Explainable AI for Causal Inference and Heterogeneous Treatment Effect Estimation via AI/ML: A conceptual framework for late phase clinical trials

AI & ML SIG: Artificial Intelligence updates and applications

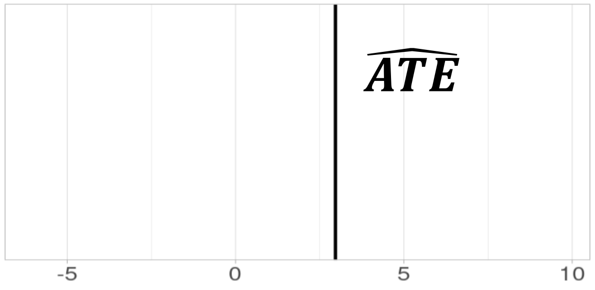
PSI Conference 2025, London

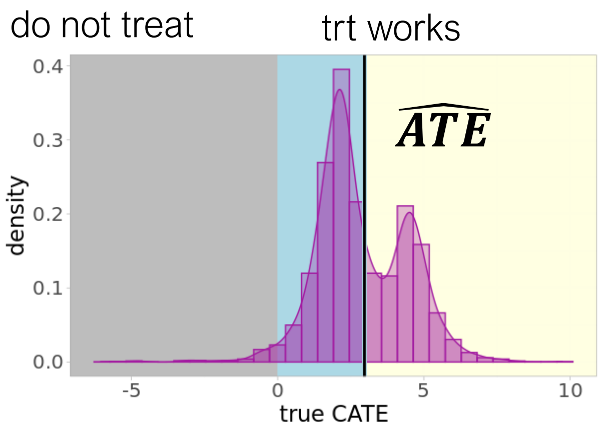
Lukas Lasota (ClinStat), Frauke Hennig (Chrestos)
Karl Köchert (Sanofi), Eliana Garcia Cossio (Bayer)

"We are a community dedicated to leading and promoting the use of statistics within the healthcare industry for the benefit of patients."

Background

- In standard RCT the Average Treatment Effect (ATE) is estimated, assuming homogeneity
- However, not everyone benefits equally from a treatment. #heterogeneity #precisionmedicine
- Conditional average treatment effect (CATE) refers to the ATE in specific subgroups of individuals conditioned on their characteristics or covariates #subgroupeffects





Background

- Machine learning (ML)-based causal inference allow CATE estimation in the presence of **complex and high dimensional covariate profiles**
  - Direct CATE estimation #causalforest #causalboost
  - Indirect CATE estimation (decomposition of CATE in multiple ML models, aka base learners) #metalearner

- ML models ( ) are tailored for prediction tasks
  - How to understand and trust CATE estimation (causal construct to estimate treatment effect)?
  - How to detect covariates driving treatment effect heterogeneity (TEH)?
  - How transferable is explainable AI (**xAI**) to CATE modelling?

Framework: xAI for CATE

CATE estimation via ML

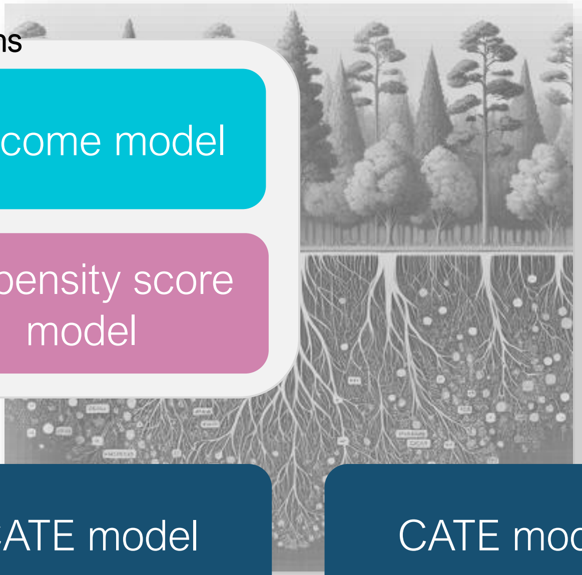
Indirect
Meta-Learners,
e.g. R-Learner

Direct
Causal Forest

Nuisance functions

Outcome model

Propensity score model



Base Learner / Model
Random Forest

Pseudo-Outcome

CATE model

CATE model

ML model → xAI

Explainable Results for clinical decision-making

Are the results reliable?
#trust #performance

R^2, RMSE

R^2, RMSE | ROC-AUC, Accuracy

Framework: xAI for CATE

CATE estimation via ML

Indirect

Meta-Learners,
e.g. R-Learner

Direct

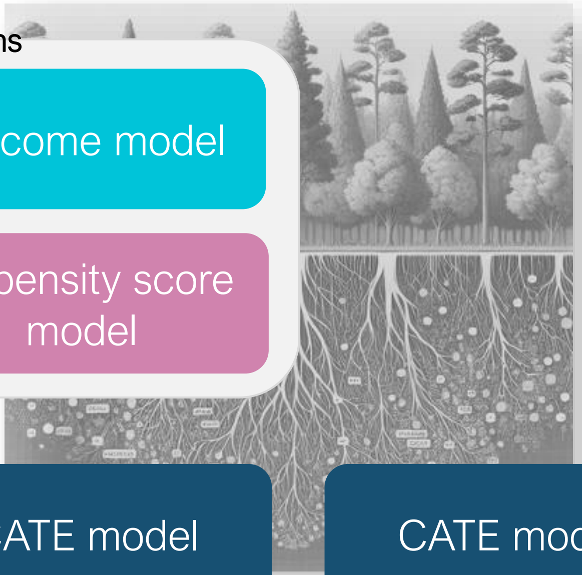
Causal Forest

Nuisance functions

Outcome model

Propensity score model

Pseudo-Outcome



Base Learner / Model
Random Forest

CATE model

CATE model

ML model → xAI

Explainable Results for clinical decision-making

Are the results reliable? **#trust #performance**	R^2, RMSE R^2, RMSE | ROC-AUC, Accuracy
Is there evidence for TEH? Reason for individualized treatment assignment?	AUPEC, global TEH-Test
What are the drivers for TEH? **#variableimportance**	permutation mean |shap| values
What are characteristics of TEH? **#dependencyplots**	PDPs shap values

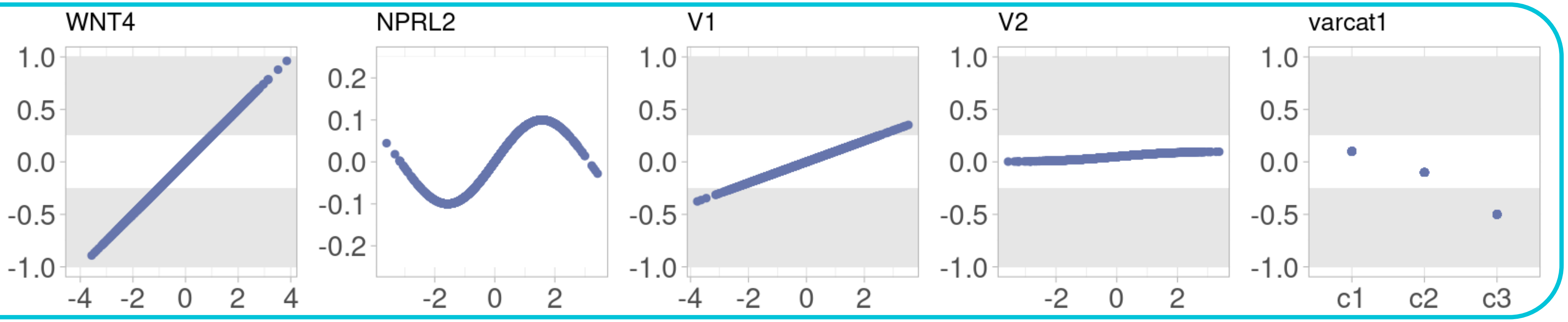
Outcome (Y) Generation

The mTOR signaling pathway was used to generate the gene expression matrix and define a correlation structure.

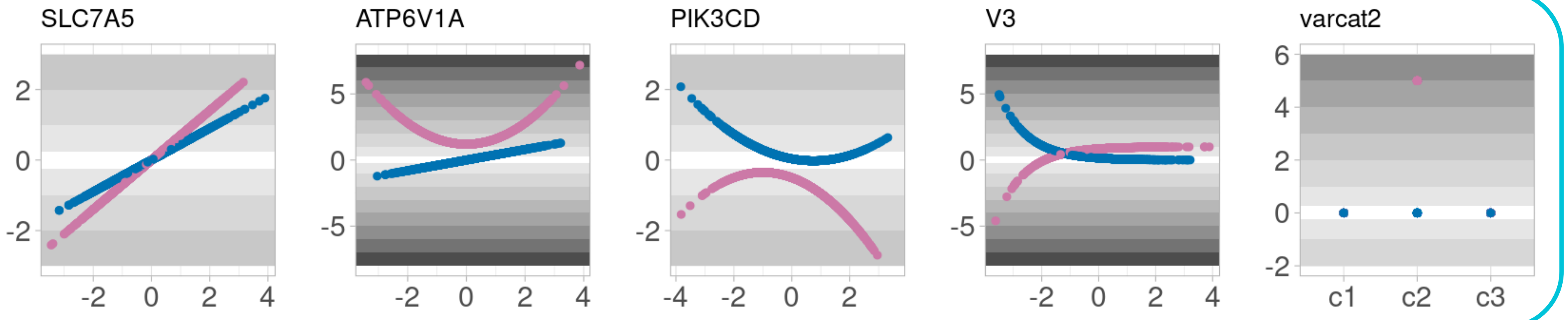
Treatment $(\beta = 0.75)$+

Visit: P041 Machine Learning in Precision Medicine: A Collaborative Approach. **PSI Biomarker SIG group.**

Prognostic:
- Gene 1 (WNT4)
- Gene 2 (NPRL2)
- Gene 3 (SLC7A5)
- Variable 1 & 2 (continuous)
- Varcat 1 (categorical)



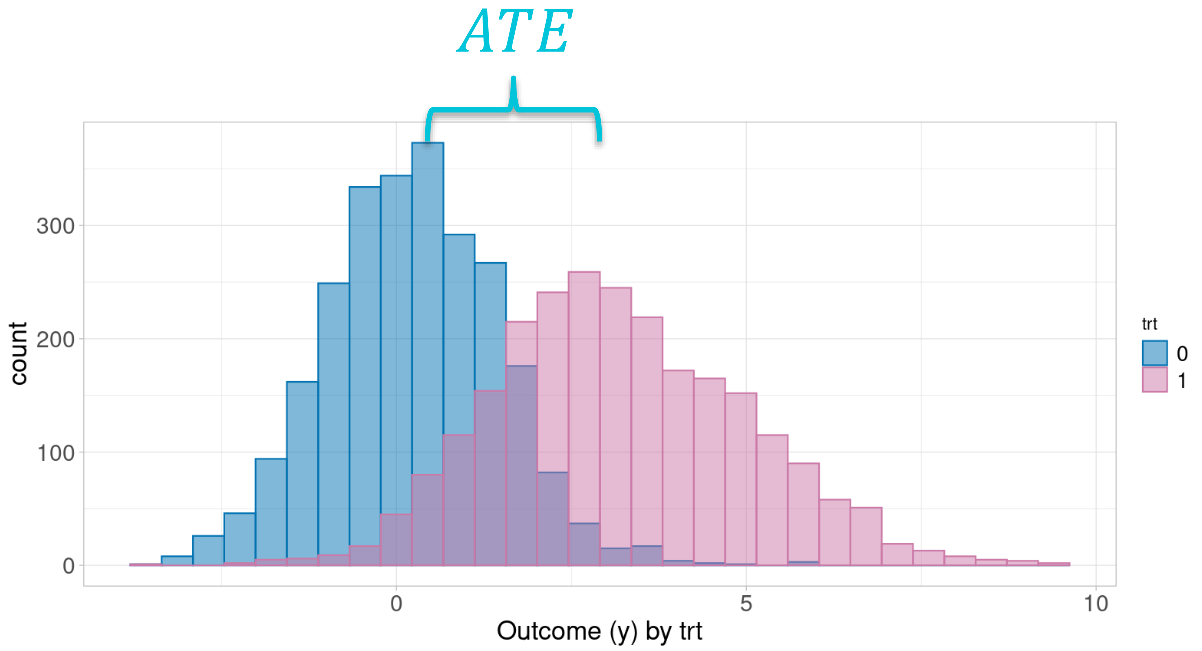
Predictive (+ prognostic):
Gene 3 (SLC7A5)
Gene 4 (ATP6V1A)
Gene 5 (PIK3CD)
Variable 3 (continuous)
VarCat 2 (categorical)



+ Noise Variables (40)

n= 5000, p = 50

Simulated average treatment effect (ATE)



Representative of continuous clinical endpoints

RCT: Average Treatment Effect

term	estimate	std.error
(Intercept)	0.234	0.030
trt1	2.967	0.043

$\widehat{ATE}$

trt increases endpoint x

CATE estimation via ML

Method of choice

Validity check

Indirect

Meta-Learners,
e.g. R-Learner

Nuisance functions

Outcome model

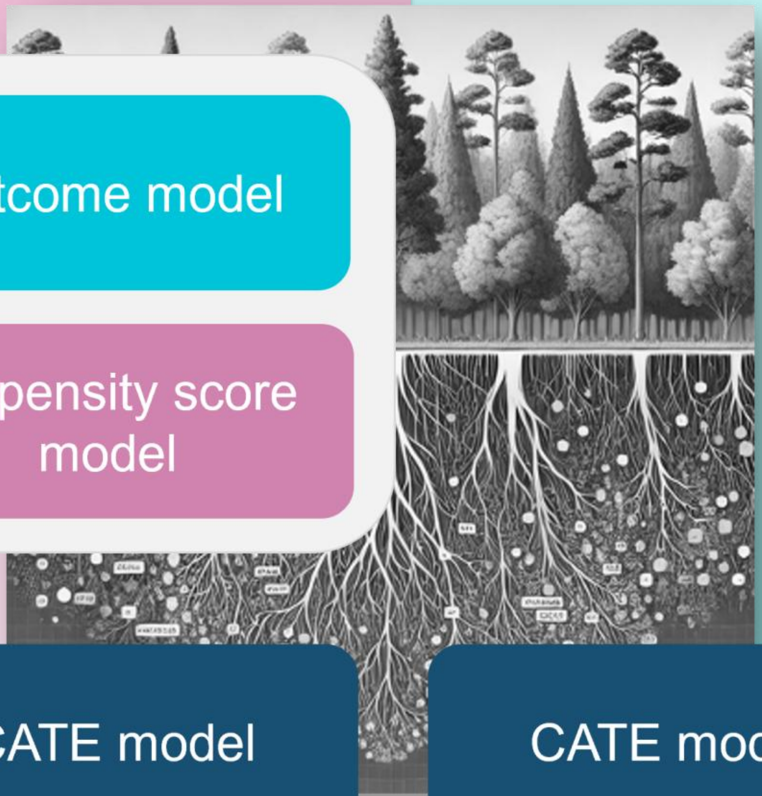
Propensity score model

Pseudo-Outcome

CATE model

Direct

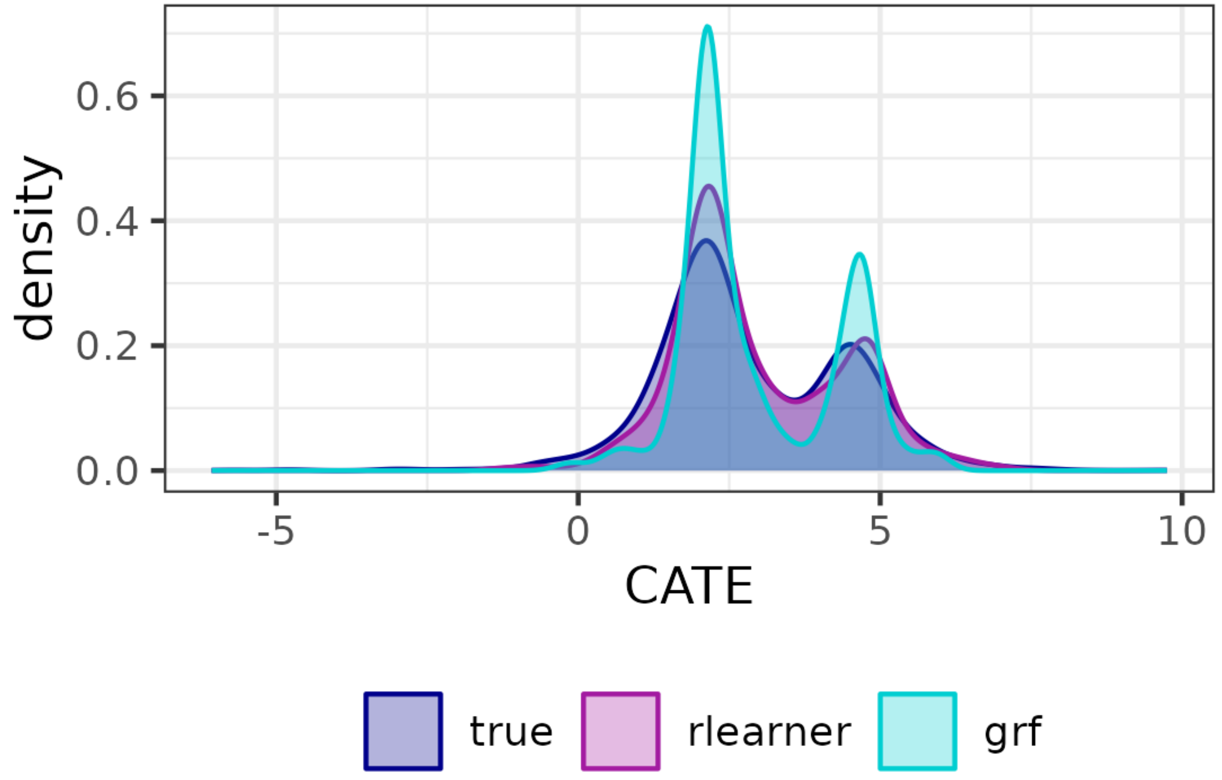
Causal Forest

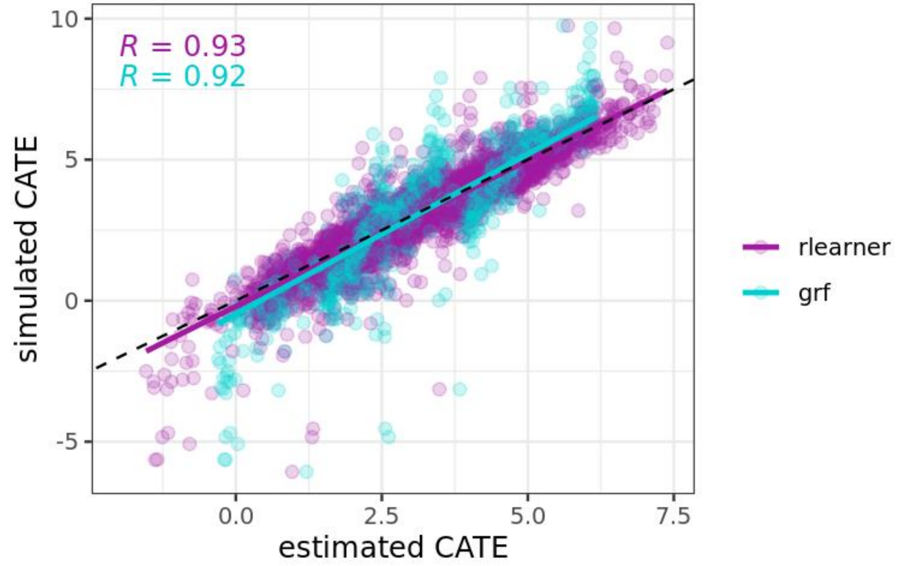


Base Learner / Model
Random Forest

CATE model

Are the results reliable? #trust #robustness





CATE	Mean (ATE)	SD
true	2.87	1.57
rlearner	2.97	1.41
grf	2.96	1.29

Are the results reliable? #performance

CATE model ($\hat{\tau}$)
Pseudo-Outcome ~ X

Metric	rlearner (oob)
R^2 (OOB)	0.30

Some explained Variance → **predictive** factors (but: #pseudo-outcome)

Nuisance functions

Outcome model ($\hat{\mu}$)
Y ~ X

Metric	Estimate
R^2 (OOB)	0.120

Some explained Variance → **prognostic** factors

Propensity score model ($\hat{\pi}$)
trt ~ X

Metric	Estimate
$ROC\ AUC$ (OOB)	0.514

Reliable random **treatment assignment**

Is there evidence for TEH?

Is the estimated CATE model useful to propose an individualized treatment rule (ITR)?

CATE model ($\hat{\tau}$)
Pseudo-Outcome ~ X

TEH?	rlearner (oob)	Causal Forest
evalITR::het.test()	<0.001	<0.001
Global Test ($\beta_2 \approx 1$)*	1.05 (<0.001)	1.15 (<0.001)
grf::test_calibration()		1.15 (<0.001)

AUPEC	rlearner (oob)	Causal Forest
evalITR::AUPEC()	0.408 (<0.001)	0.407 (<0.001)

*Using the approached proposed in: A Tutorial Introduction to Heterogeneous Treatment Effect Estimation with Meta-learners. Salditt et al., 2023.

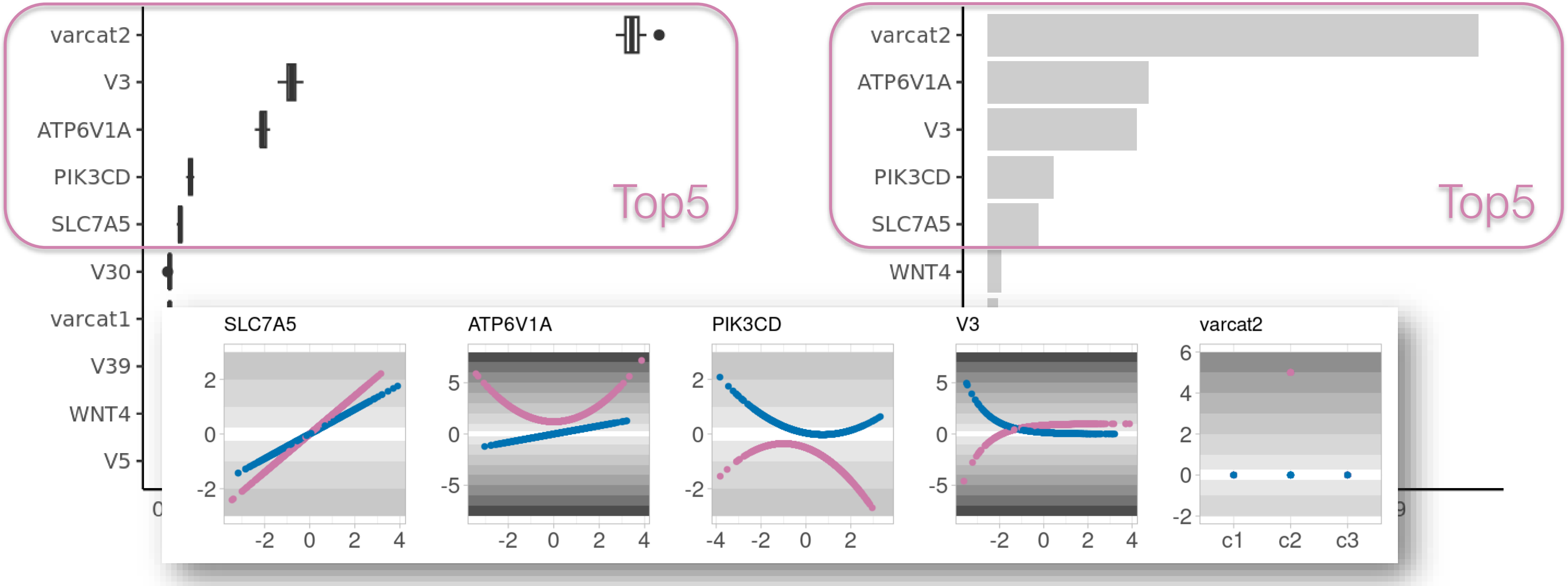
→ There is evidence of TEH

→ ITR superior to random treatment allocation
→ Model is able to rank individuals by their actual benefit from treatment

What are the drivers for TEH? #vimp

Top 10 based on permutation importance

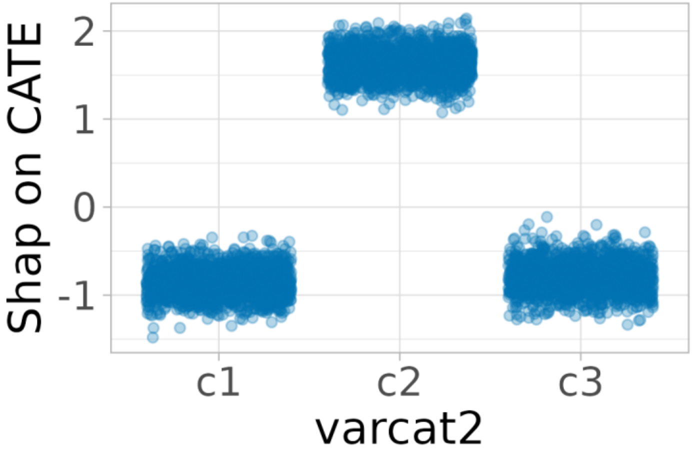
Top 10 based on mean absolute SHAP



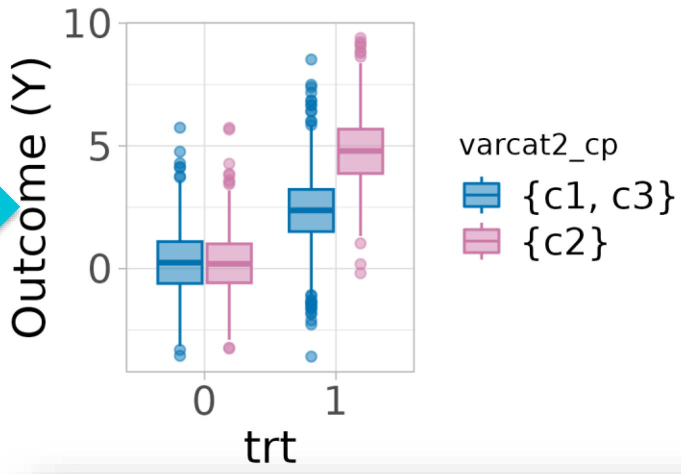
All that Glitters Is not Gold: Type-I Error Controlled Variable Selection from Clinical Trial Data. Zimmermann et al., 2024.
Modern approaches for evaluating treatment effect heterogeneity from clinical trials and observational data. Lipkovich et al., 2024.

What are characteristics of TEH? #dependencyplots #contribution

PDP-plots show similar patterns

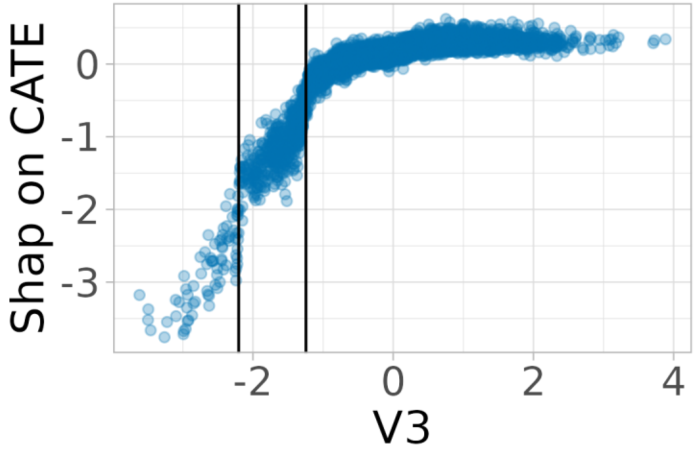


shap > 0
vs.
shap ≤ 0



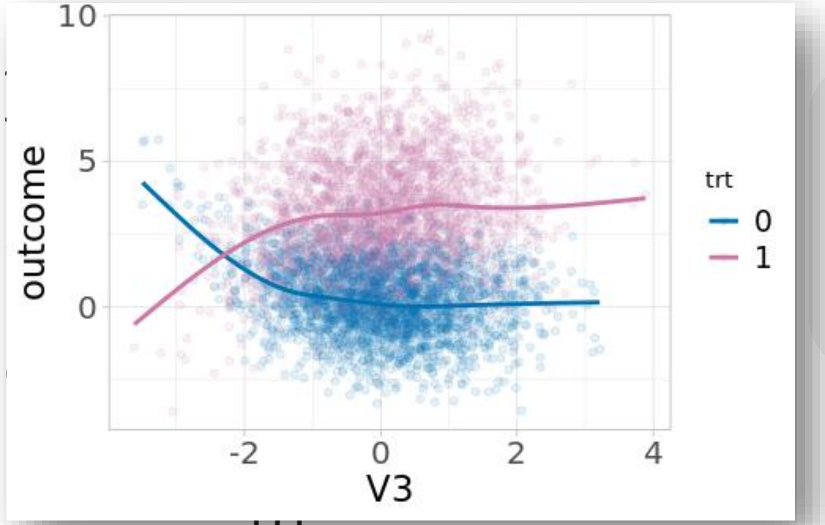
Potential Hypothesis

→ trt works better in {c2}



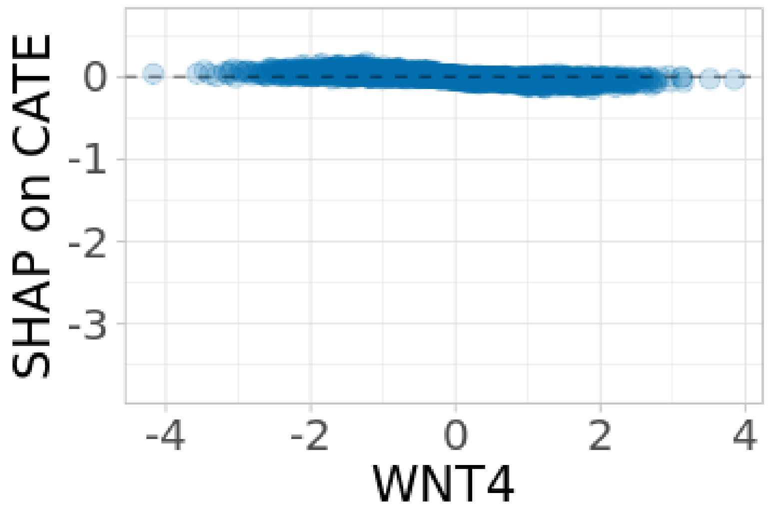
Changepoint Detection

{changepoint} R-Package



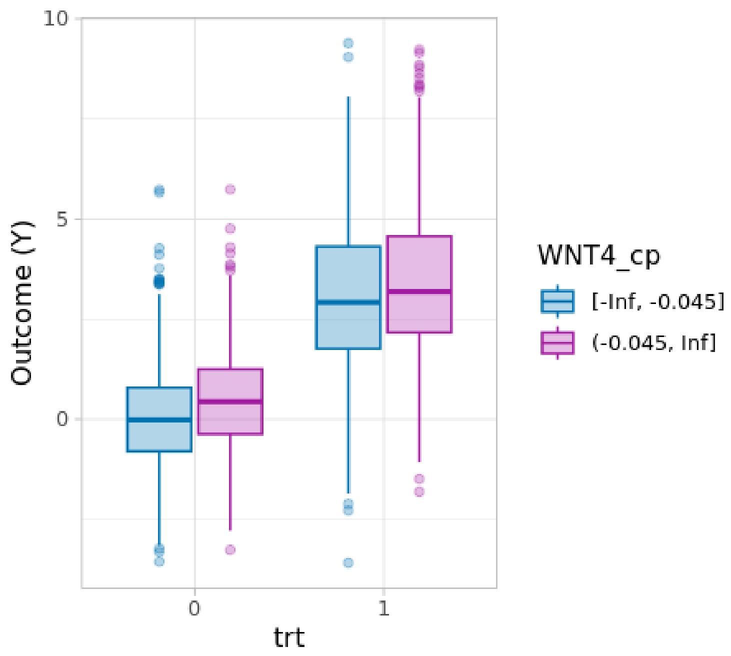
→ Why would you treat patients with V3 < -2.2?

What are characteristics of TEH?
#dependencyplots #negativecontrol



prognostic variable

no changepoints detected
→ choose median



→ No predictive value

Summary (Stop & Go)

CATE model

xAI

Explainable Results for clinical decision-making

CATE Modelling

Outcome model

Propensity score model

R^2, RMSE

ROC-AUC, Accuracy

Are the results reliable? **#trust #performance**

R^2, RMSE

Is there evidence for TEH? Reason for individualized treatment assignment?

AUPEC, GATES, global TEH-Test

What are the drivers for TEH? **#variableimportance**

permutation
mean |shap| values

What are characteristics of TEH? **#dependencyplots**

PDPs
shap values

xAI

#performance

#variableimportance

#relationships

Imbalanced Treatment assignment?

Trust in treatment randomization?

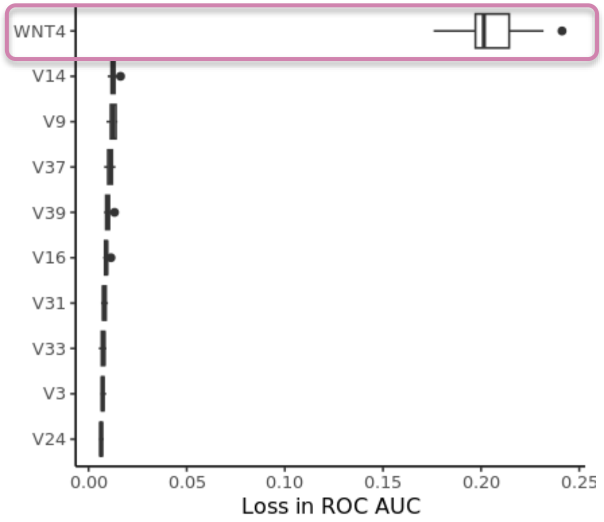
#performance

Propensity model ($\hat{\pi}$)

Metric	Estimate
ROC-AUC	0.60

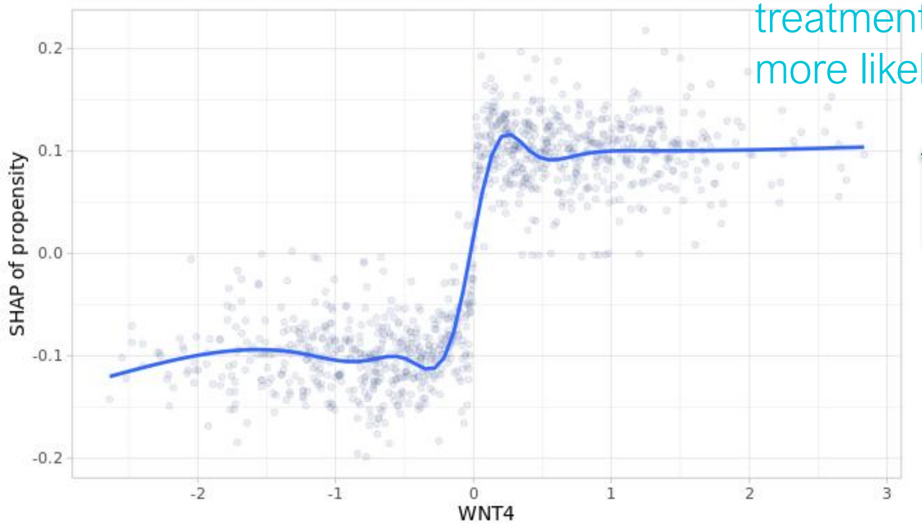
What variables led to imbalanced treatment assignment?

#variableimportance



How did random treatment assignment fail?

#relationships



treatment more likely

→ consider **WNT4** for next trial or include covariate in clinical model if it was unknown

Discussion

- Identifying drivers for TEH & characteristics using xAI:
  - Multiple choices for variable importance (e.g. permutation, SHAP)
  - Multiple ways to select drivers, e.g. top 5, knockoffs
  - Multiple ways to propose potential cut-offs & subgroups

- Multiple ways, however, can also provide robustness & trust

Discussion

- CATE is a data-driven causal construct, not an inference method
  - In order to test/proof TEH independent data is needed or at least sample splitting or cross-fitting approaches are recommended
  - The relationship between model performance/hyperparameters, CATE estimation, and FNR/FPR is a complex question, warranting further investigation
- Applying xAI methods on CATE models can help to provide insights & generate hypothesis regarding potential drivers and subgroups characteristics
- However, a few simulation cases are not sufficient to draw conclusions

Frauke Hennig (Chrestos)
frauke.hennig@chrestos.de

Lukas Lasota (ClinStat)
lukas.lasota.ext@bayer.com

Thank you

Karl.Koechert@sanofi.com
eliana.garciacossio@bayer.com