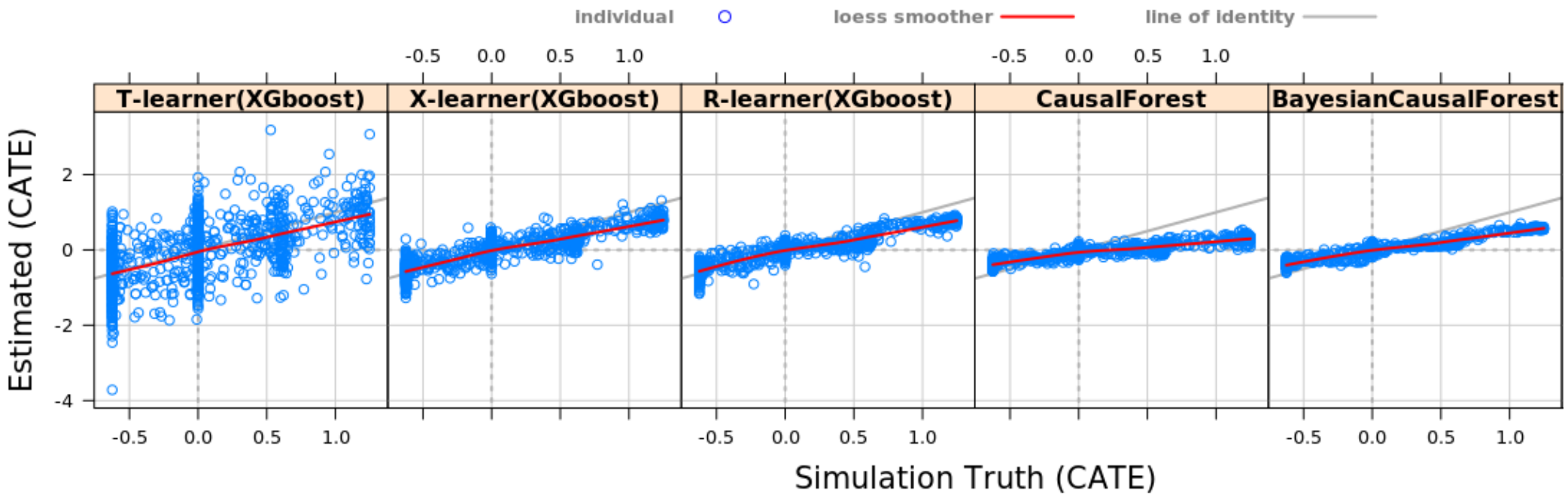
TRUE CATE
(Individual Trt. Effect)
depends on x_3 and x_4,
the ligher color=higher effect



True **S={Δ(x)>0}** has size 0.33,
True average CATE in S is 0.665.
Overall true effect = 0.0119

Illustration One Iteration (S2): Bias-Variance Trade-off

Very different performance noted across methods, e.g., watch this

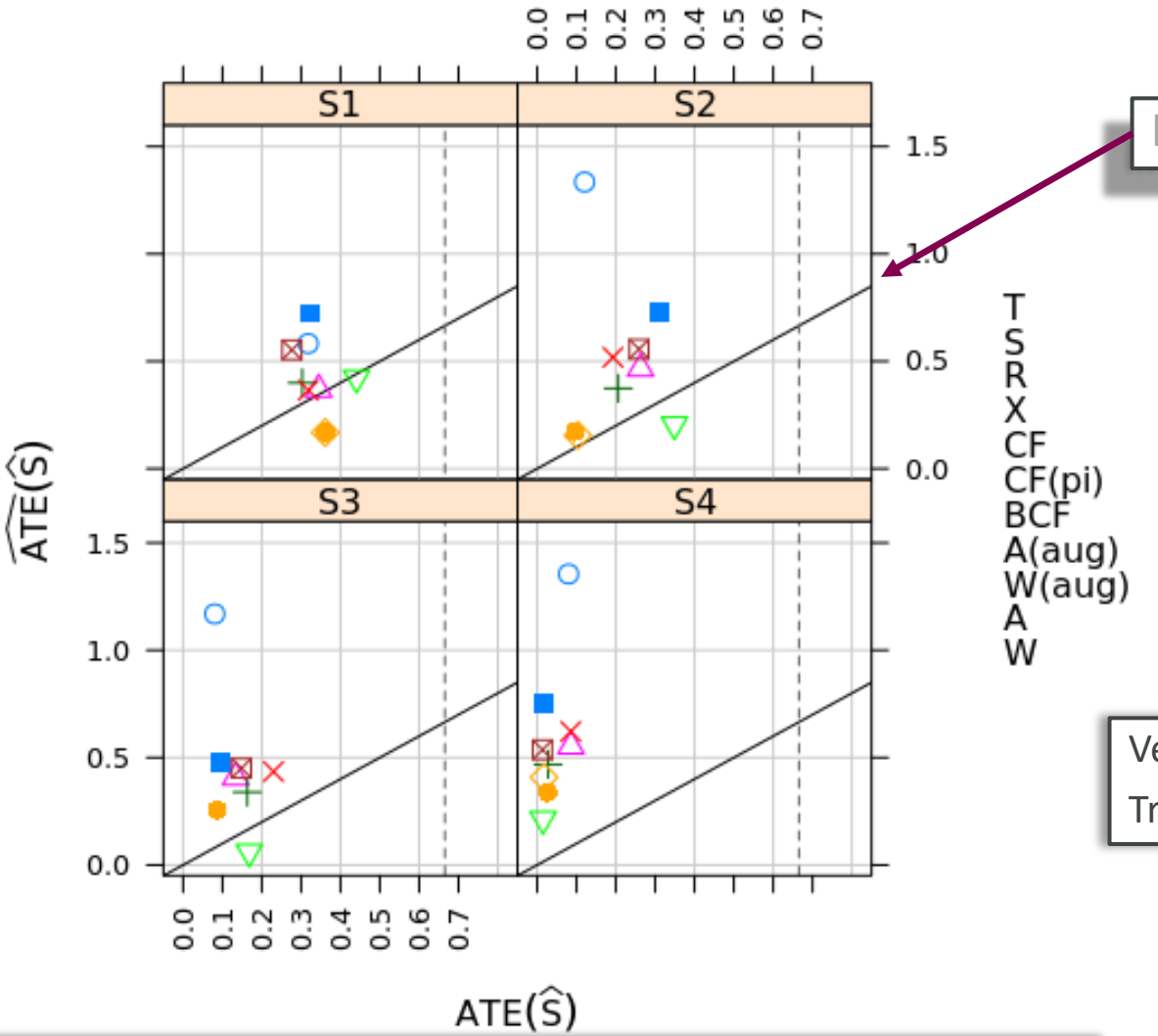


NOTE T-Learning=> Low Bias High Variance, Note Causal Forest High Bias (=hard shrinkage), Low Variance

Benchmarking re. subgroup claim $\hat{S}=\{\hat{\Delta}(x)>0\}$

Each point= averaged performance over 100 simulation iterations

Y axis:
Claimed Trt.Effect
in claimed subgroup



line of identity (Unbiased)

Vertical dashed line =
True Eff in true subgroup S={CATE>0}

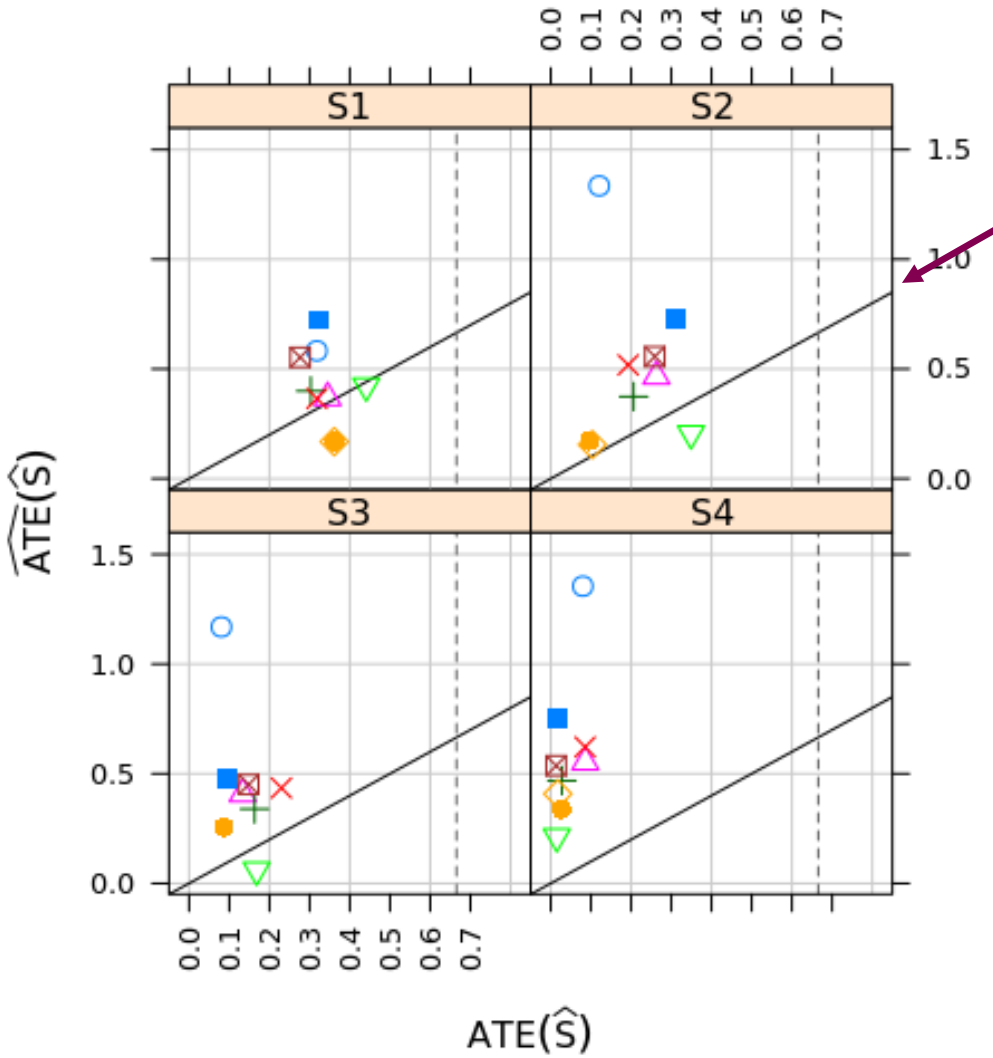
X axis: Actual Trt.Effect in claimed subgroup

Selected results:

Scenario	Method	$corr(\Delta, \hat{\Delta})$	$agree(S, \hat{S})$	$\widehat{ATE}(\hat{S})$	$ATE(\hat{S})$	$SE\{\widehat{ATE}(\hat{S})\}$	$bias\{ATE(\hat{S})\}$	η
S1	T	0.67	0.50	0.58	0.32	0.062	0.26	0.162
S1	S	0.73	0.53	0.37	0.34	0.063	0.02	0.176
S1	R	0.70	0.49	0.40	0.30	0.081	0.10	0.161

Mostly optimistic results, i.e.,

Claim>Actual



line of identity (Unbiased)

E.g., T-learning (Biased)

T
S
R
X
CF
CF(pi)
BCF
A(aug)
W(aug)
A
W

Note Causal Forest (version for observational-data)

Was off-the-chart in S3

Truncated axes:

A and W missing in action (far out ..)

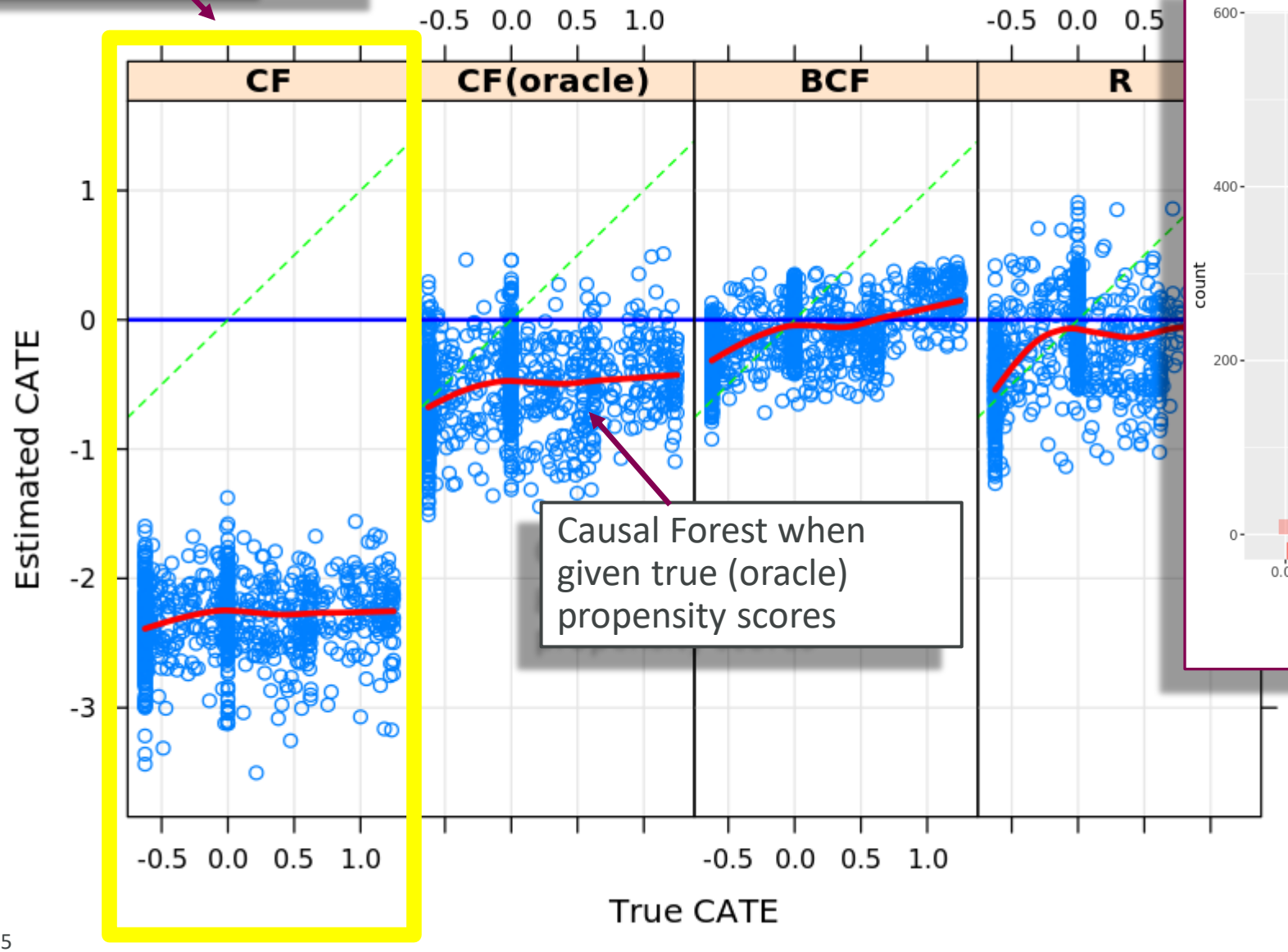
No clear winner?

FIGURE 10 *Average treatment effect (ATE) in identified subgroups by different methods across 4 scenarios (100 simulations); Y-axis displays the estimated ATE in the estimated subgroup $\hat{S}(X) = \{\hat{\Delta}(X) > 0\}$ vs the true treatment effect in $\hat{S}$ (X-axis). Vertical dotted line marks expected average effect in the true subgroup $S(X) = \{\Delta(X) > 0\}$. Notably CF gave quite spurious results in scenario S_3 with every estimate below zero by a margin (hence it is off chart). The plot is truncated and does not display grossly outlying results for the none-augmented A-learning and W-learning across all scenarios*

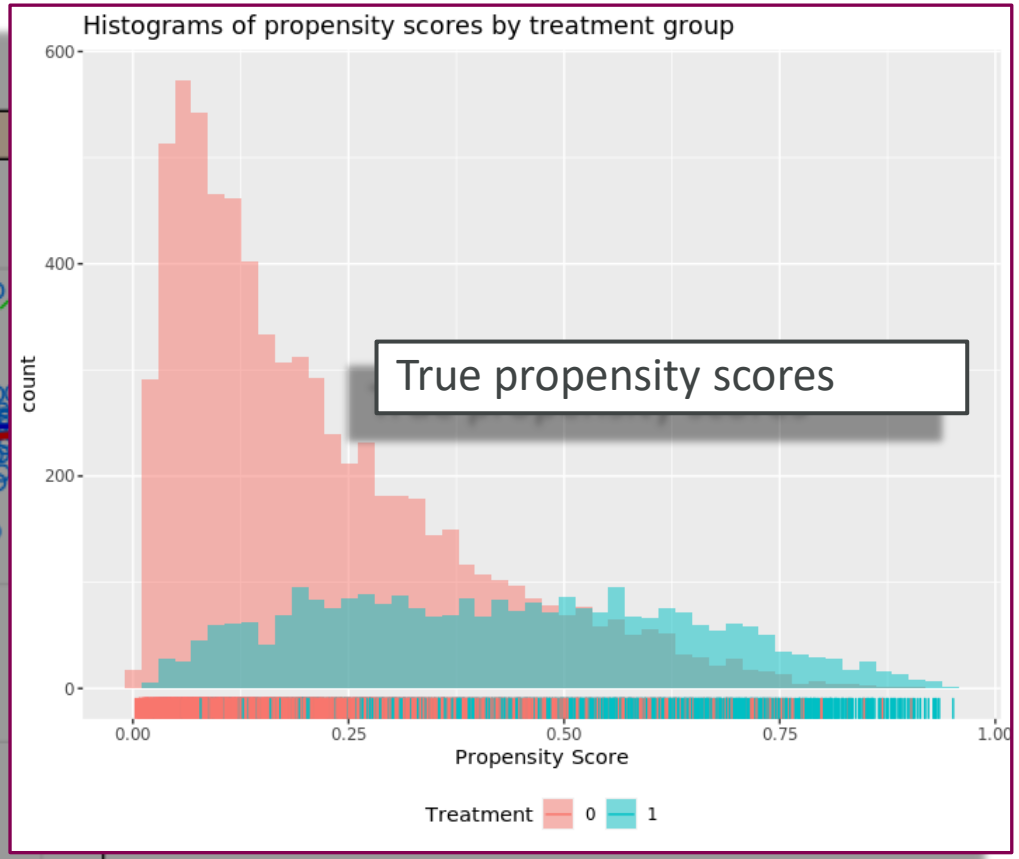
Default

Causal Forest

This page: one iteration **S3** scenario (Observational Data).



Causal Forest when given true (oracle) propensity scores

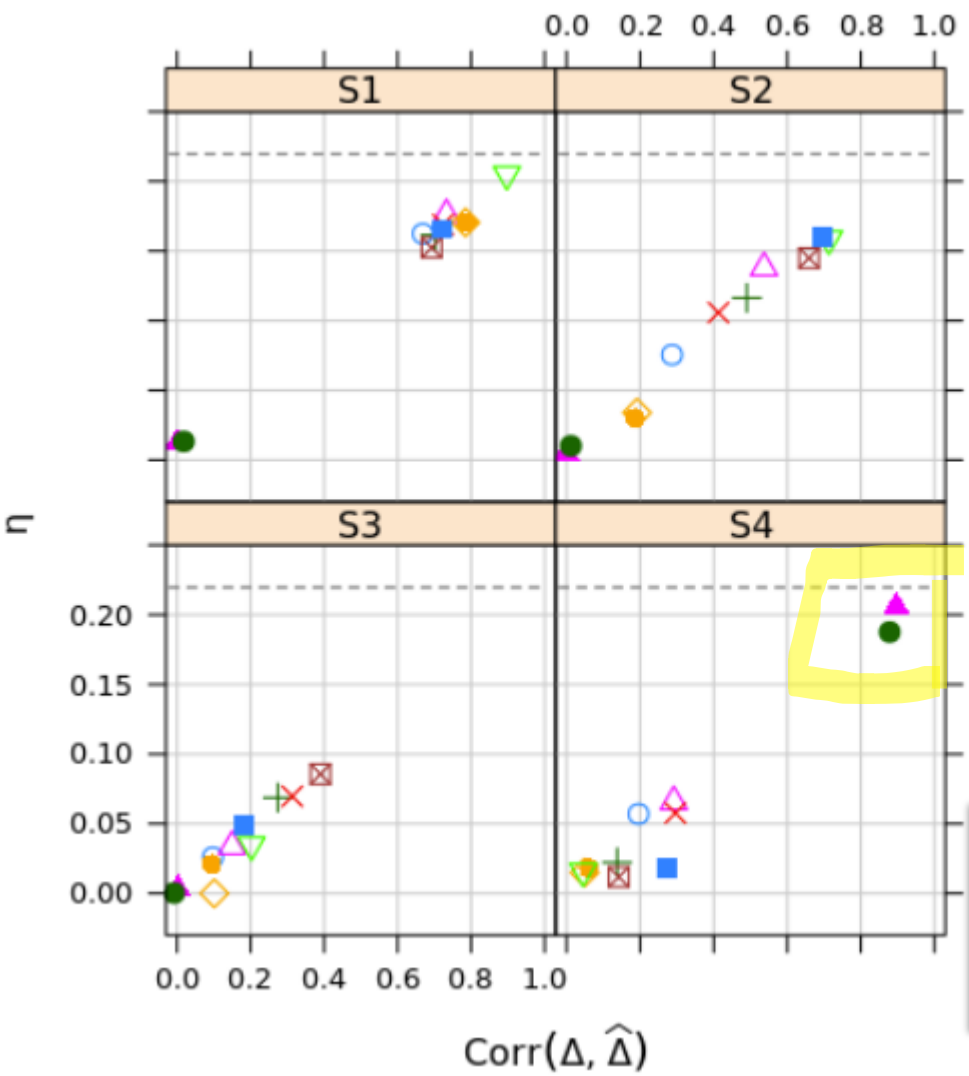


True propensity scores

More…

FIGURE 9 *Benchmarking over 100 iterations for each scenario; plotting the subgroup utility index (η metric) against the Pearson correlation between estimated and true CATE. Methods producing high values on both metrics (which are highly related) indicate good ability to recover underlying CATE as well as the subgroup of patients truly benefitting from the active treatment. The horizontal dotted line indicates the theoretically largest attainable value of the metric η = 0.22.*

Each point= averaged over 100 iterations



Watch **A** and **W**:
Strong in S4,
otherwise weak

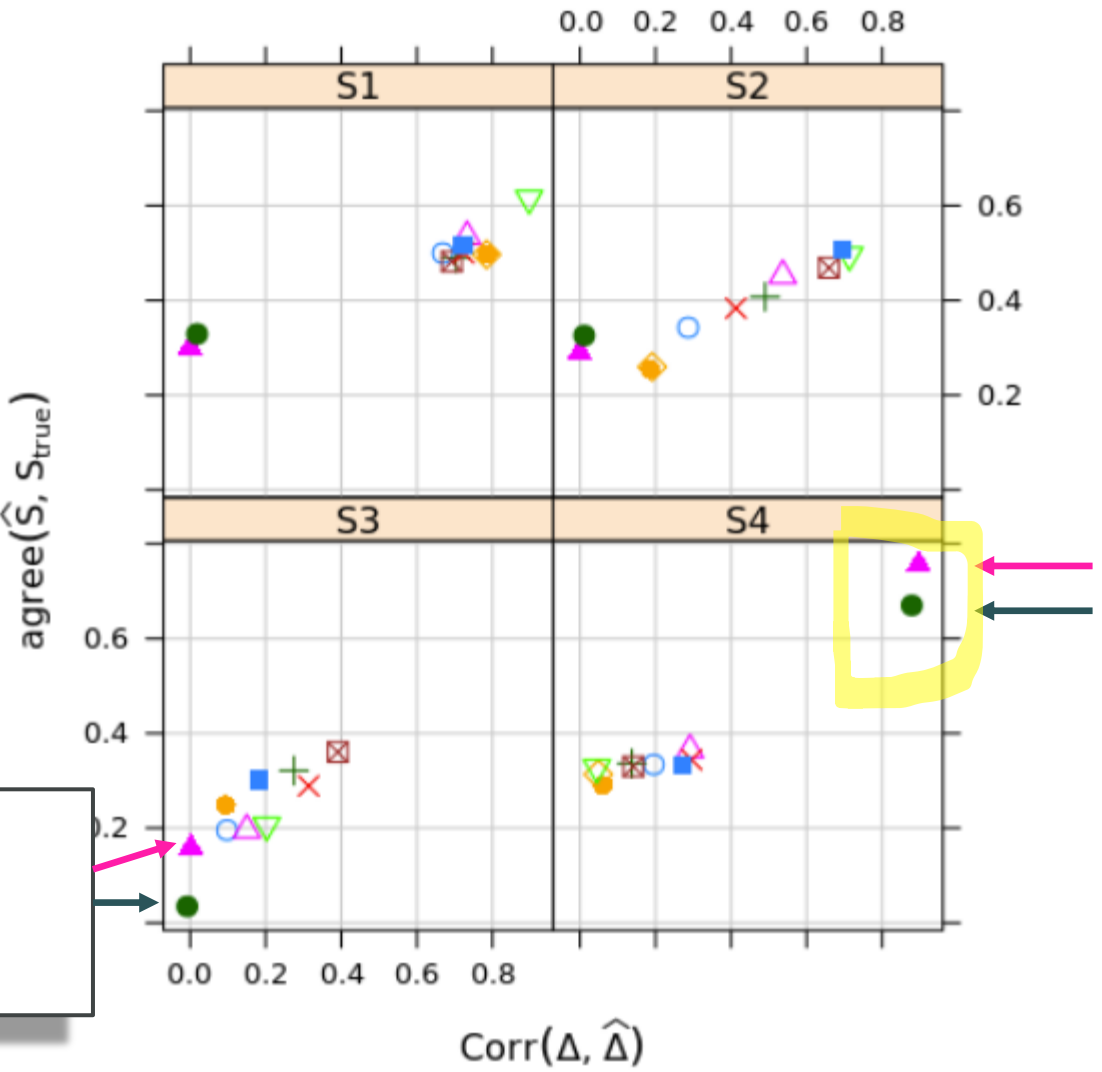


FIGURE 11 *The agreement between the true $S(X) = \{\Delta(X) > 0\}$ and identified subgroup measured by the Jaccard coefficient vs. Pearson correlation between the true $\Delta(X)$ and estimated CATE*

Summary

No clear winning method in our benchmarking; some looked more solid than others.

- The difficulties reflects how inherently hard Subgroup Discovery is

Large differences in Bias-Variance tradeoffs across methods.

Peculiar results with Causal Forest and A-Learning/Weighting Methods sometimes.

Plenty of scope for further research.

THANK YOU

References

[1] Lipkovich I, Dmitrienko A, B. D'Agostino Sr. R. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine* 2017; 36: 136–196. doi: 10.1002/sim.7064

[2] Lipkovich I, Svensson D, Ratitch B, Dmitrienko A. Overview of modern approaches for identifying and evaluating heterogeneous treatment effects from clinical data. *Clinical Trials* 2023; 20(4). doi: DOI: 10.1177/17407745231174544

[3] Lipkovich I, Svensson D, Ratitch B, Dmitrienko A. Modern approaches for evaluating treatment effect heterogeneity from clinical trials and observational data. Submitted Stat. In Medicine; *arXiv* 2023.

[4] Huling J, Menggang Y. Subgroup Identification Using the Personalized Package. https://arxiv.org/abs/1809.07905

[5] Nie X, Wager S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 2020; 108(2): 299–319

[6] Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 2018; 113(523): 1228–1242.

[7] Gutierrez P, Gerardy JY. Causal inference and uplift modeling. A review of the literature. *JMLR:Workshop and Conference Proceedings* 2016; 67.

[8] Foster JC, Taylor JM, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Statistics in Medicine* 2011; 30(24): 2867–2880.

Example of recent CATE approach: R-learning

Example of $\hat{f} = argmin_{\{f \in C\}}(\boldsymbol{L}(Y, Trt, \boldsymbol{x}; f))$ renders **Δ(x)** i.e., estimates CATE:

Nie Wager 2020 [4]: *'decomposition, **R1**esiduals on **R2**esiduals'*

$$\hat{\Delta}(\mathbf{x}) = argmin_{\{\boldsymbol{f} \in C\}} \left(\frac{1}{n} \Sigma_i \overbrace{\left(Y_i - \hat{m}^{\{-i\}}\right)}^{\text{R1}} - \overbrace{\left(T_i - \hat{\pi}^{\{-i\}}\right) \boldsymbol{f}(\boldsymbol{x_i})}^{\text{R2}} \right)^2$$

R1= Residuals: (Outcome – Outcome.model) ("$\hat{m}^{\{-i\}}$" = cross-fitted prognostic model) R2= Residuals: (Treatment – Treatment.propensity.model) ("$\hat{\pi}^{\{-i\}}$" = cross-fitted prop.scores)

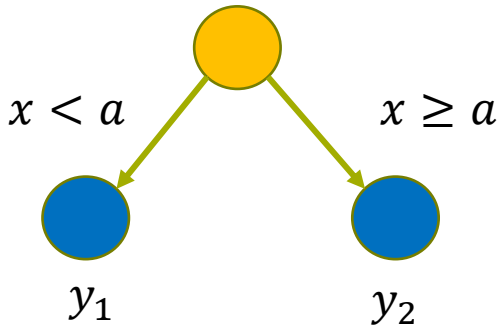
Possible to rewrite expression to $\frac{1}{n} \Sigma_i \left(w_i (Y_i^* - \boldsymbol{f}(\boldsymbol{x_i}))^2 \right)$ with Y_i^* a 'modified outcome', and weights =residual trt-propensities.

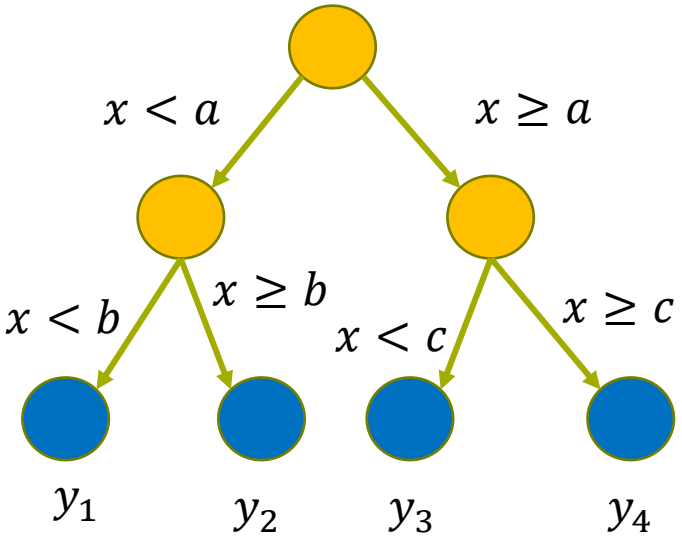
- Off-the-shelf "standard" **XGBOOST can estimate this (**squared Loss & weights).

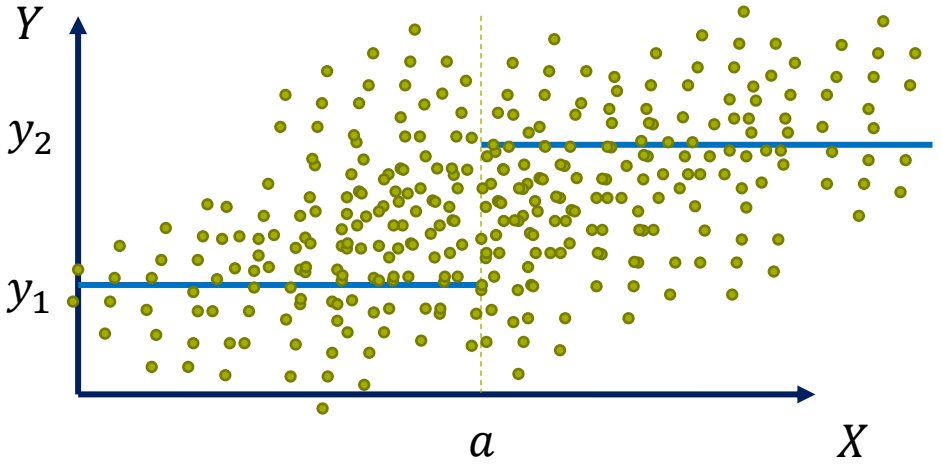
A regression tree

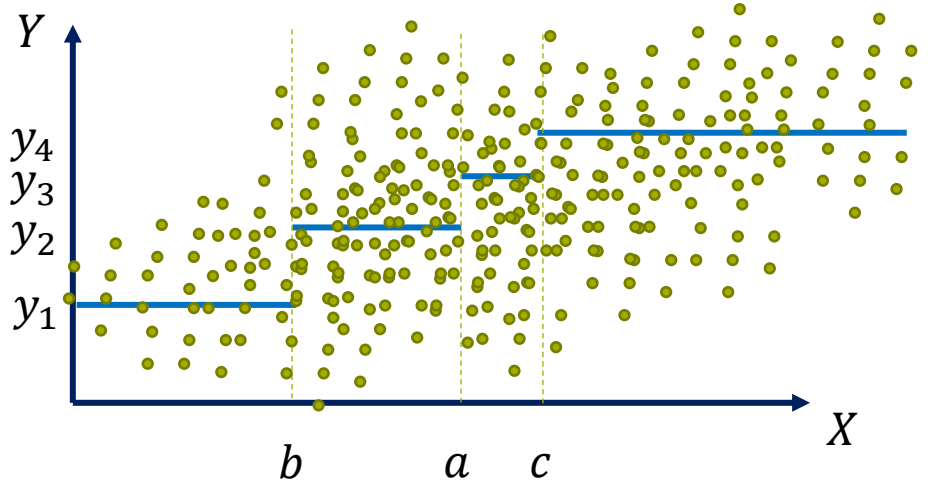
Slide kindly shared by Stefan Franzén (AstraZeneca)

Treatment effect heterogeneity – a practical example [PSI2022]









A Causal tree

Slide kindly shared by Stefan Franzén (AstraZeneca)

Treatment effect heterogeneity – a practical example [PSI2022]