Machine Learning in Precision Medicine: A Collaborative Approach

The PSI Biomarkers SIG Machine Learning / AI workstream group

Laura Schlieker

June 16, 2026

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

Background

Assessment of predictive biomarkers is essential in precision medicine

Advances in high-throughput technologies enable simultaneous measurement of up to millions of biomarkers

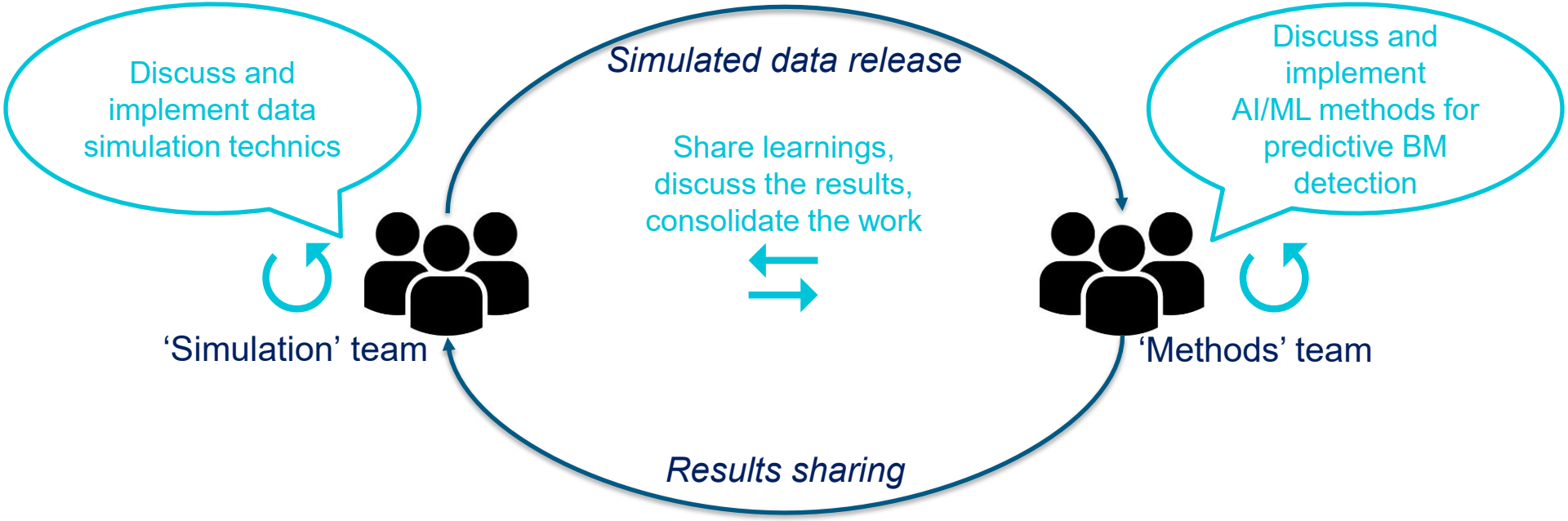
Traditional approaches struggle with scale, complexity and non-linear structure

We combine causal inference with Machine Learning (ML) and modern Artificial Intelligence (AI)

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

A biomarker/ML ESIG working group initiative

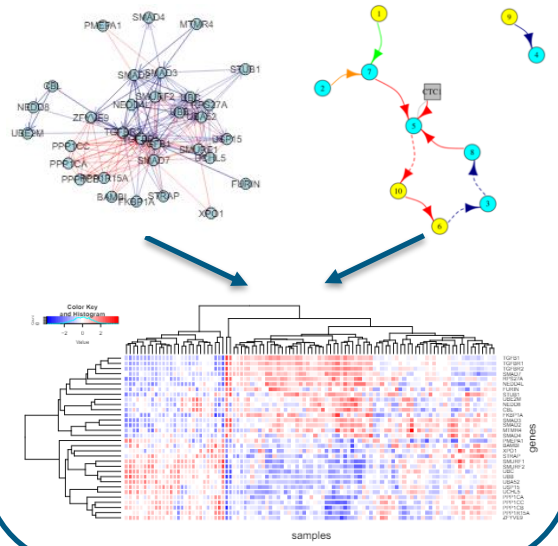
A Kaggle-like approach to continuously create synergies and learnings with fun.



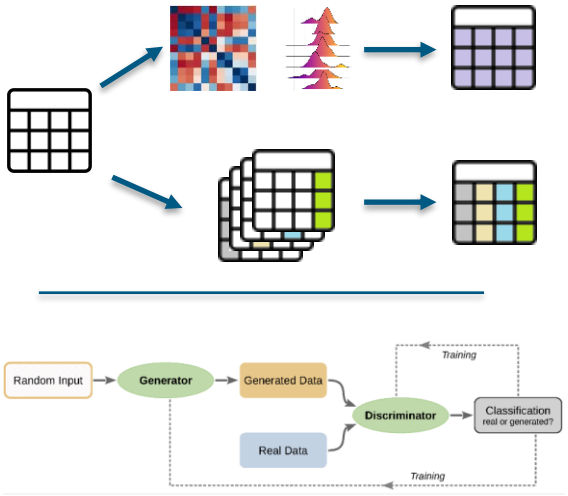
"We are a community dedicated to leading and promoting the use of statistics within the healthcare industry for the benefit of patients."

Simulate data from a complex and realistic biological setting

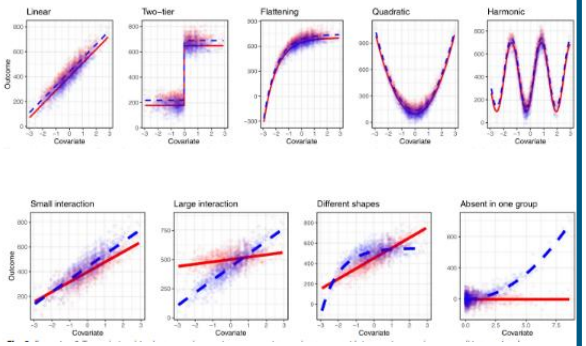
Simulation based on existing biological networks[1,2]



Simulation based on existing biological data[3,4]



Simulation based on artificial data generation[5,6]



1. graphsim R package (JOSS, 2020); 2. sismonr R package (Bioinformatics, 2020); 3. Schulz A et al. (BMC Medical Research Methodology, 2017); 4. RGAN R package (2022); 5. Tackney MS et al. (Trials, 2023)

Characteristics of simulated datasets

Characteristic	simScenario1	simScenario2
Dimensionality	n=500, p=305	n=500, p=2126
True effects	6 predictive + prognostic 3 prognostic only Only continuous	6 predictive, 4 prognostic miRNA pathway genes with small prognostic effect size Mix continuous, categorical
Functional form	Linear and quadratic forms	Linear and complex non-linear
KEGG pathways	PI3K-AKT (hsa04151) signalling pathway	Hepatocellular carcinoma (hsa05225), miRNAs in cancer (hsa05206), ErbB signalling (hsa04012) and mTOR signalling (hsa04150)

The data generating process was repeated 100 times

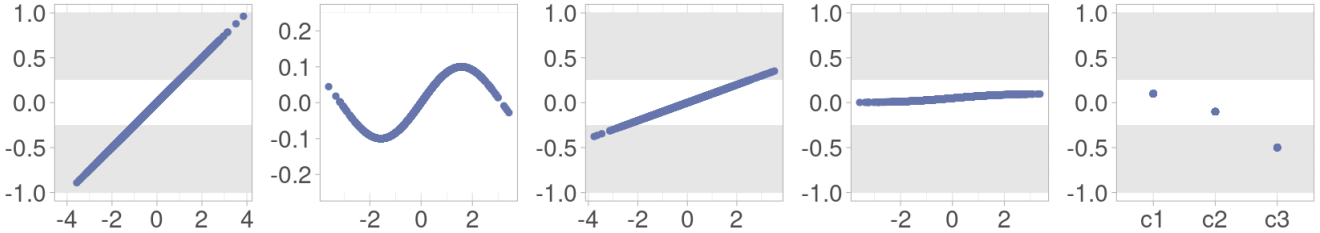
Outcome (Y) Generation Scheme

Various KEGG signaling pathway networks were used to generate a gene expression matrix (X) with a specific correlation structure. Each simulation scenario contained a selection of complex associations between a "biomarker" and outcome, with or without an additional interaction with treatment TRT (-> cause of TEH).

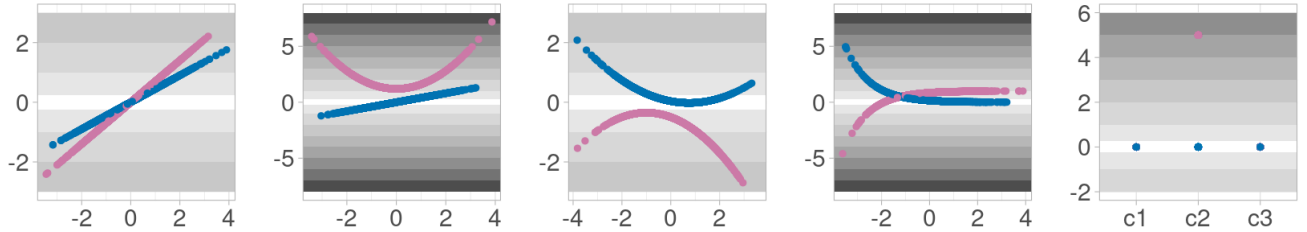
$$Y \sim \beta_0 + \beta_{TRT} * TRT + f_{prog}(X) + f_{pred}(X, TRT) + \varepsilon$$

$$\varepsilon \sim N(0; \sigma^2)$$

Example of functional form for prognostic biomarkers



Example of functional form for predictive biomarkers



Control arm Experimental arm

Methods

AI/ML-based Causal Inference allows **CATE estimation** in the presence of complex and high-dimensional covariates.

CATE estimation	Causal Inference framework	Base Learner
Indirect	S-Learner	Penalized Regression
Indirect	R-Learner	XGBoost
Indirect	Double-Robust	TabPFN (version 2.5*)
Direct	Causal Forest	

*TabPFN version 3 expected to perform much better with higher dimensionality data

Evaluation of AI/ML-based Causal Inference

For each method and each scenario, we evaluated:

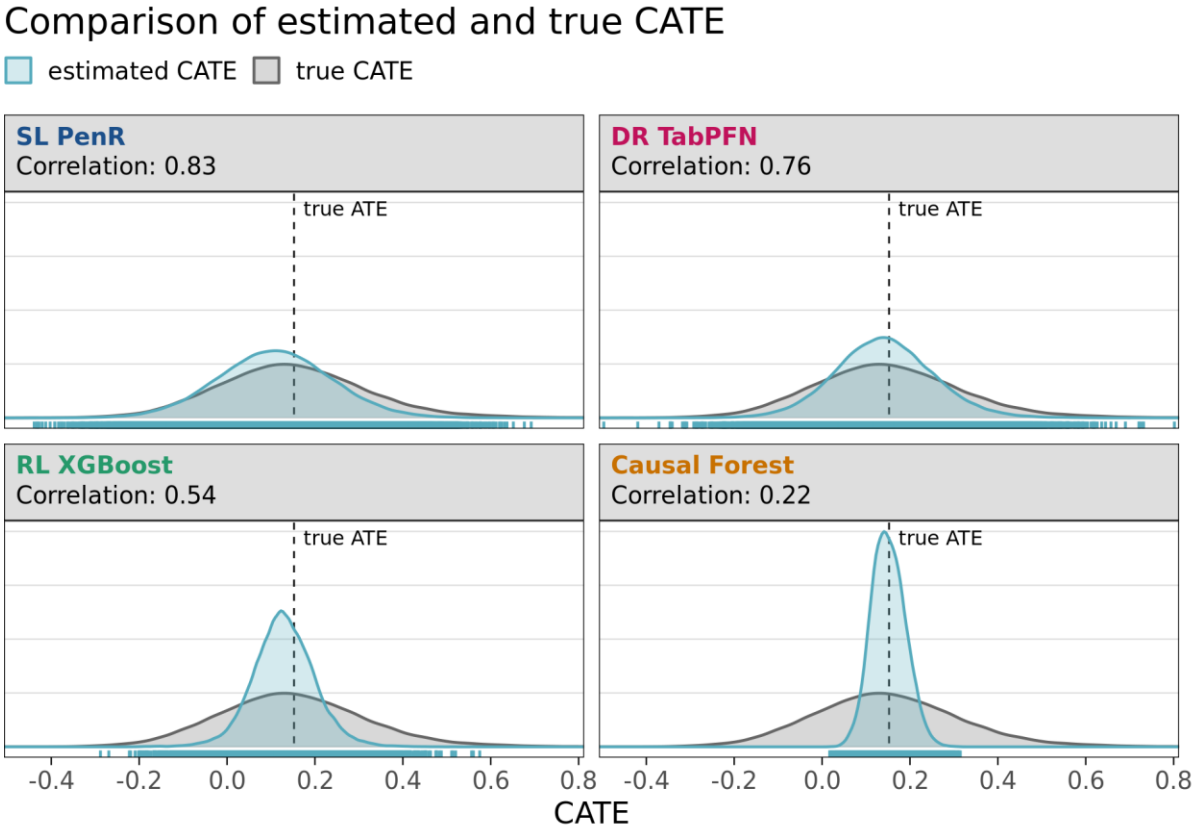
How reliable are the **CATE estimates**?

→ Spearman correlation between true and estimated CATE

Are **true predictive effects** identified?

→ TOP6 variables based on SHAP importance across simulations

simScenario1 - CATE performance

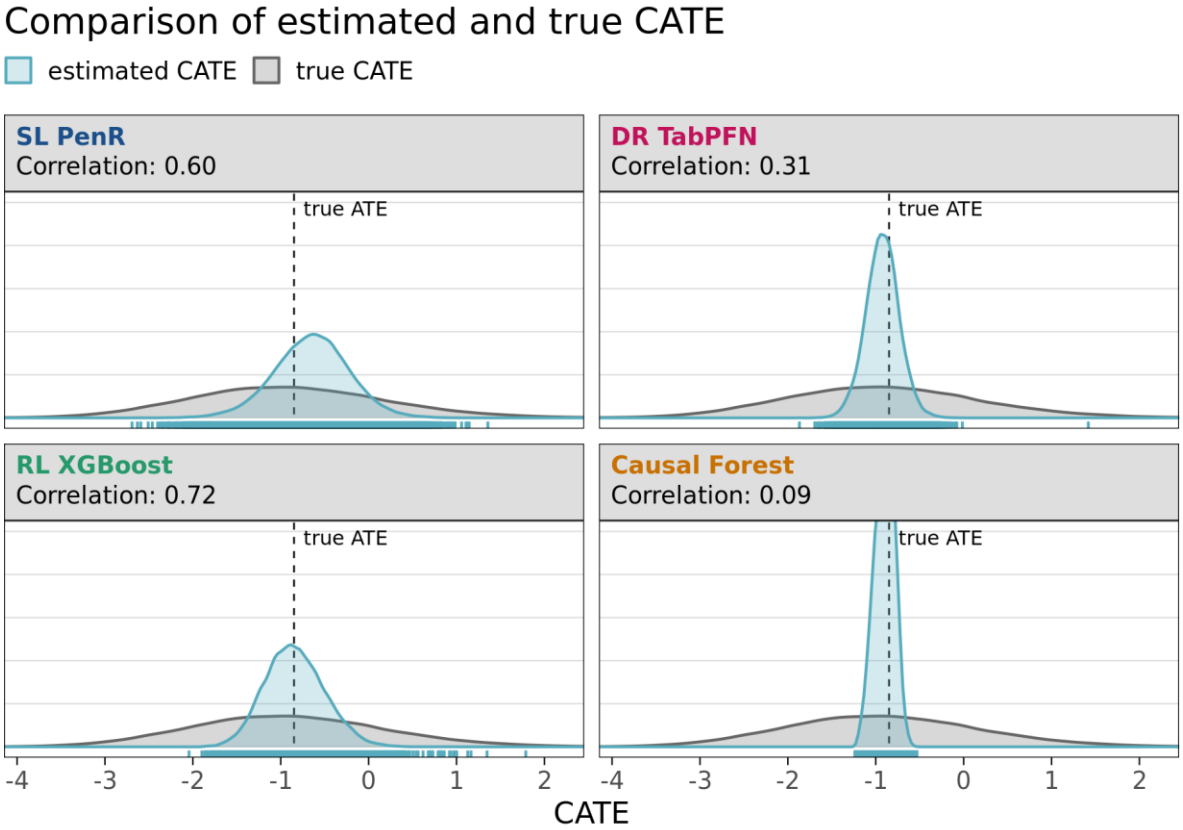


Method ranking
by performance

Spearman correlation
with true CATE

1. **SL PenR**
2. **DR TabPFN**
3. **RL XGBoost**
4. **Causal Forest**

simScenario2 - CATE performance

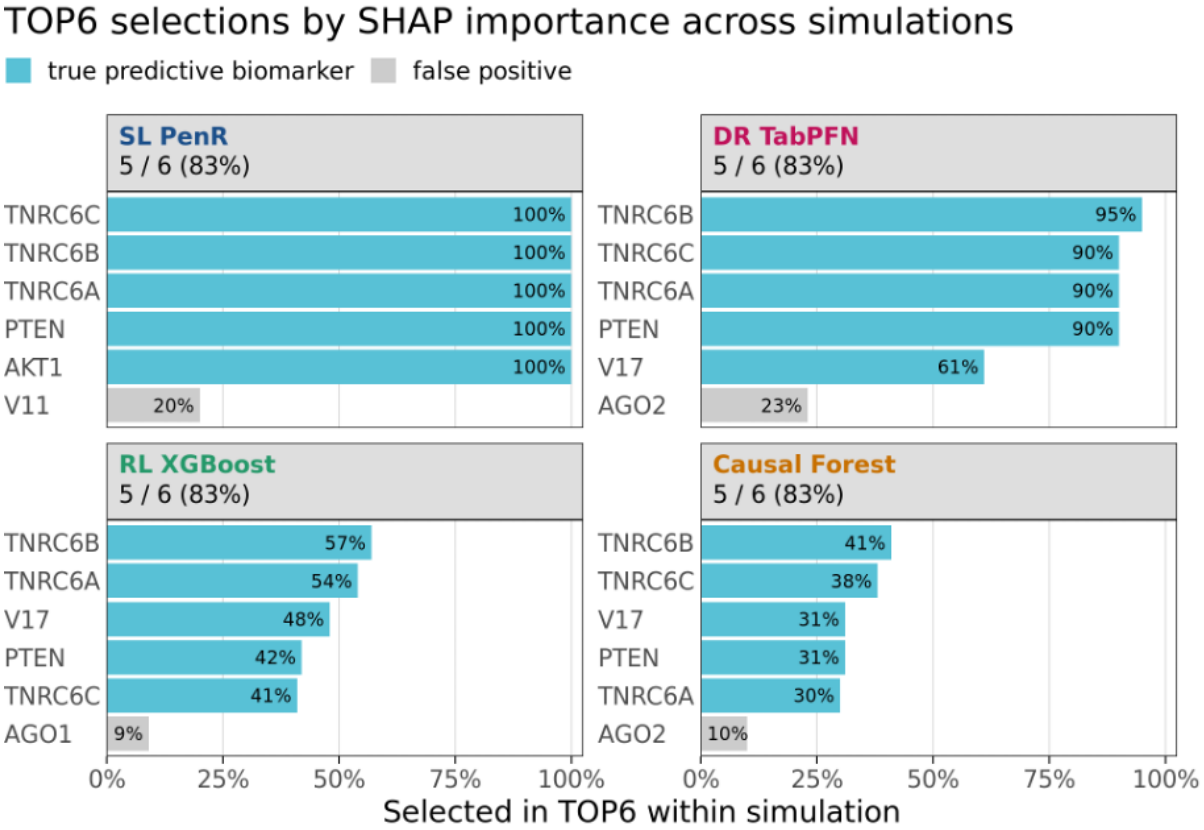


Method ranking
by performance

Spearman correlation
with true CATE

1. RL XGBoost
2. SL PenR
3. DR TabPFN
4. Causal Forest

simScenario1 – Marker detection accuracy

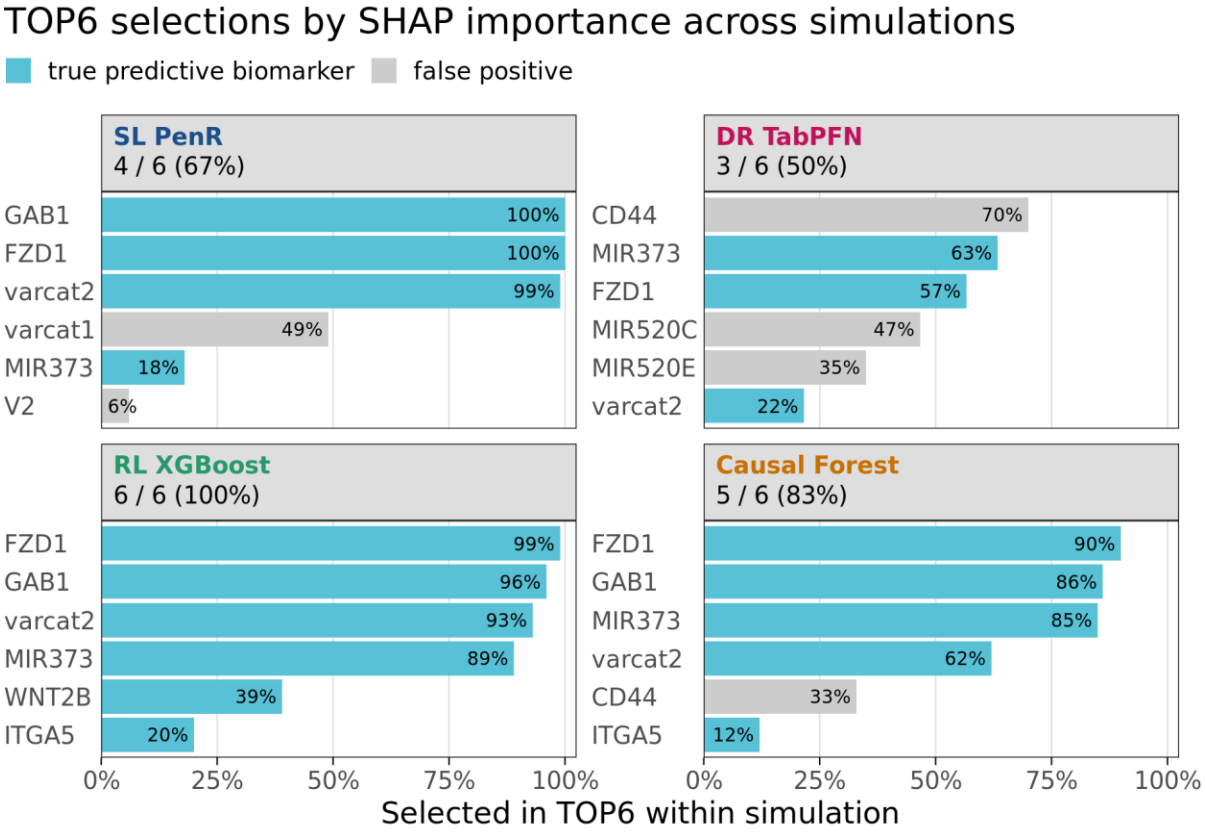


Method ranking
by detection accuracy

TOP6 selection in SHAP importance

1. SL PenR
2. DR TabPFN
3. RL XGBoost
4. Causal Forest

simScenario2 – Marker detection accuracy



Method ranking
by detection accuracy

TOP6 selection in SHAP importance

1. RL XGBoost
2. Causal Forest
3. SL PenR
4. DR TabPFN

Summary

- In **simScenario1** all methods detect 83% of true predictive BMs incl. quadratic effect, with most accurate CATE estimates for **SL PenR** and **DR TabPFN** (version 2.5)
- In high-dimensional **simScenario2, RL XGBoost** was the clear winner. **SL PenR** also performed well overall, but detected only linear effects
- In **simScenario2 Causal Forest** worked well to identify complex functional relationships, but CATE was estimated toward mean ATE for both scenarios

Conclusion

Limitation: simulations with 500 patients per scenario show only a snapshot of typical sample sizes across various trial phases and indications

Conclusions:

- **Investigated ML-based causal inference methods** performed overall quite well in analyzed scenarios, but need further investigation and – ideally – improvement
- **DR TabPFN** (version 2.5) did not outperform the other methods regarding CATE estimation and identification of true predictive variables

Outlook

- ✓ **TabPFN** with quasi-instant in-context learning is attractive for fast hypothesis testing with acceptable performance
- ✓ Foundation models can generalize **without task-specific training data**, whereas classical ML methods require training data for hyperparameter tuning.

What‘s next?

Explore whether **foundation models are competitive** with classical ML methods in a **clinical trial** setting, specifically with small sample sizes seen in Phase I/II

Contributors

Karl Köchert (Sanofi)
Nils Ternès (Sanofi)
Laura Schlieker (Staburo GmbH)
Hugo Hadjur (Saryga)
Ruben Sethi (Stratum Bio)
Eliana Garcia Cossio (Bayer AG)
Antigoni Elefsinioti (Bayer AG)
Adam Skubala (Bayer AG)
Maike Ahrens (Evidenze Germany GmbH)
Sebastian Voß (Evidenze Germany GmbH)

Thank you

schlieker@staburo.de
nils.ternes@sanofi.com
karl.koechert@sanofi.com

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

Data generation process

Simulation 1

```
y <- 0.1*X$trt +
  0.1*X$TNRC6A + 0.05*X$trt*X$TNRC6A + ###
  0.1*X$PTEN + 0.05*X$trt*X$PTEN +
  0.1*X$AKT1 + 0.05*X$trt*X$AKT1 +      ### main and interaction effects
  0.1*X$TNRC6B + 0.05*X$trt*X$TNRC6B +
  0.1*X$TNRC6C + 0.05*X$trt*X$TNRC6C +    ### below only main effects
  0.2*X$V11 +
  0.025*linkSin(X$V22) +
  0.005*linkX3(X$V23) +
  0.05*linkpredx2(x = X$V17,trt = X$trt) +
  noiseVector ## add error
```

Simulation 2

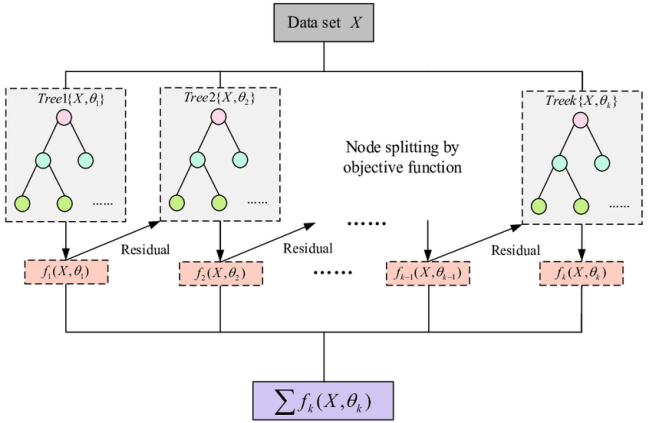
```
y <-  0.2*X$trt +
  0.2 * linkXprog(X$TGFA, fx = "x^2") +
  0.2 * linkXprog(X$HOXD10, fx = "x") +
  0.2 * linkXprog(X$MIR143, fx = "1/(1+exp(-x))") +
  0.2*X$V1 +
  0.01*linkXprog(X$V2, fx = "10 - 15*exp(-x)") +
  ifelse(X$varcat1 == "c1", yes = 0.2,
         no = ifelse(X$varcat1 == "c2", yes = -0.1, no = -0.5)) +
  0.5 * linkXpred(x = X$FZD1, trt = X$trt, fxy = "x", fxn = "0") +
  0.5 * linkXpred(x = X$GAB1, trt = X$trt, fxy = "x", fxn = "0") +
  0.5 * linkXpred(x = X$WNT2B, trt = X$trt, fxy = "x^2", fxn = "ifelse(abs(x) < 1.5, 0, 2)") +
  0.5 * linkXpred(x = X$MIR373,trt = X$trt, fxy = "1/(1+exp(-x))", fxn = "x") +
  0.5 * linkXpred(x = X$ITGA5 - 5, trt = X$trt, fxy = "x*(x>0)", fxn = "x") +
  ifelse(X$varcat2 == "c2" & X$trt == 1, yes = 1, no = 0) +
  activePrognostic_miRNA$genes +
  noiseVector
```

Brief overview of methods

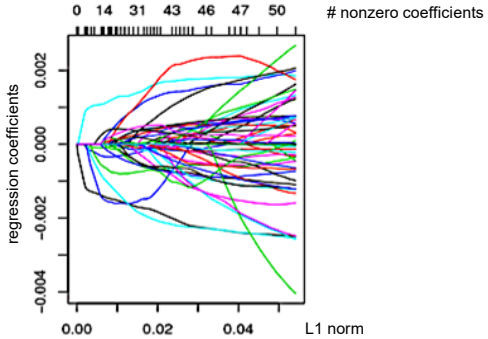
Method	Summary	Type	Variable Selection	HTE Assessment
Penalized Regression	Regularized linear model that shrinks or zeroes out coefficients to select relevant variables	Parametric / Linear	Native (L1 zero coefficients)	Limited (interaction terms needed manually)
XGBoost	Gradient boosting ensemble of trees that iteratively corrects residuals for high predictive accuracy	Non-parametric / Ensemble	Via feature importance (indirect)	Possible via meta-learners (S/T/R/DR)
TabPFN	Transformer-based in-context learner pre-trained on synthetic tabular datasets, performing inference without retraining	Non-parametric / Transformer	No native selection; attention weights as proxy	Possible via meta-learners but not natively causal
Causal Forest	Honest random forest adapted for causal inference, estimating heterogeneous treatment effects (HTE) at the individual level	Non-parametric / Ensemble + Causal	Via variable importance splitting (indirect)	Native — core purpose of the method

Brief overview of base learner methods

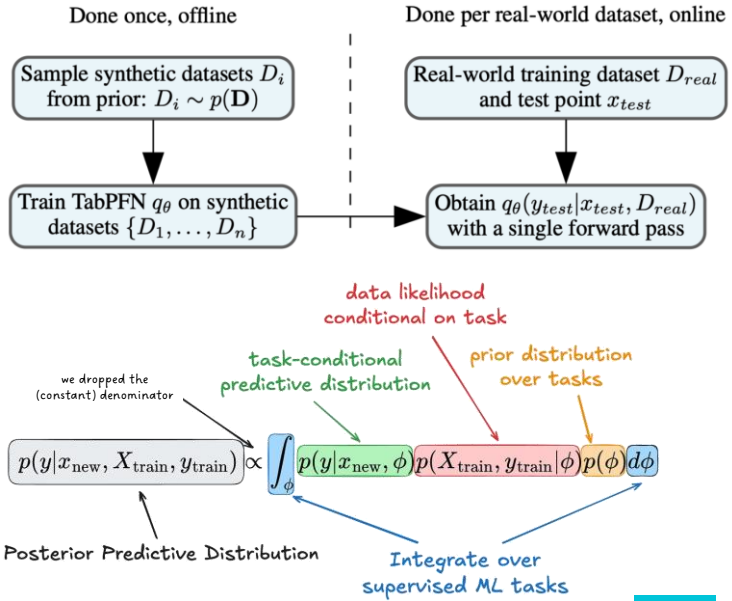
XGboost



Penalized regression



tabPFN



ML Approaches for CATE estimation

$CATE$ (conditional average treatment effect): $E[Y_i(1) - Y_i(0)|X_i = x] = E[\tau_i|X_i = x] = \tau(x)$

Decompose CATE Estimation

→ Meta-learners: ensemble of ML base learners

simple	complex
functions on the outcome (Y)	Nuisance functions & functions on pseudo–outcome (φ)
T-learner	X-learner
S-learner	DR-learner
	R-learner

Estimate CATE directly

- Causal Forests (generalized random forest - GRF)
- **Causal Boosting**

A Tutorial Introduction to Heterogeneous Treatment Effect Estimation with Meta-learners. Salditt et al., 2023.
Modern approaches for evaluating treatment effect heterogeneity from clinical trials and observational data. Lipkovich et al., 2024.

Meta-learners

S-Learner

1. Fit: $\hat{Y} = f(T, X)$
2. Predict two scenarios:
- $\hat{Y}(1, X)$ = prediction when T=1
- $\hat{Y}(0, X)$ = prediction when T=0
3. Estimate treatment effect:

$\hat{\tau}(X) = \hat{Y}(1, X) - \hat{Y}(0, X)$

R-Learner

1. Estimate outcome model:

$\hat{m}(X) = E[Y|X]$

2. Estimate treatment model:

$\hat{e}(X) = E[T|X]$

3. Compute residuals:
- $\tilde{Y} = Y - \hat{m}(X)$ (outcome residuals)
- $\tilde{T} = T - \hat{e}(X)$ (treatment residuals)

4. Fit: $\hat{\tau}(X) = g(\tilde{T}, X)$ on residuals
5. Estimate treatment effect from this model

DR-Learner

1. Estimate outcome model:

$\hat{m}(X) = E[Y|X]$

2. Estimate treatment propensity:

$\hat{e}(X) = E[T|X]$

3. Compute doubly robust residuals:
- $\tilde{Y}\_DR = Y - \hat{m}(X) + (T - \hat{e}(X)) \times (Y - \hat{m}(X)) / (\hat{e}(X)(1-\hat{e}(X)))$
- $\tilde{T}\_DR = T - \hat{e}(X)$

4. Fit: $\hat{\tau}(X) = g(\tilde{T}\_DR, \tilde{Y}\_DR, X)$
5. Estimate treatment effect

Key reference: Salditt, M., Eckes, T. & Nestler, S. A Tutorial Introduction to Heterogeneous Treatment Effect Estimation with Meta-learners. *Adm Policy Ment Health* **51**, 650–673 (2024). https://doi.org/10.1007/s10488-023-01303-9

CATE Estimation

R-learner | Minimizing R-loss

Outcome Model

$\mu(X) =$ Y ~ X → $\hat{\mu}$

Propensity Score

$\pi(x) = trt \sim X \rightarrow \hat{\pi}$

Pseudo-Outcome (based on residuals):

$$\psi_{RL}(X_i) = {(Y_i - \hat{\mu}(X_i))}/{(trt_i - \hat{\pi})}$$

Model:

$\tau(X) = \psi_{RL} \sim X$ with weights $w = (trt_i - \hat{\pi})^2$

GRF | causal random forest repurposed as an adaptive nearest neighbor finder to detect TEH

- Combines multiple trees to reduce variance in estimates
- Creates partitions to maximize treatment effect hetereogenity
- Uses separate samples for tree structure and leaf estimates (honest splitting)
- uses nuisance functions to estimate the conditional mean of the outcome and treatment assignment.

Direct estimation using the forest structure.

CATE based on Augmented Inverse Probability Weighted (AIPW) estimator (doubly robust)

$R - Loss\ (doubly\ robust)$:

$$L_R[\tau] = \frac{1}{n}\sum_{i=1}^{n}(trt_i - \hat{\pi})^2\,[\psi_{RL}(X_i) - \tau(X_i)]^2$$

CATE $\hat{\tau}$

Background: TabPFN

2.2 Design — 3-Fold DR Cross-Fitting

Folds: A, B, C (k=3, stratified by treatment W).

For each holdout fold H (3 iterations), 2 rounds swap nuisance ↔ CATE roles:

Round	Nuisance trains on	Pseudo built on	CATE trains on	Predicts on
1	F1	F2	F2	H
2	F2	F1	F1	H

Total: 3 holdouts × 2 rounds = 6 cross-fitting combinations. All three roles (nuisance-train, pseudo/CATE-train, holdout) use **distinct folds** in every round.

Nuisance fold fits per round (3 TabPFN models):

- $\hat{\pi}(X)$: propensity model (TabPFNClassifier) on all patients in nuisance fold → predict for pseudo fold
- $\hat{\mu}_1(X)$: outcome model (TabPFNClassifier/Regressor) on **treated** patients only → predict for pseudo fold
- $\hat{\mu}_0(X)$: outcome model on **control** patients only → predict for pseudo fold

DR pseudo-outcome formula (Kennedy 2020):

$$\hat{\psi}_i = \hat{\mu}_1(x_i) - \hat{\mu}_0(x_i) + \frac{A_i\big(Y_i - \hat{\mu}_1(x_i)\big)}{\hat{\pi}(x_i)} - \frac{(1 - A_i)\big(Y_i - \hat{\mu}_0(x_i)\big)}{1 - \hat{\pi}(x_i)}$$

Double robust: consistent if **either** $(\hat{\mu}_1, \hat{\mu}_0)$ **or** $\hat{\pi}$ is correctly specified.

Background: SHAP for TabPFN

- SHAP estimation similar to the Svensson et al. 2025 approach:
  - For each dataset take predicted CATE and fit another tabPFN model to estimate SHAP.
  - No Xgboost as in the paper but the idea of using the predicted CATE as depended variable to estimate SHAP was applied
  - https://arxiv.org/abs/2505.01145
- TabPFN not trained for more than 2000 features, this comes with risk of performance

Background: XGBoost hyperparameter tuning

- Using the 2025 hyperparameters, the outcome and propensity models in the RIearner yielded good R-squared values
- For the CATE model part of the RIearner, the 2025 parameters produced some overfitting:
  - Based on sensitivity analysis was run, selected the set of hyperparameters that maximizes the average of the 5 R-squared values from the 5 first datasets of scenario 1 (R-squared of predicted vs actual CATE values).
  - Iterated once and refined the grid (manually).