Utilizing Shapley Values to Identify Predictive Biomarkers in CATE Modelling

Erik Hermansson[1]

In collaboration with David Svensson[1], Ilya Lipkovich[4], Nikos Nikolaou[2] and Konstantinos Sechidis[3]

1. AstraZeneca, Statistical innovation
2. UCL, London, UK
3. Novartis, Basel, Basel, Switzerland
4. Eli Lilly and Company, Indianapolis, Indiana, USA

Rubin's Potential Outcome framework:

Following Rubin's Potential Outcome Framework [7], each patient has two potential outcomes, denoted as Y(0) and Y(1), corresponding to Trt=0 and Trt=1, respectively. 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 $\Delta(\mathbf{x}) := \mathbb{E}[Y^{(1)} - Y^{(0)} | \mathbf{X}=\mathbf{x}]$, 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 ...

Subj	Trt	$Y^{(0)}$	$Y^{(1)}$	ITE
1	1	?	3	?
2	1	?	1	?
3	0	2	?	?
4	1	?	1	?
5	0	3	?	?
6	0	0	?	?

Stressing

- Target is CATE

... for a patient represented by these covariates

$$\Delta(\mathbf{x}) := E[Y(1) - Y(0) \mid X=\mathbf{x}]$$ as a (multivariate) function of $\mathbf{x}=(x1,...,xp)$

Expected (individual) trt. Effect ...

- Might be considered when a trial fails to convince in the average sense.
  - 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?
- Interestingly, several 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”

Prognostic vs Predictive variables

- Target is CATE

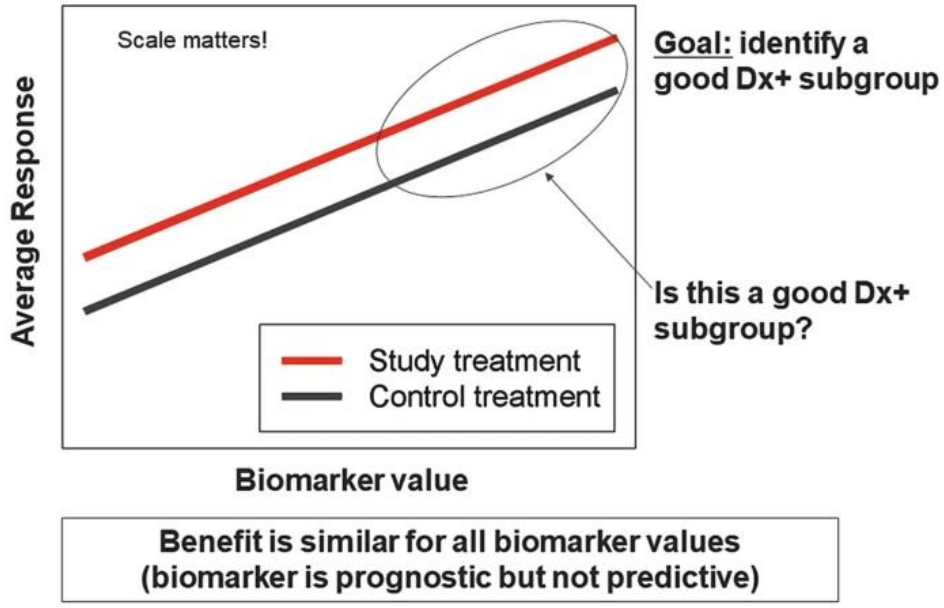
… for a patient represented by these covariates

$$\Delta(\mathbf{x}) := E[Y(1) - Y(0) \mid X=\mathbf{x}]$$ as a (multivariate) function of $\mathbf{x}=(x1,\ldots,xp)$

Expected (individual) trt. Effect …

A variable is predictive if CATE varies systematically; conversely, prognostic variables maintain a constant effect.

Fig. 11.2



How to estimate CATE

Method	Number of models	Propensity?	CATE modelling
T-Learner	2	Implicit	$Y^{(0)}$ and $Y^{(1)}$
S-lear			$Y^{(0)}$ and $Y^{(1)}$
X-lear			$Y^{(0)}$ and $Y^{(1)}$
Causa			$Y^{(0)}$ and $Y^{(1)}$
R-lear			Modified outcome
Doubly Robust		Explicit	Modified outcome

Modified Outcome methods target CATE directly without modelling each potential outcome.
- e.g., multiply Y with 2*Treatment (-1 or 1) and regress that against X

1. fit f(x) on extended data set
2. $\Delta(\mathbf{x}) = f(x, Trt = 1) - f(x, Trt = 0)$

See Jacob [3] on a survey of modern approaches for CATE modelling

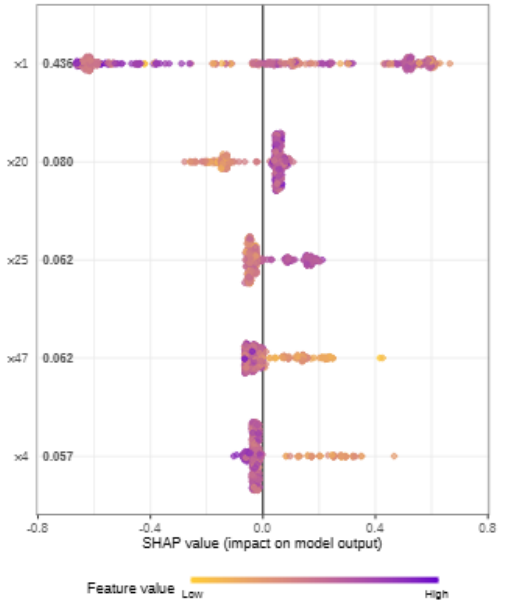
Using SHAP to identify predictive biomarkers

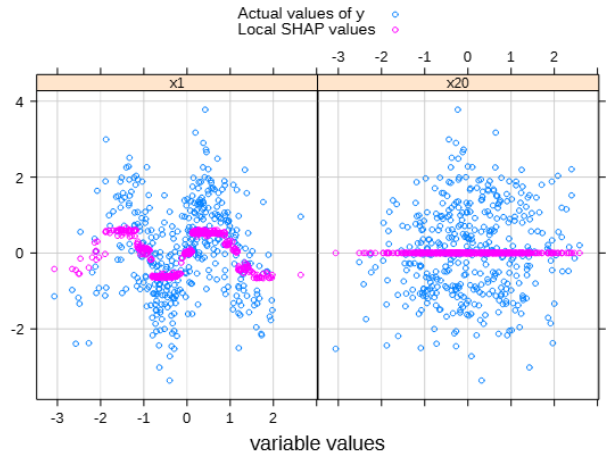
- Note: important to note the difference between predictive and prognostic
  - Y = x_1 + trt + trt*x_2

- Given estimates of CATE, we regress it against the baseline variables: $\Delta(\mathbf{x}) \sim \mathbf{x}$

  using an xgboost model and derive SHAP from it.

- We can now use SHAP values to estimate which covariates have the largest impact on CATE.
  - Instance level SHAP – *How does the importance depend on the covariate value*
  - Global SHAP – *How important is the covariate compared to other*





Note on SHAPLEY values

- Shapley values is a game theoretical concept, "SHAP" is the version for variable importance in Machine Learning.
  - Popular, current standard now. Honest estimate of model importance
  - But the model might still be wrong!

- In many cases we need an additional model to get SHAP values, but for the modified outcome models we can model them directly without an extra model
  - However, this does not impact performance!
  - Many different SHAP methods available, many very computationally expensive. Treeshap is however a good performer and fast to compute.

- .

Simulation Landscape: S2-S3

Simulation	No. Prognostic x	Trial Type	TRT assignment	Predictive x
S2	**19**	**RCT**	3:1 rand (more active)	x3, x4
S3	**19**	Observational	Prognostic 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

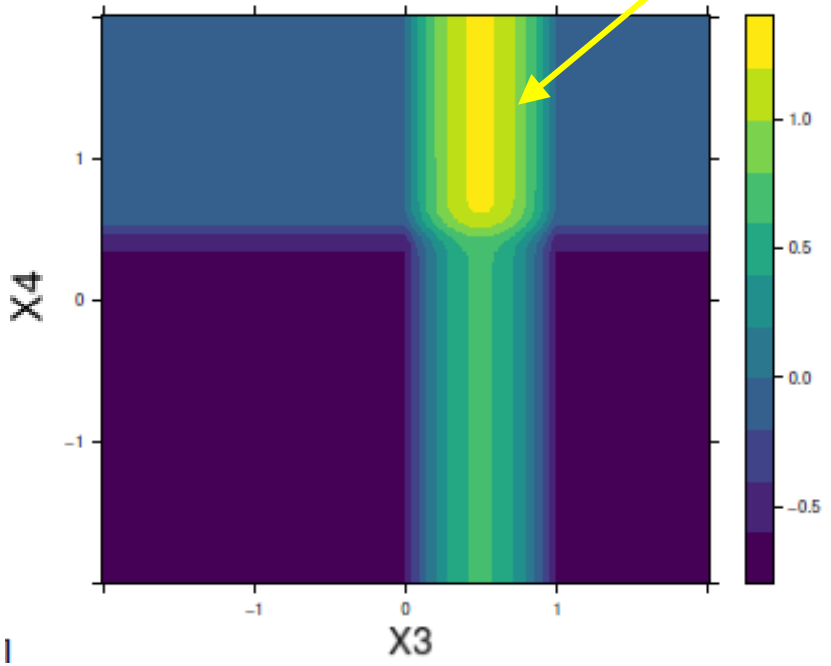
True Treatment Effects?

- non-linear, non-monotone

True **S={Δ(x)>0}** has size 0.33,

True average CATE in S is 0.665.

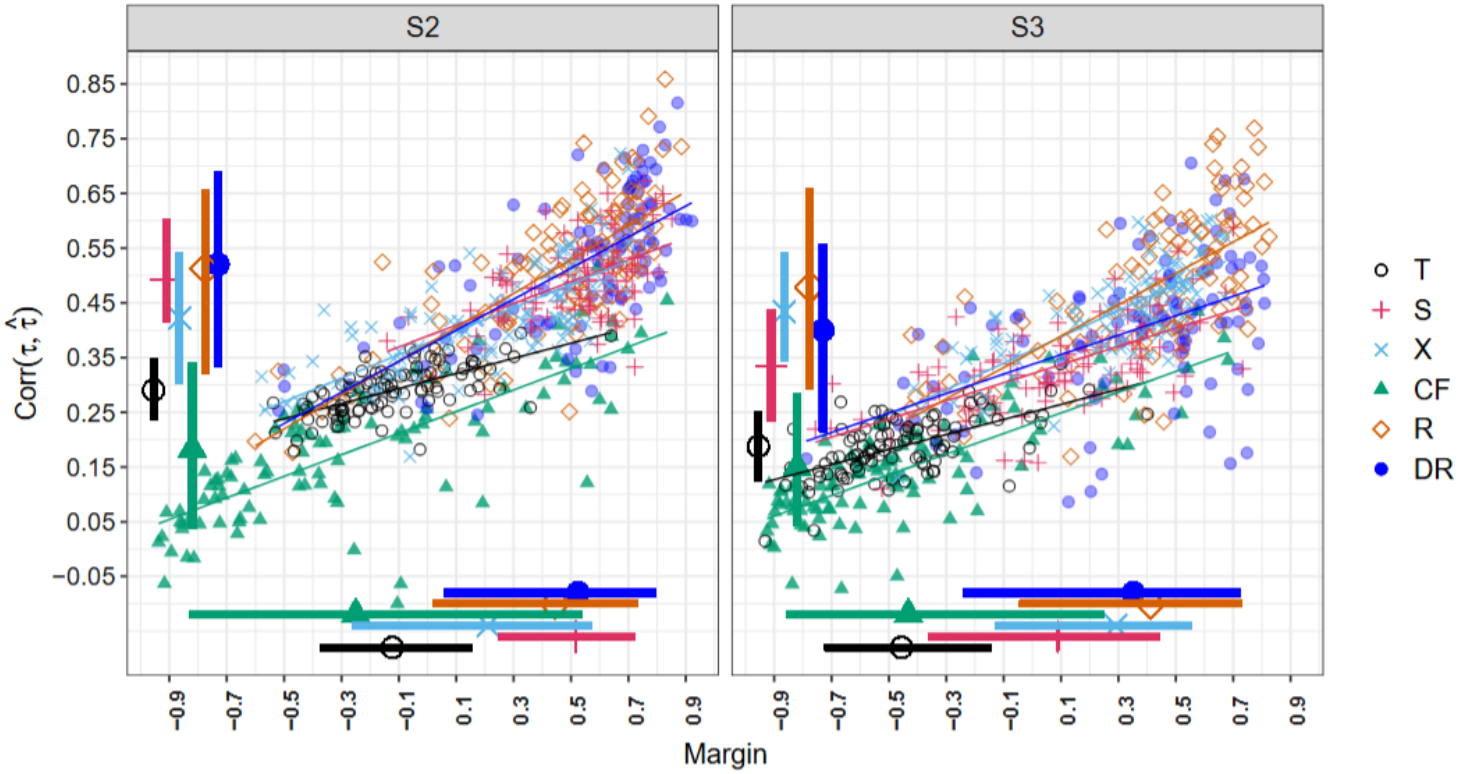
Overall true effect = 0.0119



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

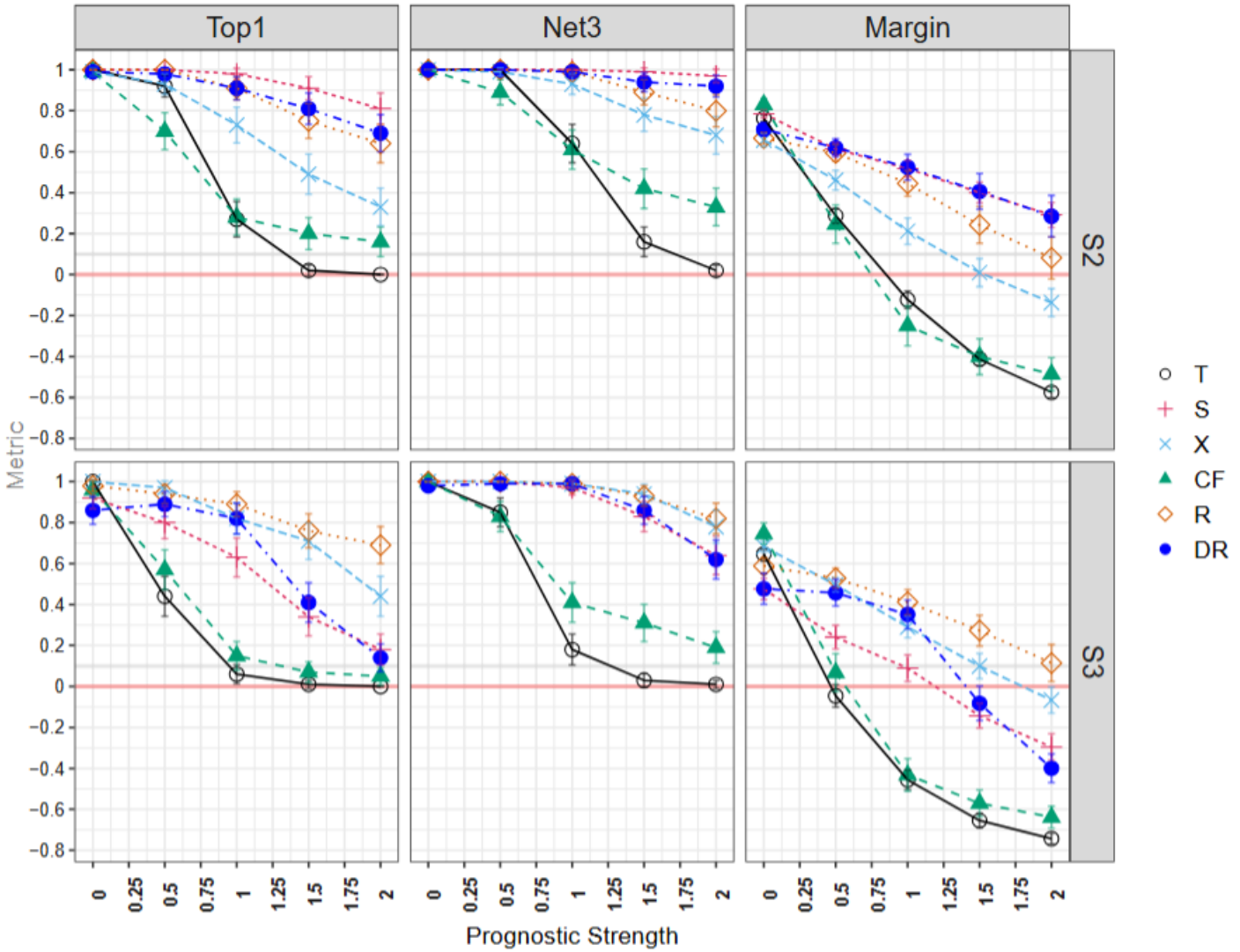
Estimating CATE correctly is vital!

- Without a good estimate of CATE, we can not separate between predictive and prognostic covariates
- Margin = How large is the separation in Shap values between the predictive and prognostic covariates
- Some models have negative margin...



Simulations

- Top 1 – how often do we select a predictive biomarker
- Top 3 – how often is a predictive biomarker in the top 3
- Margin – Ability to distinguish predictive and prognostic biomarkers
- Grey line – performance of random guess for top 1



Some concluding thoughts

- Hard to tell which model is the best
  - R and DR are strong performers, but can be complex to implement
    - S learner was strong on non-RWE data
  - *But easier to say which to avoid!*
    - Causal forest is not 'honest'? Surprisingly bad
    - T learner performs badly, mostly due to regularization biases as one model is fitted to each arm

      and prognostic effects gets mismodelled. This has been shown in Hermansson [1] and Lipkovich [4] as well.
  - **SHAP values is not a panacea** ("insights" vs "inference")
    - Explains the fitted model regardless of how good it is. [Garbage in, garbage out]
    - (and hard to assess a CATE model in the practice; fundamentally unobservable target)
    - Does not identify a subgroup
      - Other methods does this directly (GUIDE, MOB etc)

References

1. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep. *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems*.
2. Hermansson, E. &. (2021). On Discovering Treatment-Effect Modifiers Using Virtual Twins and Causal Forest ML in the Presence of Prognostic Biomarkers. *International Conference on Computational Science and Its Applications*.
3. Jacob, D. (2021). CATE meets ML. *Digital Finance*.
4. Lipkovich, I., Svensson, D., Ratitch, B., & Dmitrienko, A. (2023). Overview of modern approaches for identifying and evaluating heterogeneous treatment effects from clinical data. *Clin Trials*.
5. Ruberg, S. J. (2021). Assessing and communicating heterogeneity of treatment effects for patient subpopulations: The hardest problem there is. *Pharm Stat.*
6. Sechidis K, S. S. (2024). WATCH: A Workflow to Assess Treatment Effect Heterogeneity in Drug Development for Clinical Trial Sponsors. *Pharm Stat.*
7. Rubin, D. B. (2005). Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*

AstraZeneca PLC, 1 Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge, CB2 0AA, UK
+44(0)203 749 5000
www.astrazeneca.com