StaDISC: Stable Discovery of Interpretable Subgroups via Calibration

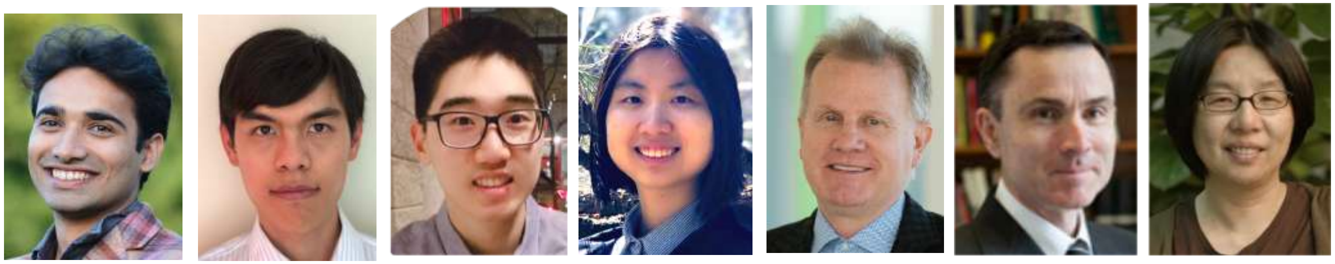
Raaz Dwivedi
EECS, UC Berkeley

ETH Young Data Science Researcher Seminar Zurich

September 25, 2020

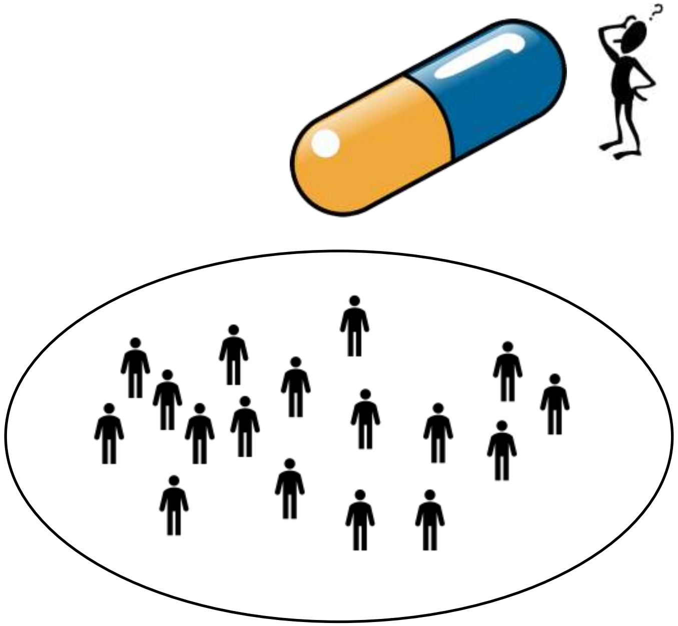
A collaborative project

- Raaz Dwivedi, Yan Shuo Tan, Briton Park, Mian Wei, Kevin Horgan, David Madigan, Bin Yu



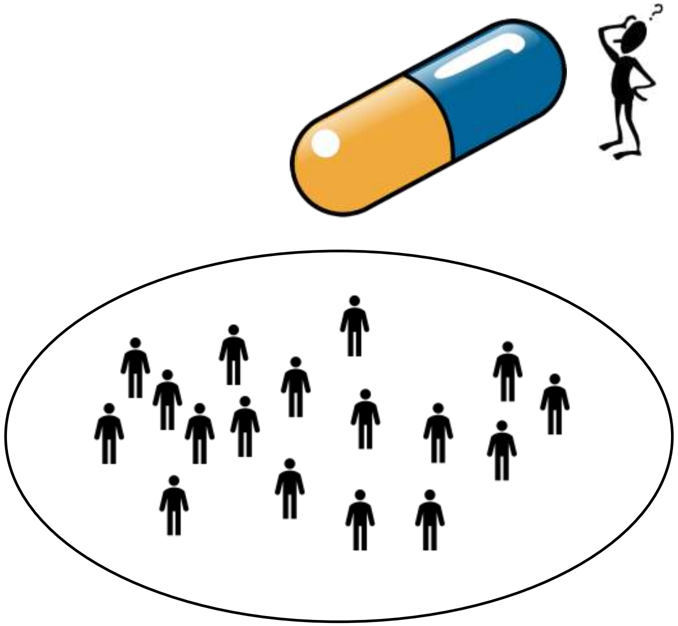
Stable discovery of interpretable subgroups via calibration in causal studies.
Accepted at International Statistical Review
Preprint available at arXiv:2008.10109

Heterogeneous treatment effects (HTE)



$Y_i(1)$	$Y_i(0)$
2	?
?	5
6	?
?	5
3	?
?	2
…	…

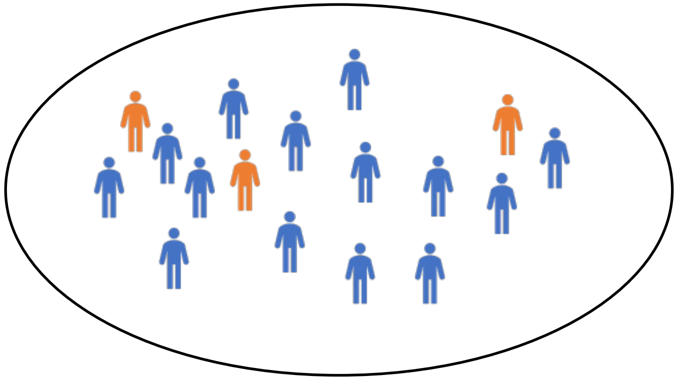
Heterogeneous treatment effects (HTE)



$Y_i(1)$	$Y_i(0)$
2	?
?	5
6	?
?	5
3	?
?	2
…	…

$\hat{\tau}_{ATE} = 0.3$ 95% CI: (-0.1,0.7) ☹

Heterogeneous treatment effects (HTE)



$Y_i(1)$	$Y_i(0)$
2	?
?	5
6	?
?	5
3	?
?	2
…	…

$\mathcal{G}$

$\hat{\tau}_{\mathcal{G}} = -1.7$ 95% CI: (-2.3,-1.1) ☺

Heterogeneous treatment effects (HTE)

- The treatment effect of drugs, public policies, advertisements, are often heterogeneous

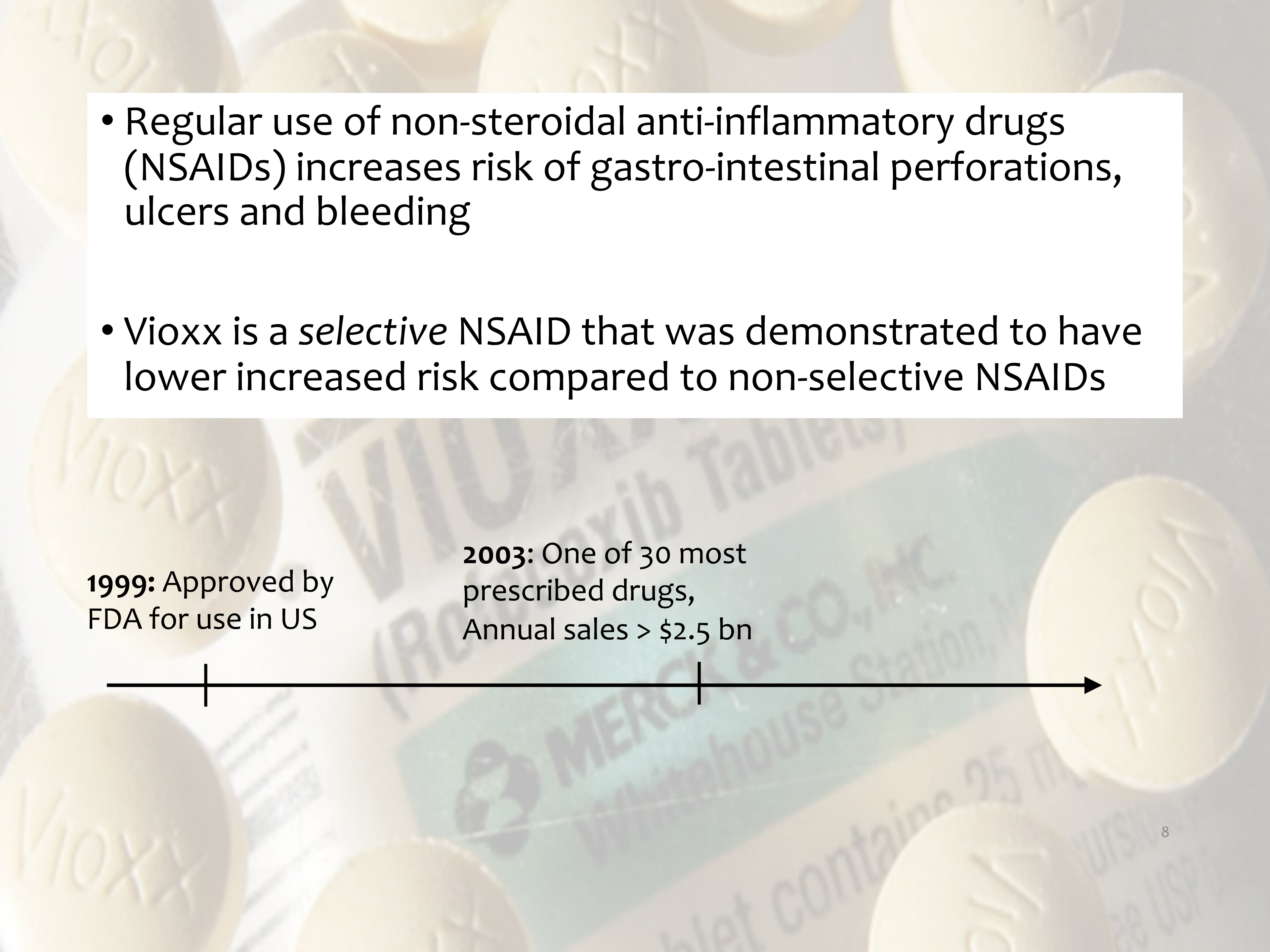
- Being able to identify a subgroup that benefits/is harmed disproportionately allows us to **target interventions**

- This work addresses HTE in **randomized experiments**

Vioxx
VIOXX® 25 mg
(Rofecoxib Tablets)
MERCK & CO., INC.
Whitehouse Station, N
blet contains 25 m
ursu
ee USP

- Regular use of non-steroidal anti-inflammatory drugs (NSAIDs) increases risk of gastro-intestinal perforations, ulcers and bleeding

- Vioxx is a *selective* NSAID that was demonstrated to have lower increased risk compared to non-selective NSAIDs

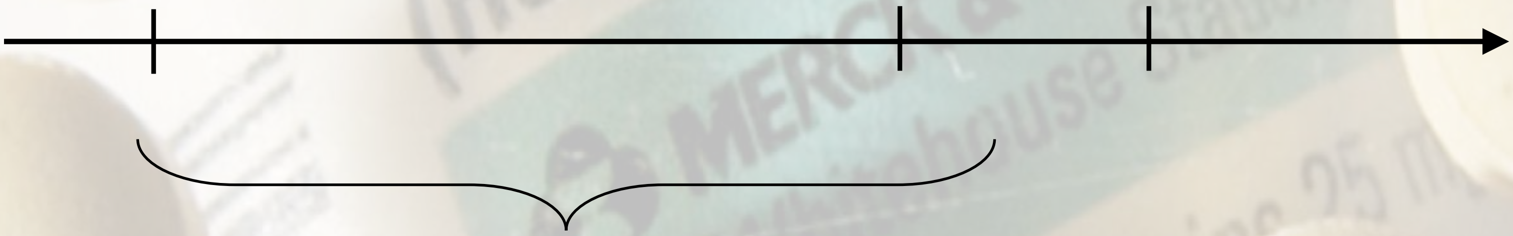


1999: Approved by FDA for use in US

2003: One of 30 most prescribed drugs, Annual sales > $2.5 bn

- Regular use of non-steroidal anti-inflammatory drugs (NSAIDs) increases risk of gastro-intestinal perforations, ulcers and bleeding

- Vioxx is a *selective* NSAID that was demonstrated to have lower increased risk compared to non-selective NSAIDs



1999: Approved by FDA for use in US

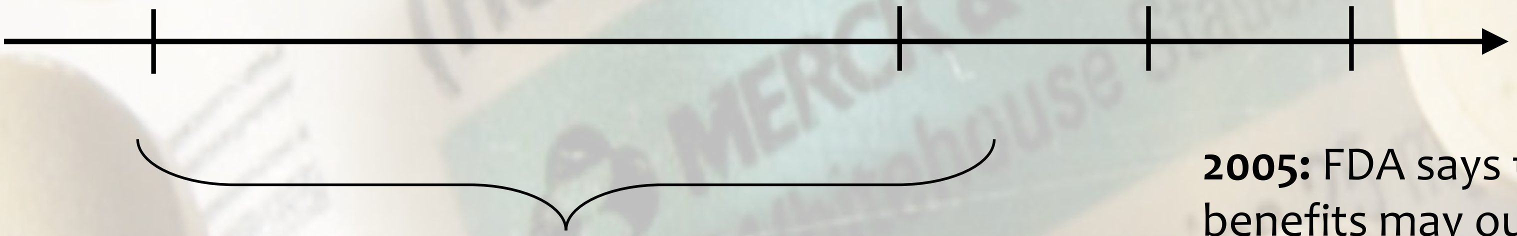
2003: One of 30 most prescribed drugs, Annual sales > $2.5 bn

2004: Merck withdraws Vioxx from market

2001-2004: Study found that Vioxx increased the risk of thrombotic cardiovascular events

- Regular use of non-steroidal anti-inflammatory drugs (NSAIDs) increases risk of gastro-intestinal perforations, ulcers and bleeding

- Vioxx is a *selective* NSAID that was demonstrated to have lower increased risk compared to non-selective NSAIDs



1999: Approved by FDA for use in US

2003: One of 30 most prescribed drugs, Annual sales > $2.5 bn

2004: Merck withdraws Vioxx from market

2001-2004: Study found that Vioxx increased the risk of thrombotic cardiovascular events

2005: FDA says that benefits may outweigh risks, may return to market

The VIGOR study:
VIoxx GI Outcomes Research

- 1999-2000 Randomized controlled by Merck with a **8076 patients** who had rheumatoid arthritis

- Treatment arm: **Vioxx** vs Control arm: **Naproxen**

Outcome	ATE	Base rate
Gastro-intestinal (GI) event	-1.6%	2.2%
Thrombotic cardiovascular (TC) event	0.6%	0.7%

Bombardier et al.. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group. *New England Journal of Medicine*, 343(21):1520–1528, 2000

The VIGOR Study

- Authors also found:
  - Relative risk for GI event of 0.5 with 95% CI (0.3, 0.6)
  - On 14 pre-identified subgroups, relative risk not significantly different
- Most patients (98%) did not have substantial protocol violations
- For simplicity:
  - We will ignore compliance and time-to-event
  - We consider treatment efficacy in terms of ATE (rather than relative risk)

Bombardier et al.. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group. *New England Journal of Medicine*, 343(21):1520–1528, 2000

Research questions

Can we find subgroups of patients for which Vioxx’s effects are disproportionate for the two outcomes?

How do we validate our findings?

Neyman-Rubin framework

- Assume a superpopulation $\left(X_i, T_i, Y_i(1), Y_i(0)\right) \sim_{i.i.d.} \mathbb{P}$
- Randomized experiment:
  - $Y_i(T_i), X_i | T_i = a$ has same distribution as $(Y_i(a), X_i)$ for $a = 0{,}1$
- ATE: $\tau_{ATE} = \mathbb{E}_{\mathbb{P}}[Y_i(1) - Y_i(0)]$
- Conditional Average Treatment Effect (CATE):
  - $\tau(x) \coloneqq \mathbb{E}[Y_i(1) - Y_i(0) | X = x]$
- Subgroup CATE: Given a subgroup $\mathcal{G} \subset \mathcal{X}$
  - $\tau_{\mathcal{G}} \coloneqq \mathbb{E}[Y_i(1) - Y_i(0) | X \in \mathcal{G}] = \mathbb{E}[\tau(X) | X \in \mathcal{G}]$
- Goal: Find **interpretable** $\mathcal{G}$ for which $\tau_{\mathcal{G}}$ is **larger** than τ_{ATE}.

How to estimate the HTE?
Subgroup Analysis

- Compute subgroup CATE on a pre-determined list of subgroups

- Ignores potential heterogeneity

- Naive subgroup search: Combinatorial explosion of number of possible subgroups

... Byar ’85, Dixon-Simon ’91,
Assmann et al. ’00, Peck ’03, Imbens-Wooldridge ’09,
Lipkovich et al. ’11, Athey -Imbens ’16 …

How to estimate the HTE?
CATE modeling

- Estimate $\hat{\tau}(x)$ from samples, use $\hat{\tau}(x)$ to identify subgroups

- How to estimate CATE (non-parametrically)
  - Metalearner framework
    - T-learner [Foster et al. ’11, Imai-Ratkovic ’13, Bloniarz et al. ’16..]
    - X-learner [Kunzel et al. ’19]
    - R-learner [Nie-Wager ’20]
  - Tree-based methods
    - Causal tree [Athey-Imbens ’16]
    - Causal forest [Wager-Athey ’18]
    - BART [Hill ’12]

Problems with CATE modeling

- Numerous modeling choices
  - Meta-learner, base learner, hyperparameters
- Model validation hard due to missing data
  - Existing schemes: Proxy loss functions
  - Require uncheckable assumptions for theoretical guarantees
  - Do not have easily interpretable scale (like R^2 or ROC AUC)
- In VIGOR: Poor signal because of event rarity
  - 2.2% for GI, 0.7% for TC

Schuler et al. ’18, Ross et al. ’09, Carini et al. ’14, Alaa-van der Schaar ’19

PCS Framework

Towards bridging the two cultures: Statistics and Machine Learning

Bin Yu and Karl Kumbier. Veridical data science. *PNAS*, 117(8):3920–3929, 2020

PCS framework

Three principles of data science : PCS

Predictability (P) (from ML)

Computability (C) (from ML)

Stability (S) (from statistics)

PCS **bridges** the two cultures:
Statistics and *machine learning*,
unifies and **expands** on their ideas

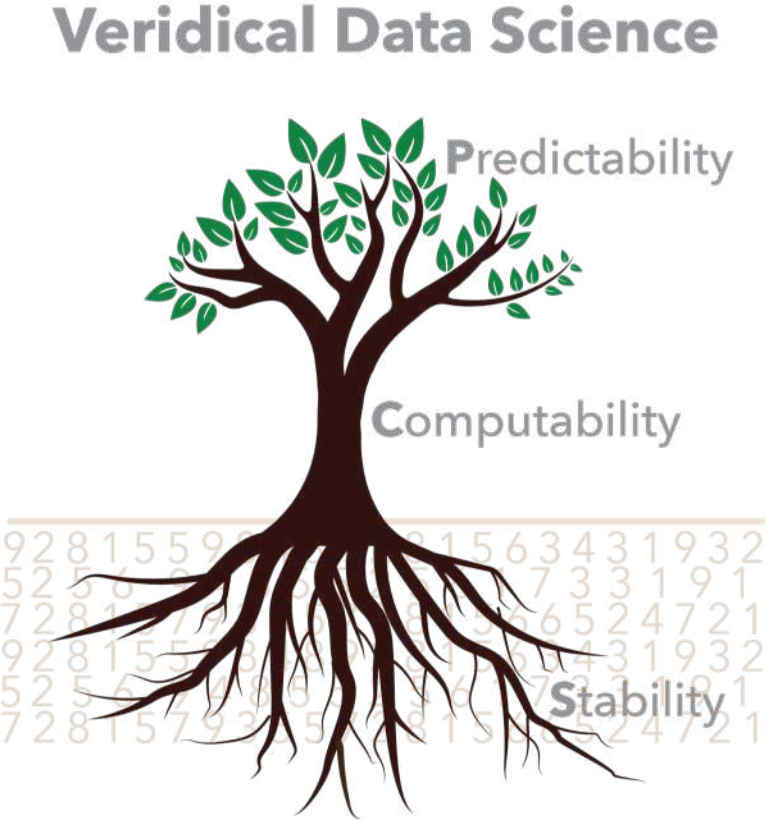


Image credit: R. Barter

Predictability for reality check
Stability tests DSLC by “shaking” every part

Shakes come from human decisions

DSLC

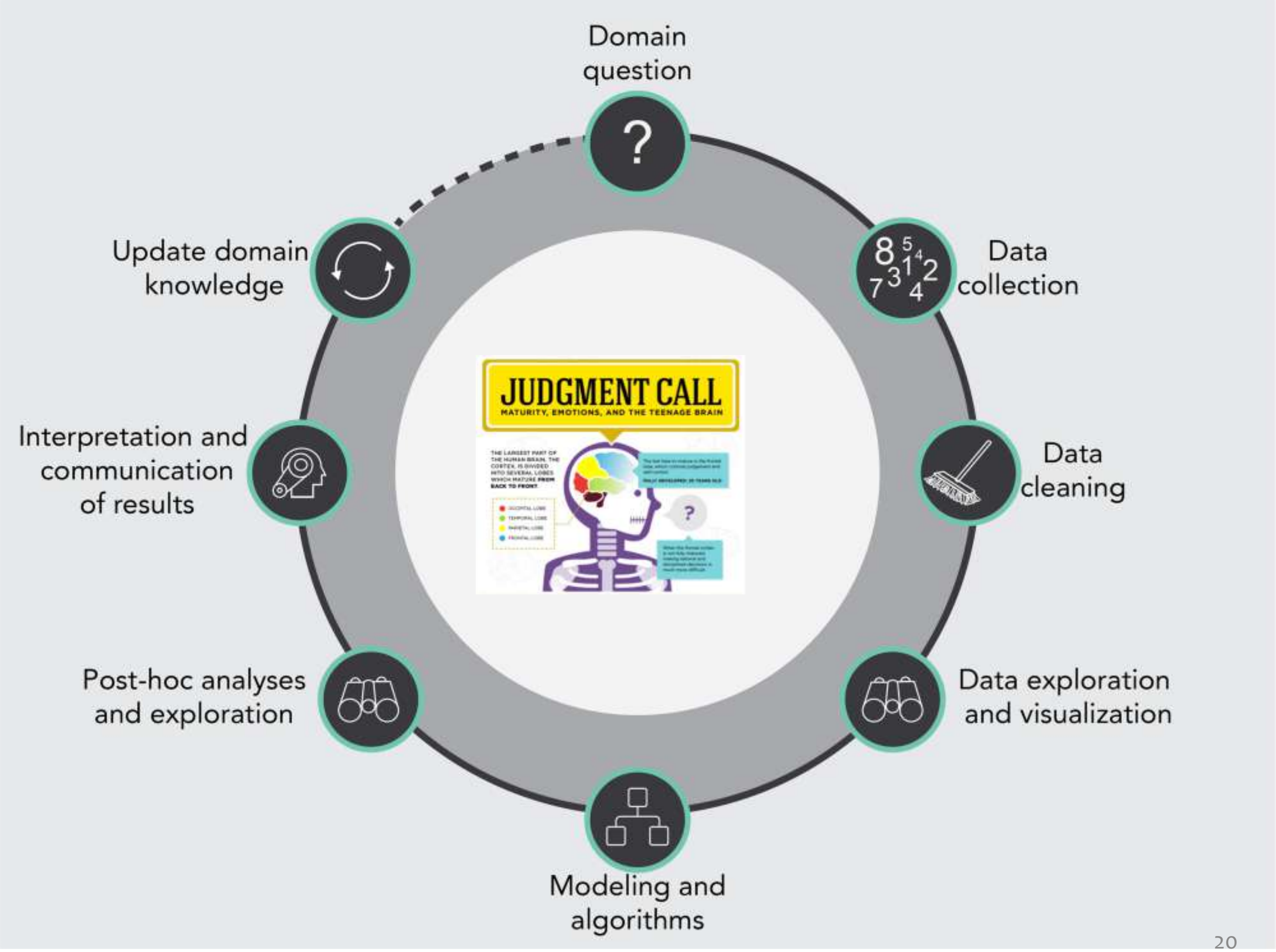
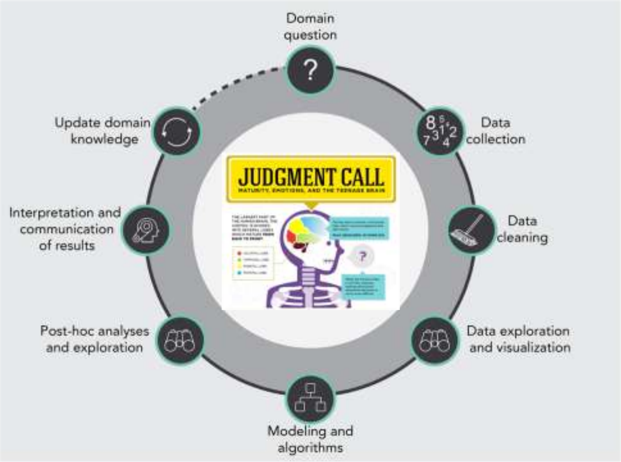

Domain question
Data collection
Data cleaning
Data exploration and visualization
Modeling and algorithms
Post-hoc analyses and exploration
Interpretation and communication of results
Update domain knowledge
JUDGMENT CALL
MATURITY, EMOTIONS, AND THE TEENAGE BRAIN

Image credits: R. Barter and toronto4kids.com

PCS workflow

- Workflow incorporates P, C, S into each step of the DSLC


Domain question
Data collection
Data cleaning
Data exploration and visualization
Modeling and algorithms
Post-hoc analyses and exploration
Interpretation and communication of results
Update domain knowledge
JUDGMENT CALL
MATURITY, EMOTIONS, AND THE TEENAGE BRAIN

- In particular, basic PCS inference applies PCS through data and model perturbations at the modeling stage (with P as a first screening step before perturbation intervals are made)

Image credits: R. Barter and toronto4kids.com

Contributions

1. Extend PCS framework from supervised learning to causal studies

2. Introduce calibration-based predictive checks for CATE models

3. Overall, develop staDISC methodology for using CATE models to find interpretable subgroups

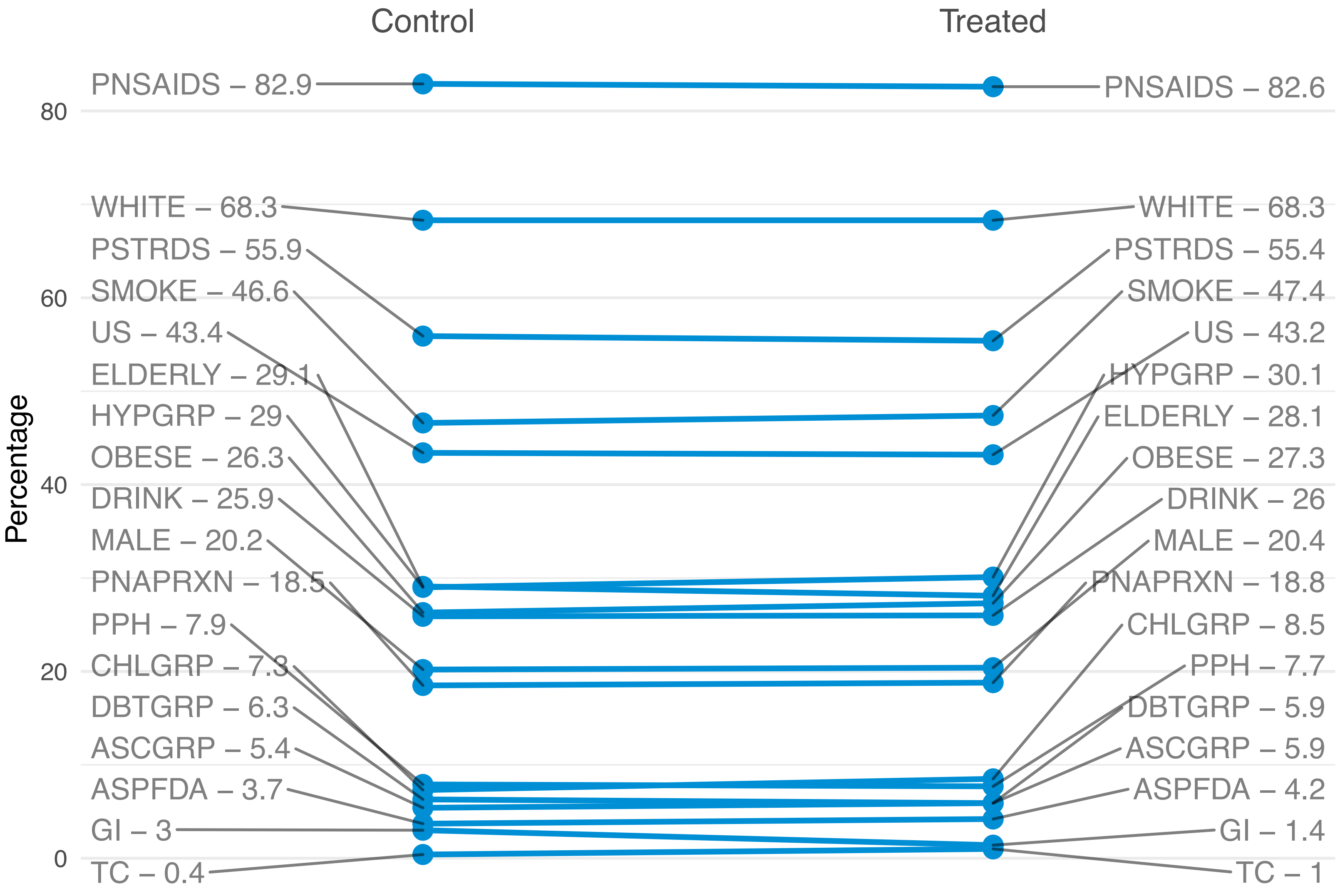
4. Case study with VIGOR, and external validation with APPROVe study

Feature engineering

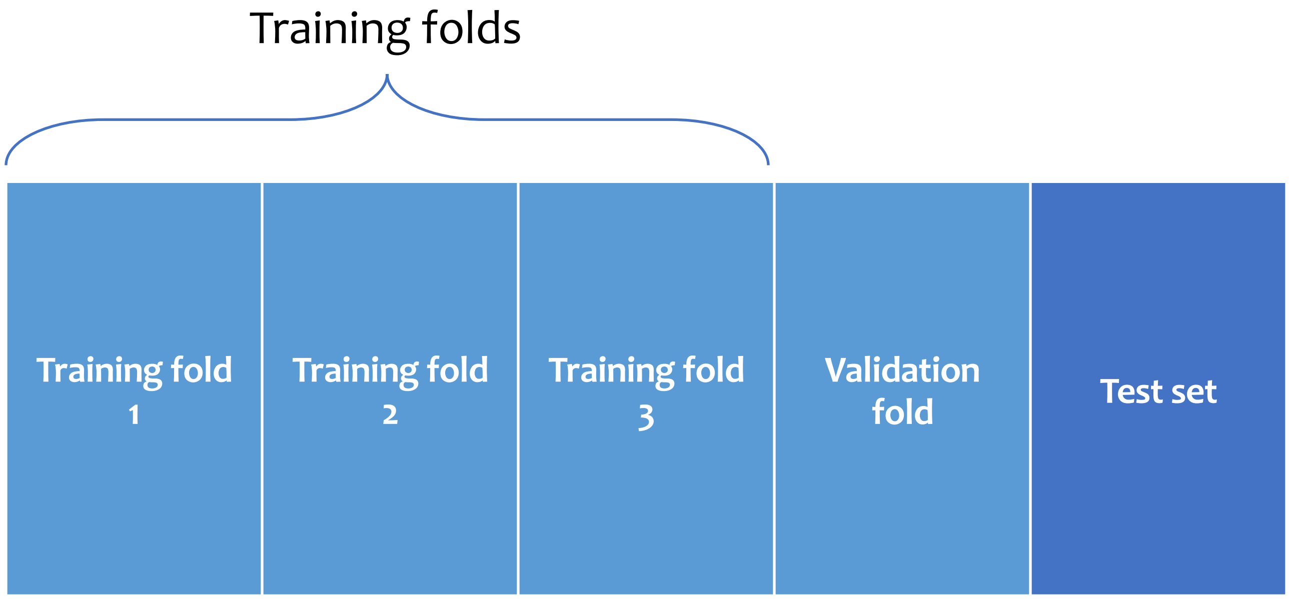
16 binary features

- Demographics (5):
  - Gender, race, country, elderly, obese
- Lifestyle risk factors (2):
  - Smoking, drinking
- Medical risk factors (9):
  - Medical history (e.g. prior history of GI event, hypertension, ..)
  - Use of other medication (e.g. use of glucorticoids/steroids, ..)

Covariate Balance in the Dataset



Data splitting



Training folds

Training fold 1

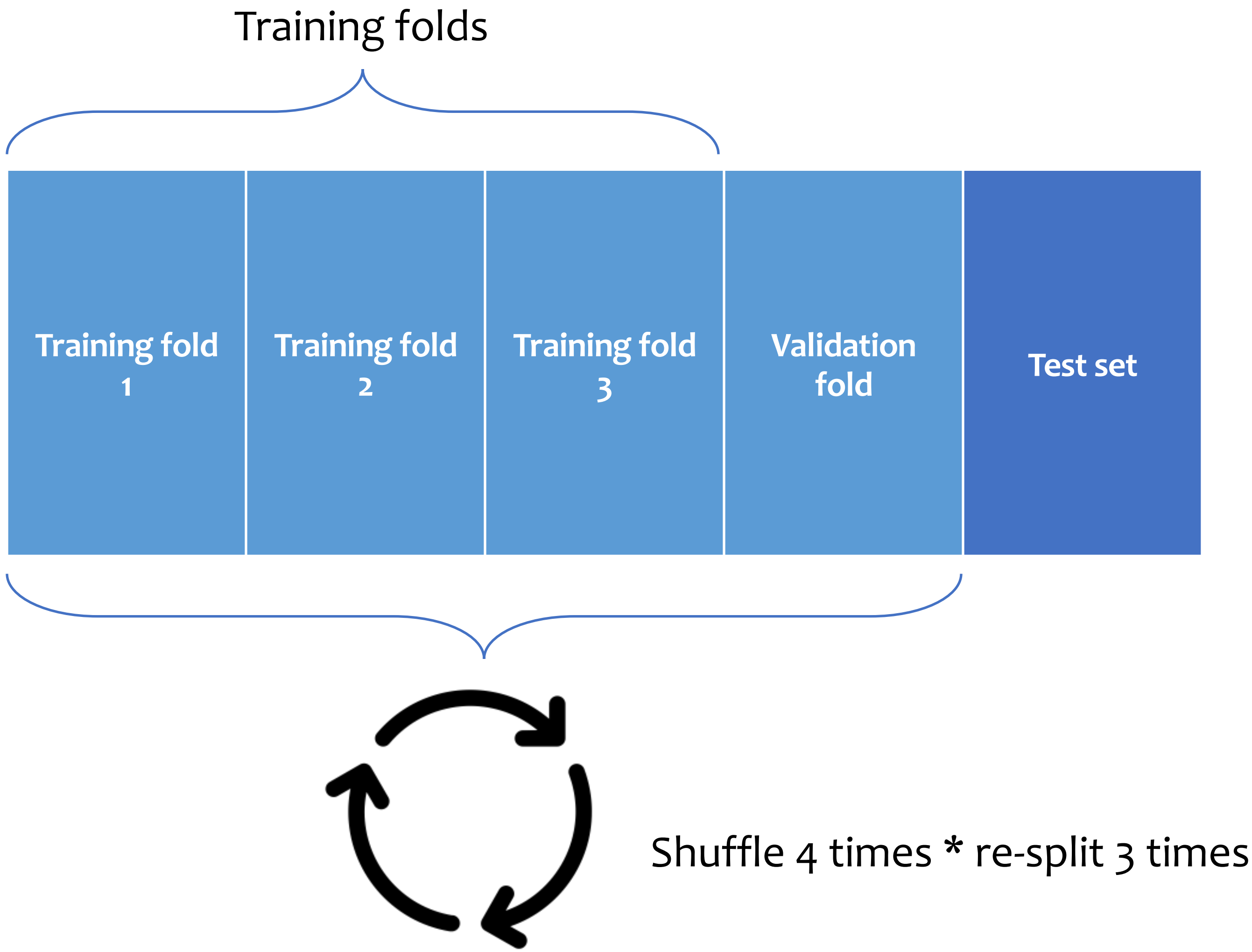
Training fold 2

Training fold 3

Validation fold

Test set

Data splitting



Training folds

Training fold 1

Training fold 2

Training fold 3

Validation fold

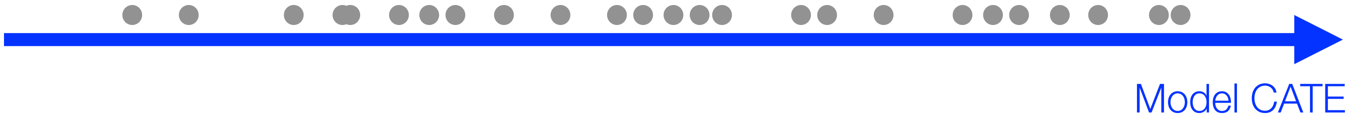
Test set

Shuffle 4 times * re-split 3 times

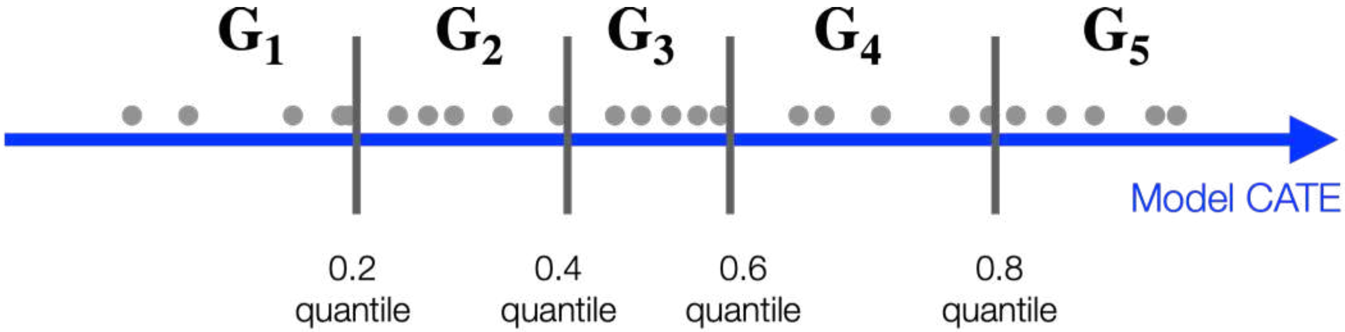
18 CATE models

- S learners
  - Random Forest, XGBoost
- T learners
  - Random Forest, XGBoost, Lasso, Logistic
- X learners
  - Outcome learner: Random Forest, XGBoost, Lasso, Logistic
  - Cross learner: Lasso
- R learners
  - {Lasso, Lasso}, {Lasso, XGB}, {RF, Lasso}, {RF, RF}
- Causal Tree
  - 2 hyperparameters
- Causal Forest
  - 2 hyperparameters

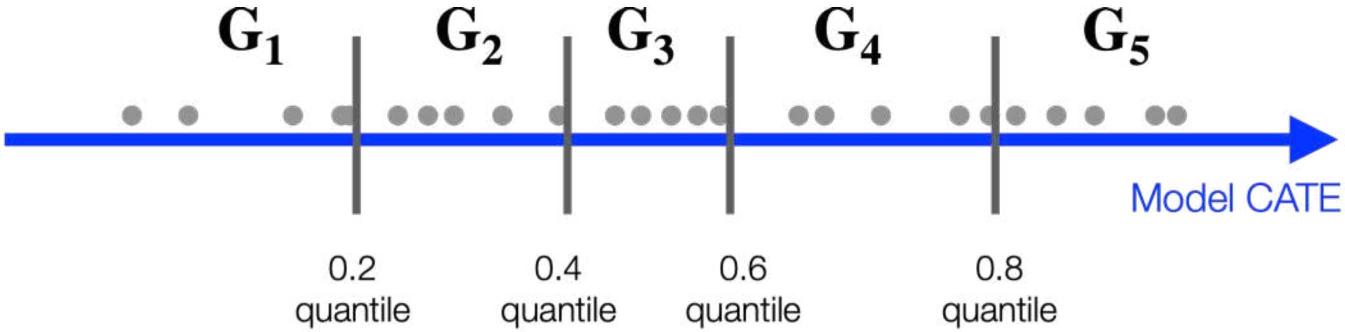
CATE modeling: Prediction check?



Prediction check via calibration



Prediction check via calibration

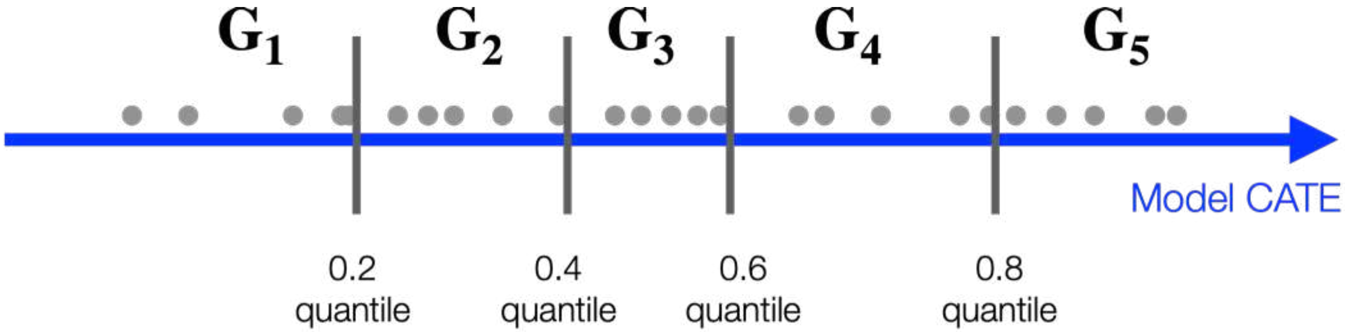


- Rich history in supervised learning for validating estimated probabilities from models for data with deterministic outcomes

- First use in weather forecasting (?!), and more recently for calibrating modern ML methods including NNs
  [Brier '50, Miller '62, Murphy '73, Dawid '82, DeGroot and Fienberg '83, …, Niculescu et al. '05, Naeini '15, Guo et al. '17, ..]

- We introduce it to causal settings but we need some proxy for "true labels"

Prediction check via calibration

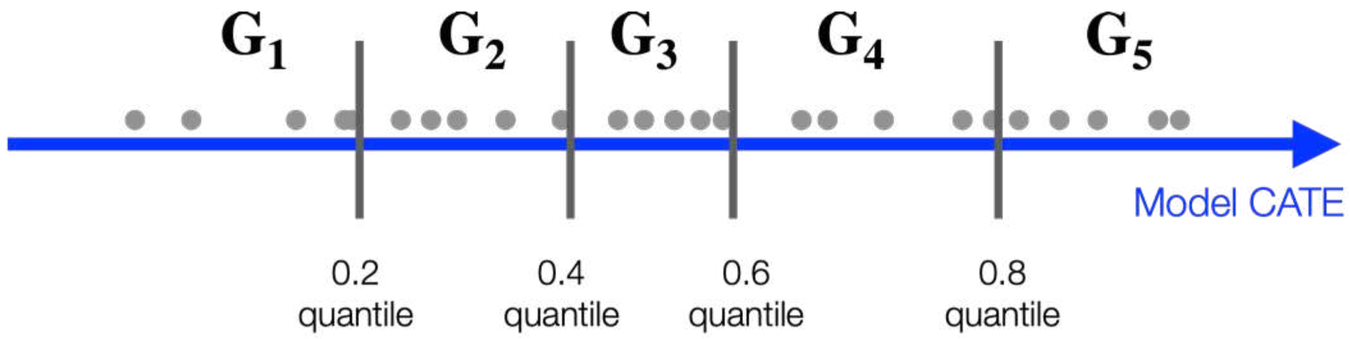


Model estimate of bin CATE

$$\overline{\mathbf{M}}_{\mathbf{G}_j \cap \mathbf{S}} := \frac{1}{|\mathbf{G}_j \cap \mathbf{S}|} \sum_{i \in \mathbf{G}_j \cap \mathbf{S}} \mathbf{M}(X_i)$$

S denotes training or validation folds.

Prediction check via calibration



Model estimate of bin CATE

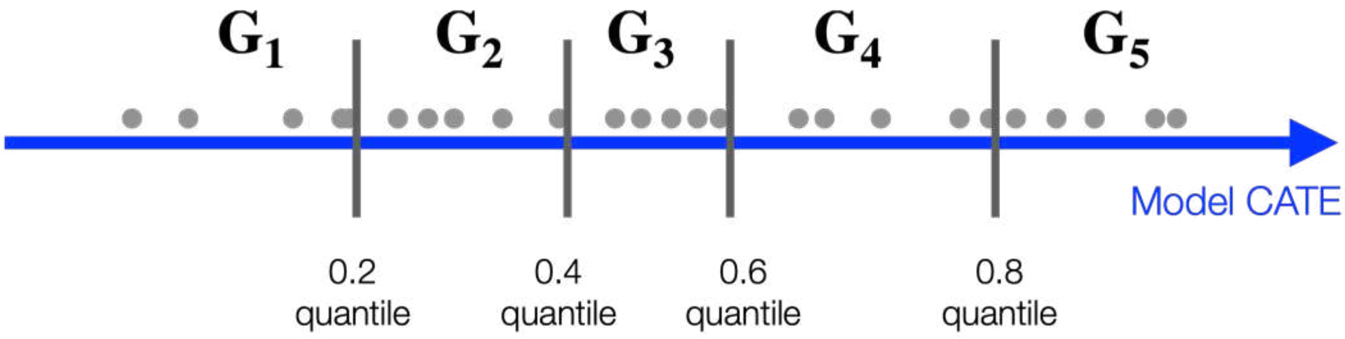
$$\overline{\mathbf{M}}_{\mathbf{G}_j \cap \mathbf{S}} := \frac{1}{|\mathbf{G}_j \cap \mathbf{S}|} \sum_{i \in \mathbf{G}_j \cap \mathbf{S}} \mathbf{M}(X_i)$$

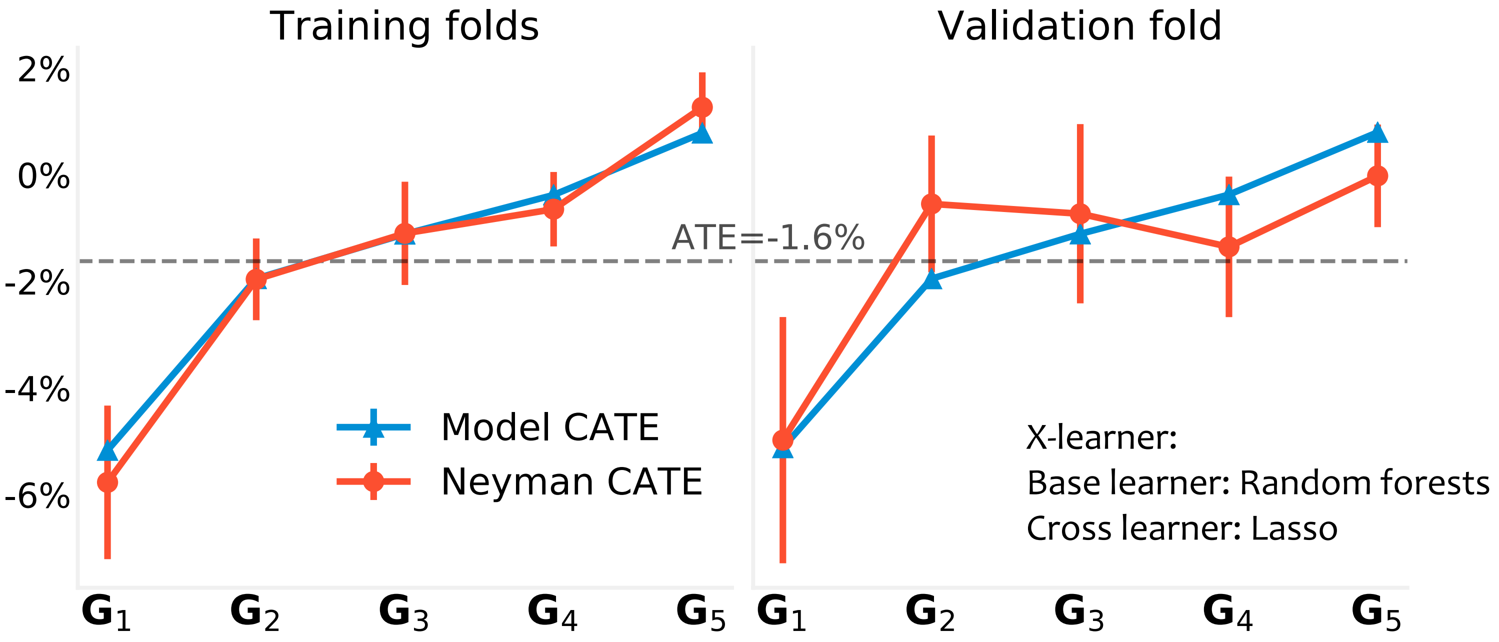
Neyman estimate of bin CATE

$$\widehat{\tau}_{\mathbf{G}_j \cap \mathbf{S}} := \frac{1}{|\mathbf{T} \cap \mathbf{G}_j \cap \mathbf{S}|} \sum_{i \in \mathbf{T} \cap \mathbf{G}_j \cap \mathbf{S}} Y_i(1) - \frac{1}{|\mathbf{C} \cap \mathbf{G}_j \cap \mathbf{S}|} \sum_{i \in \mathbf{C} \cap \mathbf{G}_j \cap \mathbf{S}} Y_i(0)$$

S denotes training or validation folds.

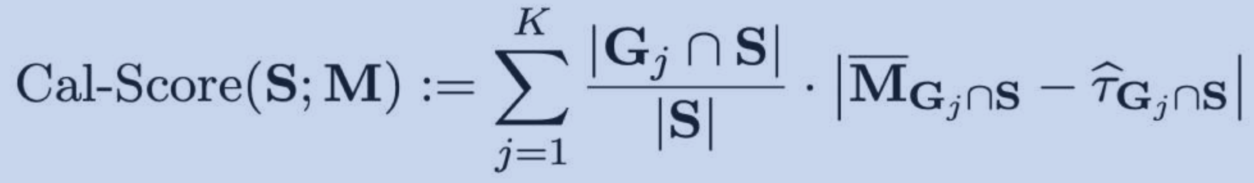
Prediction check via calibration:
Visual Assessment





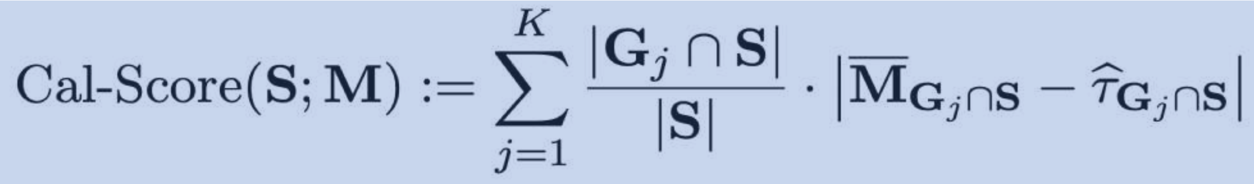
X-learner:
Base learner: Random forests
Cross learner: Lasso

Prediction check via calibration: *Quantitative assessment*

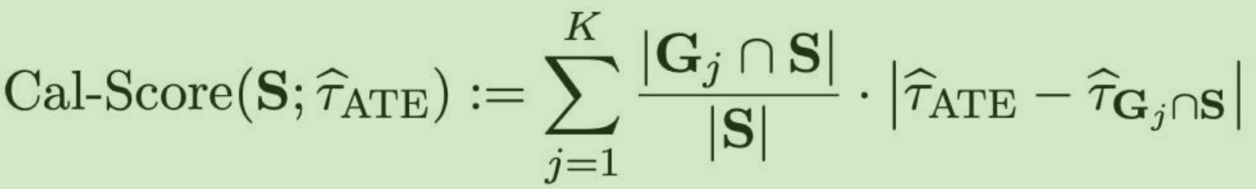


$$\text{Cal-Score}(\mathbf{S};\mathbf{M}) := \sum_{j=1}^{K} \frac{|\mathbf{G}_j \cap \mathbf{S}|}{|\mathbf{S}|} \cdot \left|\overline{\mathbf{M}}_{\mathbf{G}_j \cap \mathbf{S}} - \widehat{\tau}_{\mathbf{G}_j \cap \mathbf{S}}\right|$$

Prediction check via calibration:
Quantitative assessment



$$\text{Cal-Score}(\mathbf{S};\mathbf{M}) := \sum_{j=1}^{K} \frac{|\mathbf{G}_j \cap \mathbf{S}|}{|\mathbf{S}|} \cdot \left|\overline{\mathbf{M}}_{\mathbf{G}_j \cap \mathbf{S}} - \widehat{\tau}_{\mathbf{G}_j \cap \mathbf{S}}\right|$$



$$\text{Cal-Score}(\mathbf{S};\widehat{\tau}_{\text{ATE}}) := \sum_{j=1}^{K} \frac{|\mathbf{G}_j \cap \mathbf{S}|}{|\mathbf{S}|} \cdot \left|\widehat{\tau}_{\text{ATE}} - \widehat{\tau}_{\mathbf{G}_j \cap \mathbf{S}}\right|$$

Prediction check via calibration: *Quantitative assessment*

$$\text{Cal-Score}(\mathbf{S};\mathbf{M}) := \sum_{j=1}^{K} \frac{|\mathbf{G}_j \cap \mathbf{S}|}{|\mathbf{S}|} \cdot \left|\overline{\mathbf{M}}_{\mathbf{G}_j \cap \mathbf{S}} - \widehat{\tau}_{\mathbf{G}_j \cap \mathbf{S}}\right|$$

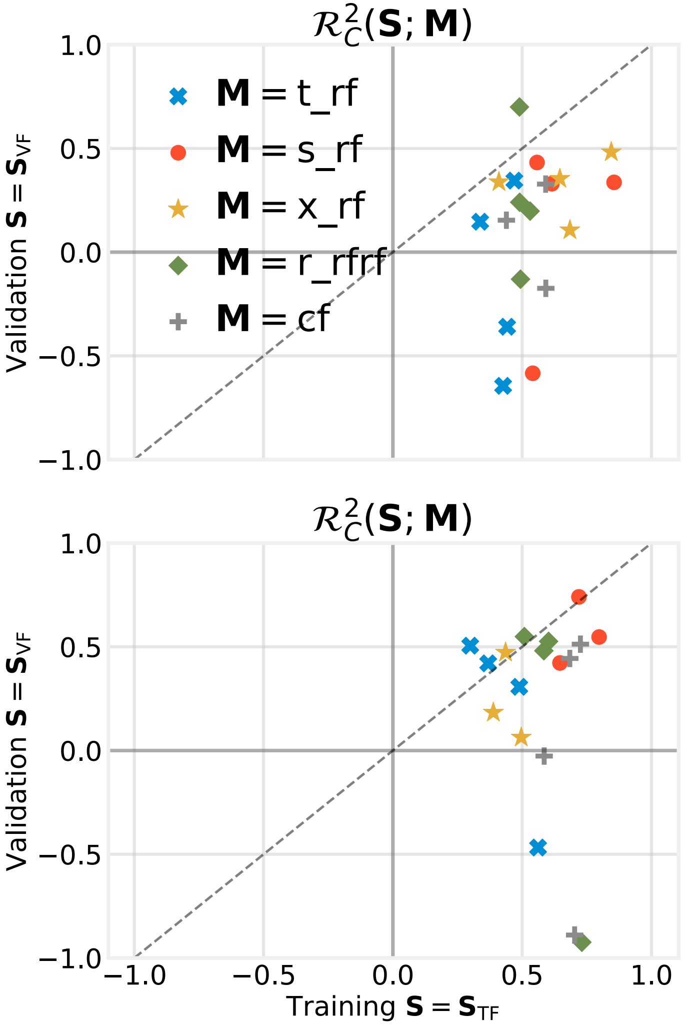
$$\text{Cal-Score}(\mathbf{S};\widehat{\tau}_{\text{ATE}}) := \sum_{j=1}^{K} \frac{|\mathbf{G}_j \cap \mathbf{S}|}{|\mathbf{S}|} \cdot \left|\widehat{\tau}_{\text{ATE}} - \widehat{\tau}_{\mathbf{G}_j \cap \mathbf{S}}\right|$$

$$\mathcal{R}^2_{\text{C}}(\mathbf{S};\mathbf{M}) := 1 - \frac{\text{Cal-Score}(\mathbf{S};\mathbf{M})}{\text{Cal-Score}(\mathbf{S};\widehat{\tau}_{\text{ATE}})}$$

- Lies in $(-\infty, 1]$
- Value close to 1 suggests good performance

Prediction check via calibration:
Poor generalization on validation set

(5 models, 4 folds)

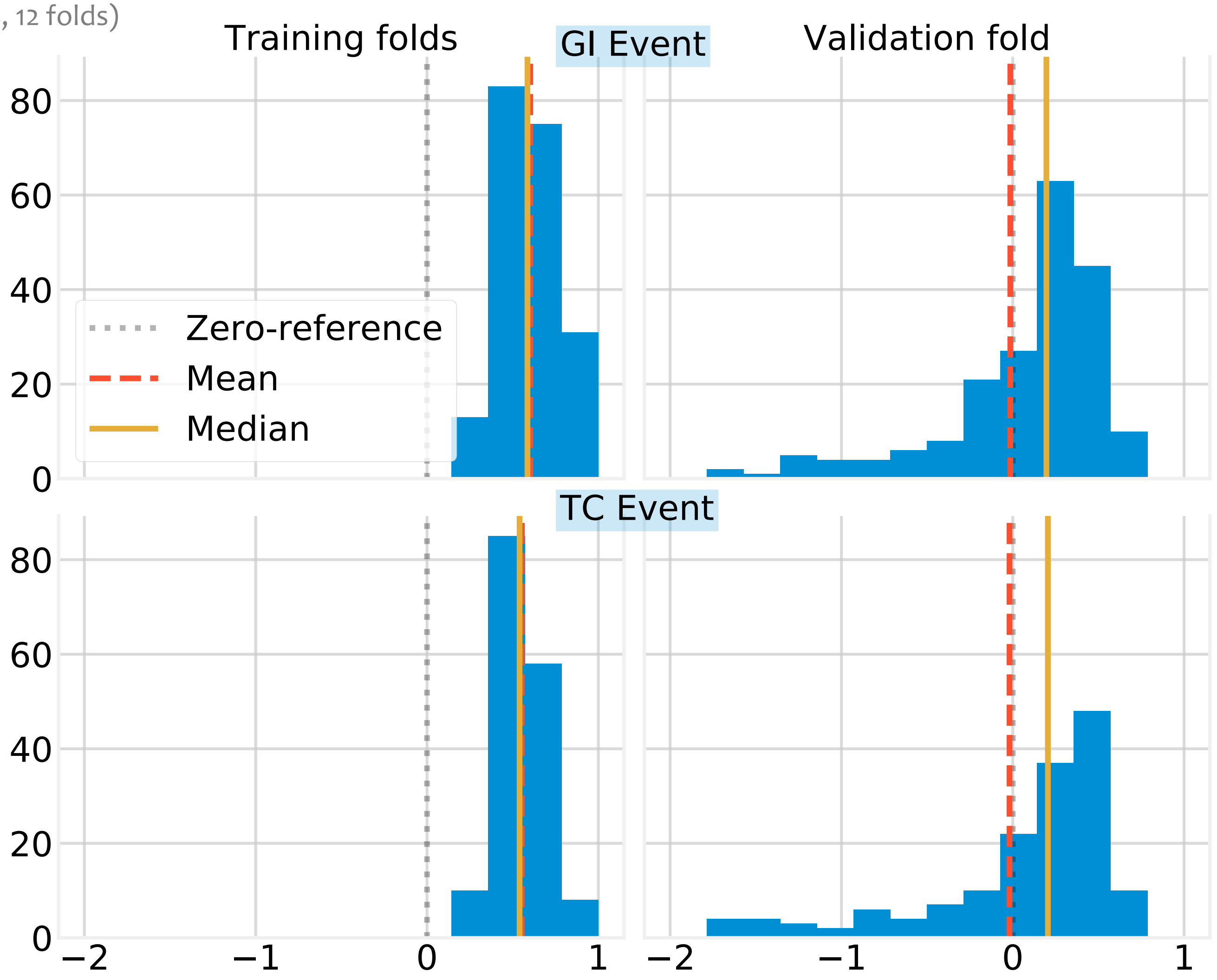


GI event

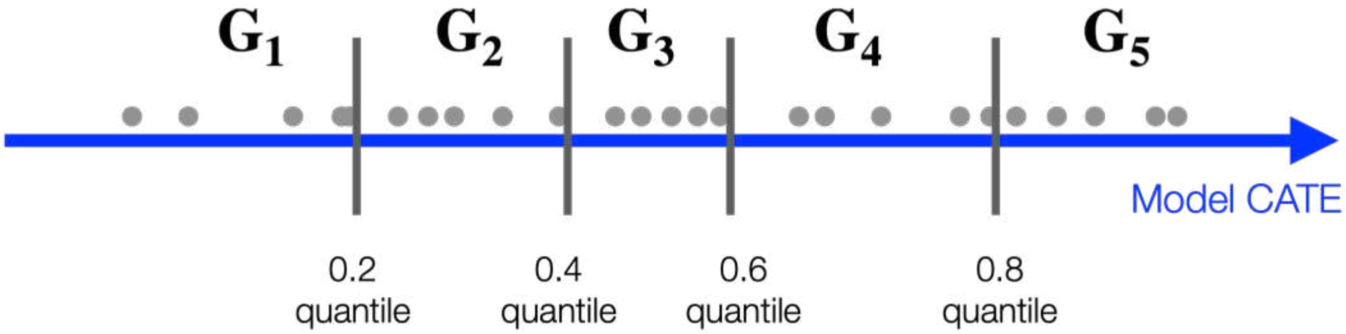
TC event

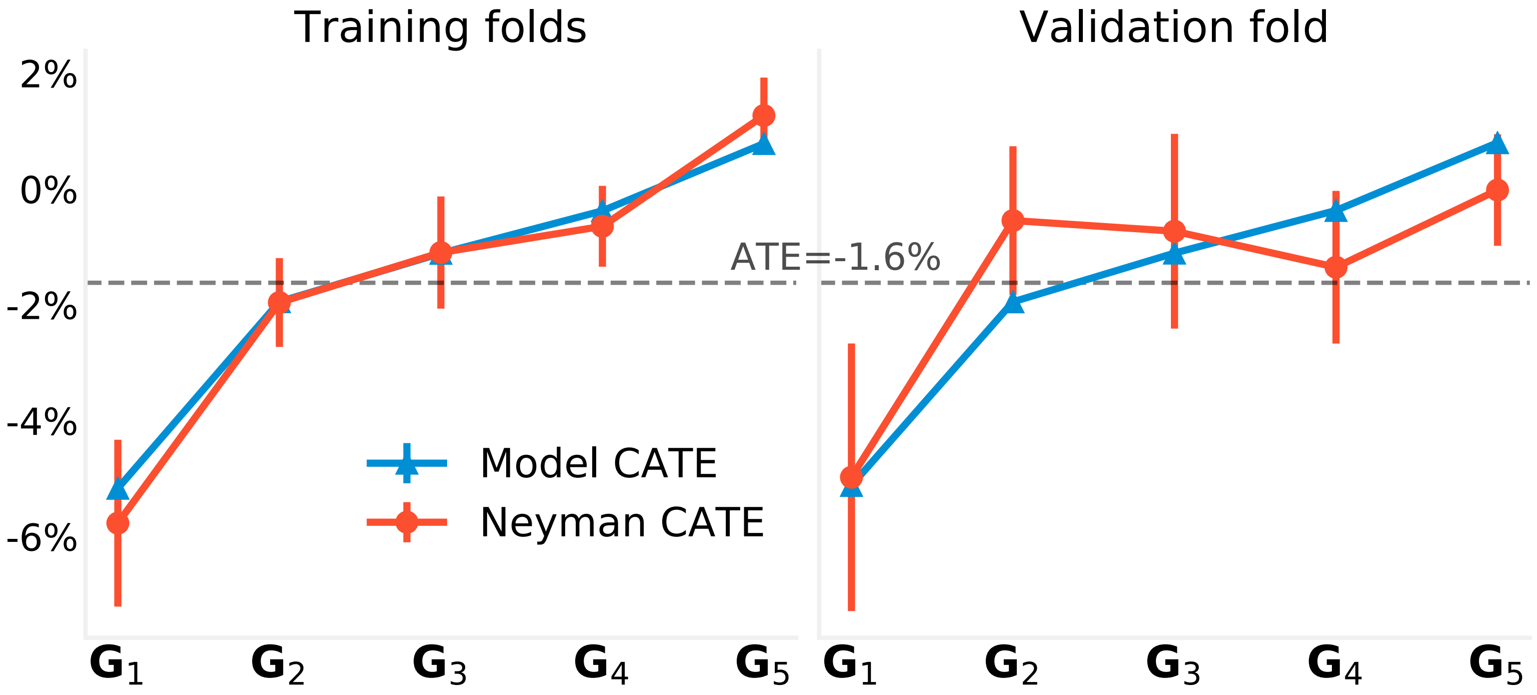
Prediction check via calibration:
Poor generalization on validation set

(18 models, 12 folds)



Prediction check via calibration: *Monotonicity*





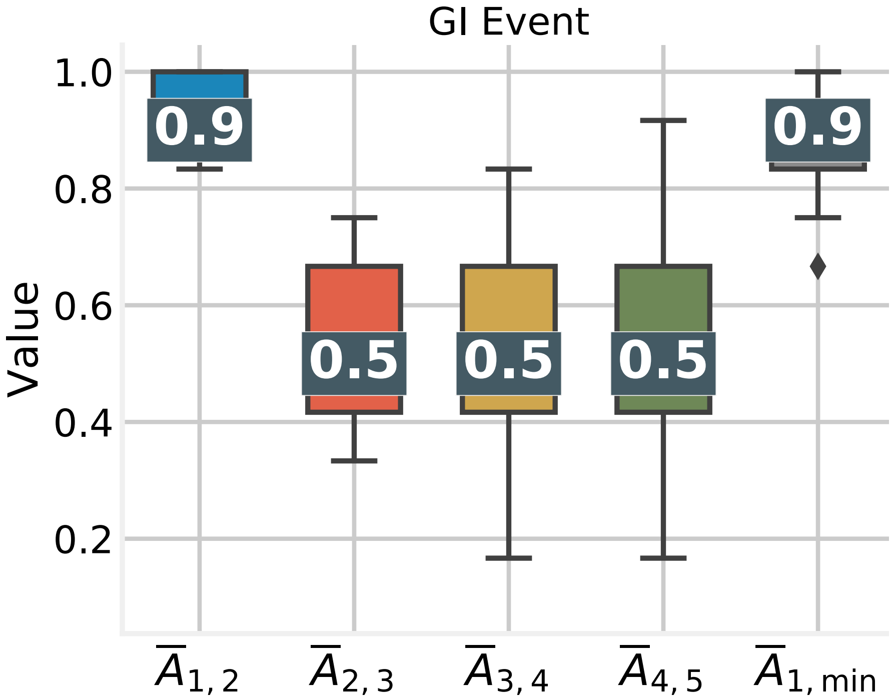
Prediction check via calibration:
Monotonicity in consecutive quantiles

$A_{j,j+1} = \text{Neyman estimate for Bin } \boldsymbol{G_j} < \text{Neyman Estimate for Bin } \boldsymbol{G_{j+1}}$

Prediction check via calibration: *Monotonicity in consecutive quantiles*

$$A_{j,j+1} = \text{Neyman estimate for Bin } \boldsymbol{G_j} < \text{Neyman Estimate for Bin } \boldsymbol{G_{j+1}}$$

(18 models, 12 folds)

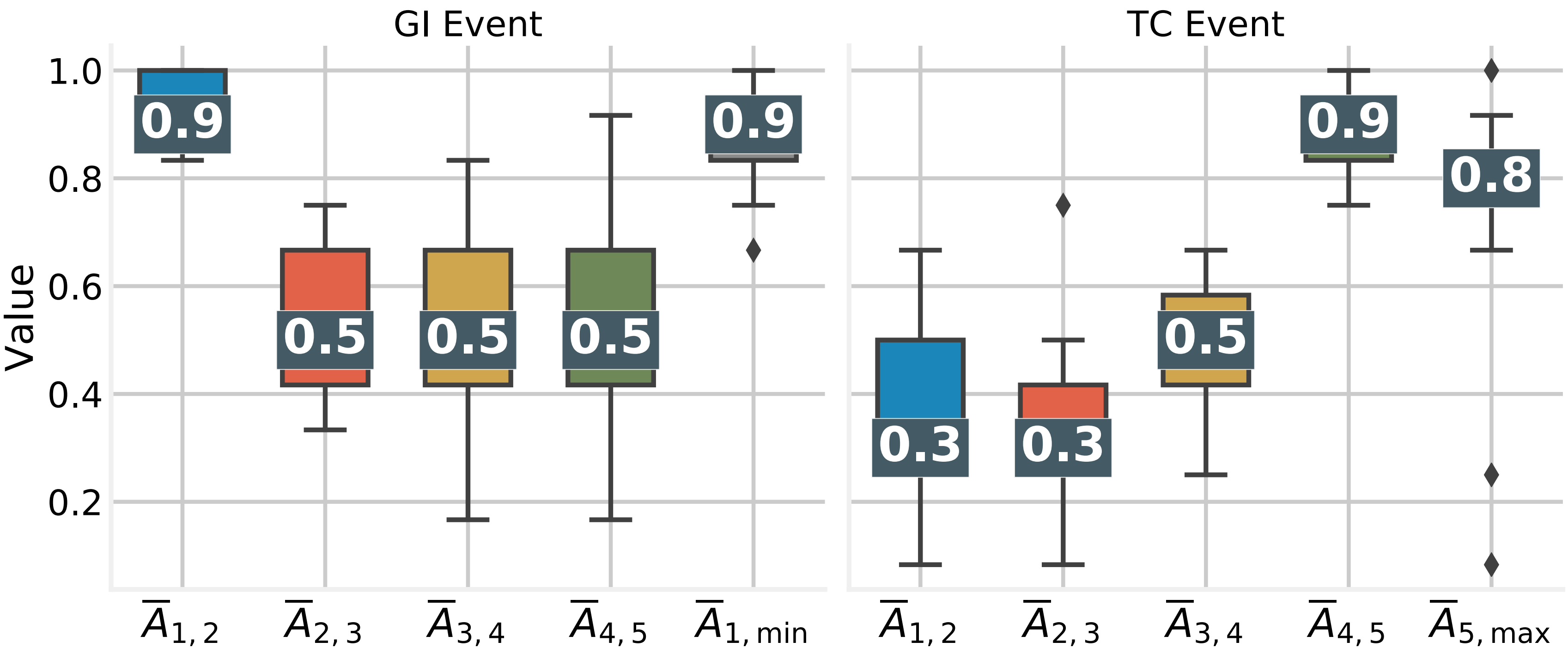


Neyman estimate for Bin $\boldsymbol{G_1}$
$= \min$ Neyman estimate for Bin $\boldsymbol{G_j}$

Prediction check via calibration: *Monotonicity in consecutive quantiles*

$$A_{j,j+1} = \text{Neyman estimate for Bin } \boldsymbol{G_j} < \text{Neyman Estimate for Bin } \boldsymbol{G_{j+1}}$$

(18 models, 12 folds)



Bottom/Top quantile-bins show promise?

Prediction check via calibration:
Take-aways *(for Vioxx dataset)*

- CATE models do not have “good generalization” on the whole dataset

- Top and bottom quantile-based subgroups seem promising

- Some CATE models better than others

- Questions:
  - How to aggregate/rank the models w.r.t. identifying subgroups?
  - Which quantile to choose?
  - How to obtain clinically interpretable subgroups?

Contributions

1. Extend PCS framework from supervised learning to causal studies

2. Introduce calibration-based predictive checks for CATE models

3. Overall, develop staDISC methodology for using CATE models to find interpretable subgroups

4. Case study with VIGOR, and external validation with APPROVe study

StaDISC: Applying PCS to CATE modeling

Stable Discovery of Interpretable Subgroups via Calibration

C

Feature Engineering
+ 18 CATE Models

P

Calibration-based
predictive screening

StaDISC: Applying PCS to CATE modeling

Stable Discovery of Interpretable Subgroups via Calibration

C

Feature Engineering + 18 CATE Models

P

Calibration-based predictive screening

S

Stability to data/model/ judgment perturbations

StaDISC: Applying PCS to CATE modeling

Stable Discovery of Interpretable Subgroups via Calibration

C

Feature Engineering
+ 18 CATE Models

P

Calibration-based predictive screening

S

Stability to data/model/judgment perturbations

Ranking and ensemble using P + S checks

StaDISC: Applying PCS to CATE modeling

Stable Discovery of Interpretable Subgroups via Calibration

C

Feature Engineering + 18 CATE Models

P

Calibration-based predictive screening

+

S

Stability to data/model/ judgment perturbations

Ranking and ensemble using P + S checks

Finding interepretable subgroups

Stability check:
The stability principle

“

A good estimator should have good performance on a slightly different dataset that could have arisen in a parallel world where a few choices were made differently.

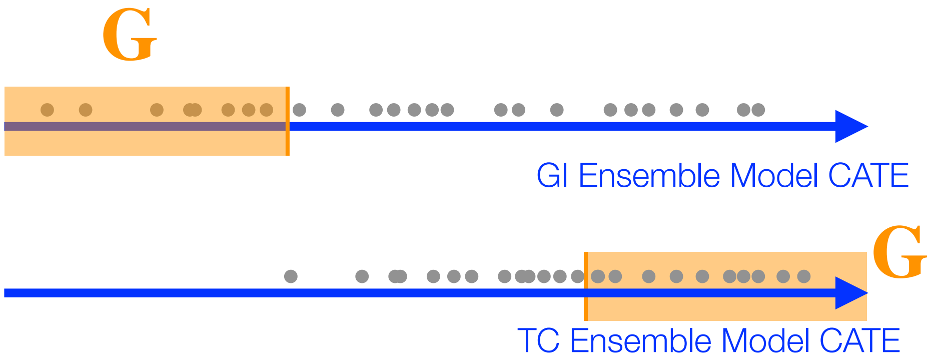
”

Stability check: *Appropriate data perturbations*

- Sampling perturbations
  - 2 additional random splits for CV
  - Enrollment time-based split
- Feature engineering perturbations
  - Different thresholds for defining “elderly” or “obese” features
  - Slightly perturbed definition of the outcome (include unconfirmed events)
- No hyperparameter tuning for the new splits/datasets

P + S check:
Top quantile-based subgroups

- Top quantile-based subgroups

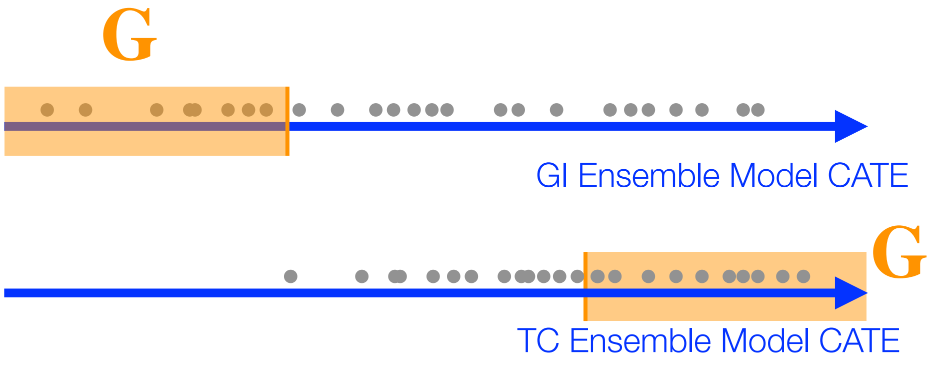


- Standardize subgroup CATE (t-statistics)

$$\mathbb{T}_{\mathbf{G}} := \frac{\widehat{\tau}_{\mathbf{G}} - \widehat{\tau}_{\mathrm{ATE}}}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\tau}_{\mathbf{G}} - \widehat{\tau}_{\mathrm{ATE}})}}$$

P + S check:
Top quantile-based subgroups

- Top quantile-based subgroups



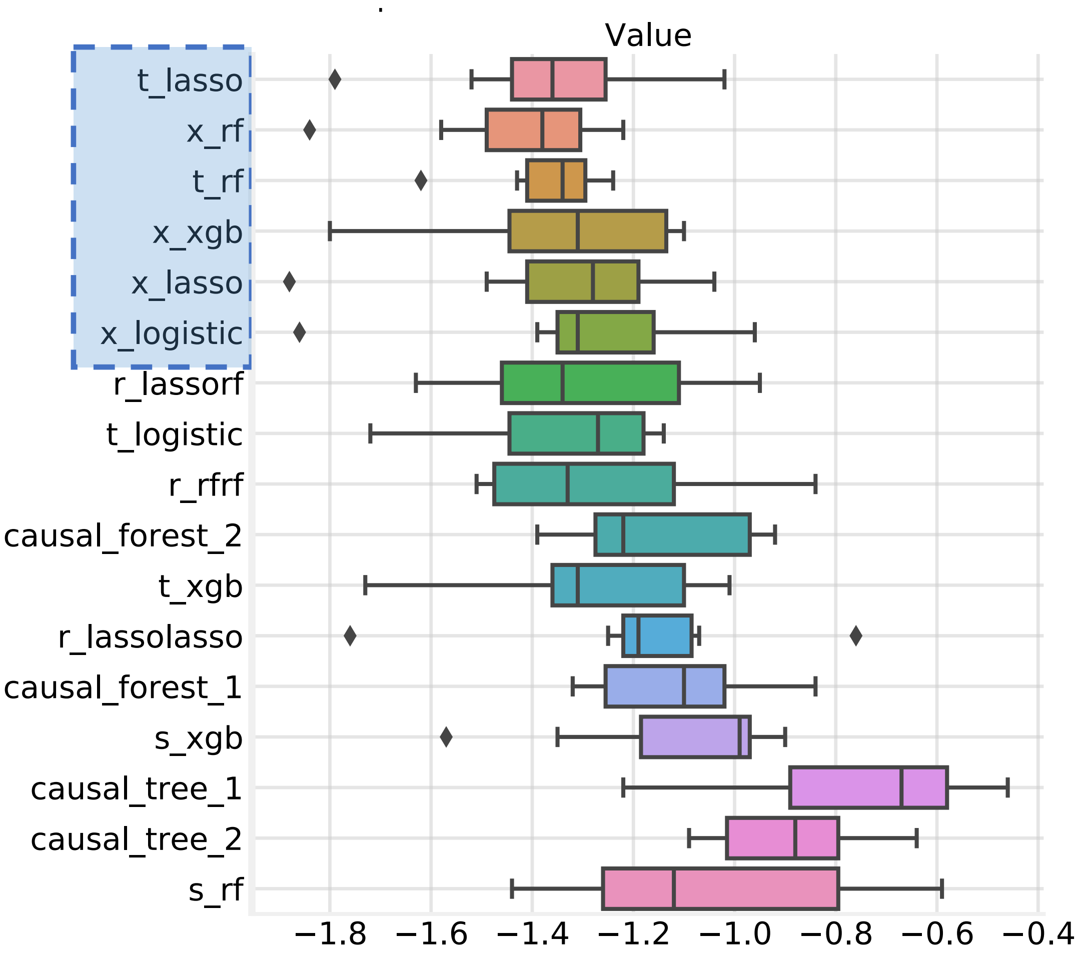
- Standardize subgroup CATE (t-statistics)

$$\mathbb{T}_{\mathbf{G}} := \frac{\widehat{\tau}_{\mathbf{G}} - \widehat{\tau}_{\mathrm{ATE}}}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\tau}_{\mathbf{G}} - \widehat{\tau}_{\mathrm{ATE}})}}$$

- For each perturbation $\mathfrak{D}$, compute avg. t-statistics across folds, and different bottom quantiles

P + S check:

Ranking the 18 CATE models based on T-statistics



(one sided) p-value vs t-statistics
0.05 --- 1.65
0.025 --- 1.96
0.001 --- 2.33

StaDISC: Applying PCS to CATE modeling

PCS

Top quantile-based subgroups of CATE Models (After P + S checks)

StaDISC: Applying PCS to CATE modeling *and* finding interpretable subgroups

PCS

Top quantile-based subgroups of CATE Models (After P + S checks)

→

Interpretability

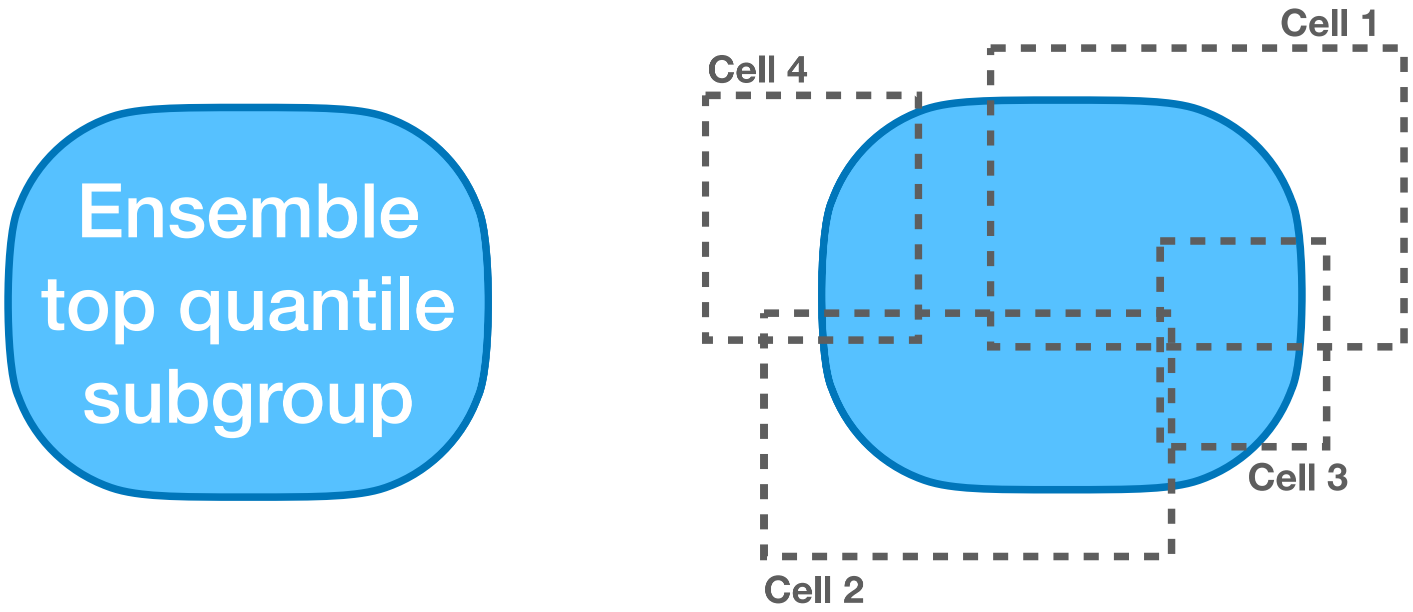
= Ensemble top models
+ Cell-Search to interpret the quantile-based subgroups

→

(Clinically) Interpretable Subgroups

Towards interpretable subgroups via cell search:

Find feature based representation of top quantiles



Desiderata:

Few stable disjoint cells---each based on few features---that have pure coverage of the quantile

StaDISC finds interpretable subgroups

Vioxx when compared to Naproxen

disproportionately **reduced GI Risk** for patients

- with history of GI
- with history of hypertension
  + prior usage of steroids
- with old age
  + prior usage of steroids

disproportionately **increased TC Risk** for patients

- with history of atherosclerosis
- with usage of aspirin indicated by FDA
- with old age and male gender*

*Poor generalization on test set, (no events)

StaDISC finds interpretable subgroups

Vioxx when compared to Naproxen

disproportionately **reduced GI Risk** for patients

- with history of GI
- with history of hypertension + prior usage of steroids
- with old age + prior usage of steroids

disproportionately **increased TC Risk** for patients

- with history of atherosclerosis
- with usage of aspirin indicated by FDA
- with old age and male gender*

*Poor generalization on test set, (no events)

Are these subgroups of more general relevance?

External validity

- RCTs are the gold standards for clinical research but…

“Between measurements based on RCTs and benefit . . . in the community there is a gulf which has been much under-estimated.”

- A L Cochrane, 1971

External validity of RCTs: "To whom do the results of this trial apply?" [Rothwell '05]

- Conclusions from one study may not be application for routine practice

- Differences in population, clinical monitoring, …

- From RCT to RCT, different outcomes of interest… .

The APPROVe study

- 2587 patients RCT during 2001-2004 by Merck
- Can Vioxx "reduce the risk of *adenomatous polyps* in individuals with a recent history of these tumors"?
- Treatment group: **Vioxx**, control group: **Placebo**
- High cardiovascular toxicity of Vioxx led to earlier termination by 2 months, and withdrawal of drug from the market

J. A. Baron et al.. Cardiovascular events associated with Rofecoxib: Final analysis of the APPROVe trial. The Lancet, 2008.

VIGOR vs APPROVe: Overview

	VIGOR	APPROVe
Duration	1999-2000 9 mon + 3 mon follow-up	2001-2004 3 yrs + 1 yr follow-up
Study Population	Patients with rheumatoid arthritis	Patients with history of colorectal polyps
Primary Focus	GI toxicity (gastrointestinal complications)	Adenomatous polyps (tumor in large intestine and rectum)
Control Arm	Naproxen	Placebo

VIGOR vs APPROVe: Overview

VIGOR Study (Control = Naproxen)	ATE	Base rate
Gastro-intestinal (GI) event	-1.6%	2.2%
Thrombotic cardiovascular (TC) event	0.6%	0.7%

APPROVe STUDY (Control = Placebo)	ATE	Base rate
Gastro-intestinal (GI) event	1.6%	0.5%
Thrombotic cardiovascular (TC) event	1.9%	2.5%

External validation: Interpretability helps!

- Clinical interpretability of our subgroups helps our attempts with external validation

``Do the subgroups found by StaDISC for the VIGOR study ***generalize*** *to the APPROVe study?"*

External validation: Interpretability helps!

- Clinical interpretability of our subgroups helps our attempts with external validation

*``Do the subgroups found by StaDISC for the VIGOR study **generalize** to the APPROVe study?”*

*`Mostly yes..... **4/6 subgroups show significant heterogeneous treatment effect** in the APPROVe study.”*

External validation of subgroups with APPROVe study

Vioxx when compared to placebo

disproportionately increased GI Risk for patients

- with history of GI
- with history of hypertension + prior usage of steroids*
- with old age + prior usage of steroids*

*Very small subgroup, no events

disproportionately increased TC Risk for patients

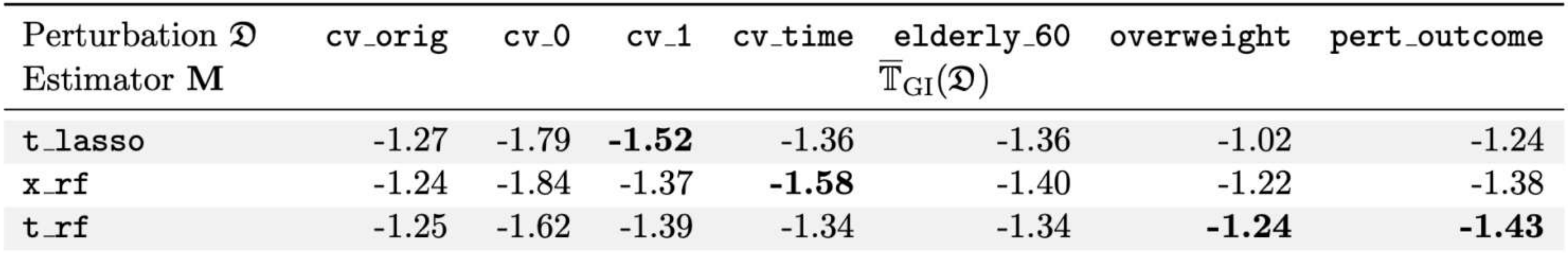
- with history of atherosclerosis
- with usage of aspirin indicated by FDA
- with old age and male gender

Contributions

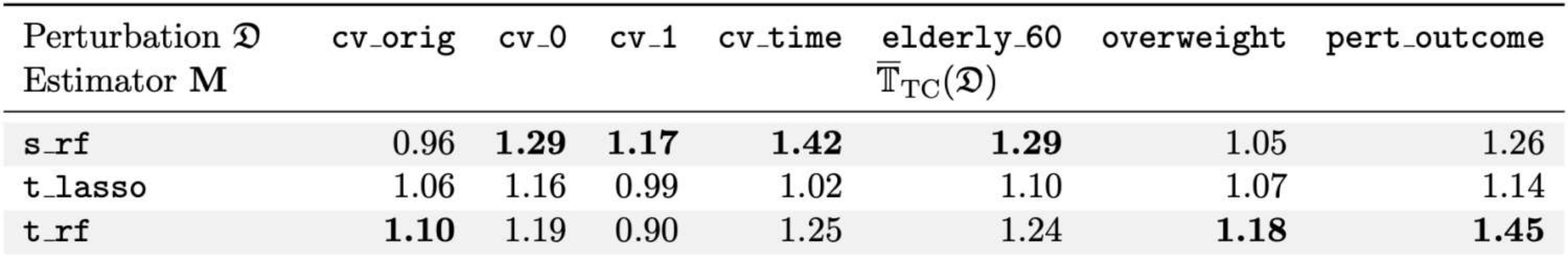
1. Extend PCS framework from supervised learning to causal studies
2. Introduce calibration-based predictive checks for CATE models
3. Overall, develop staDISC methodology for using CATE models to find interpretable subgroups
4. Case study with VIGOR study, and external validation with APPROVe study

Extra slides

P + S check:
Perturbation wise performance



Perturbation $\mathfrak{D}$	cv_orig	cv_0	cv_1	cv_time	elderly_60	overweight	pert_outcome
Estimator $\mathbf{M}$					$\overline{\mathbb{T}}_{\mathrm{GI}}(\mathfrak{D})$		
t_lasso	-1.27	-1.79	**-1.52**	-1.36	-1.36	-1.02	-1.24
x_rf	-1.24	-1.84	-1.37	**-1.58**	-1.40	-1.22	-1.38
t_rf	-1.25	-1.62	-1.39	-1.34	-1.34	**-1.24**	**-1.43**

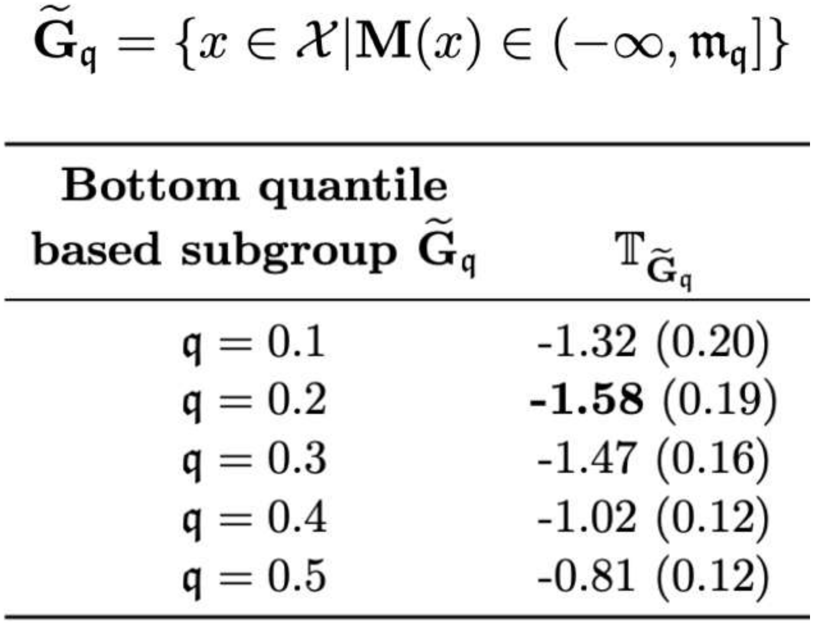


Perturbation $\mathfrak{D}$	cv_orig	cv_0	cv_1	cv_time	elderly_60	overweight	pert_outcome
Estimator $\mathbf{M}$					$\overline{\mathbb{T}}_{\mathrm{TC}}(\mathfrak{D})$		
s_rf	0.96	**1.29**	**1.17**	**1.42**	**1.29**	1.05	1.26
t_lasso	1.06	1.16	0.99	1.02	1.10	1.07	1.14
t_rf	**1.10**	1.19	0.90	1.25	1.24	**1.18**	**1.45**

(one sided) p-value vs t-statistics
0.05 --- 1.65
0.025 –- 1.96
0.001 --- 2.33

Which quantile group to interpret?

Find predictive and stable ones via t-statistics

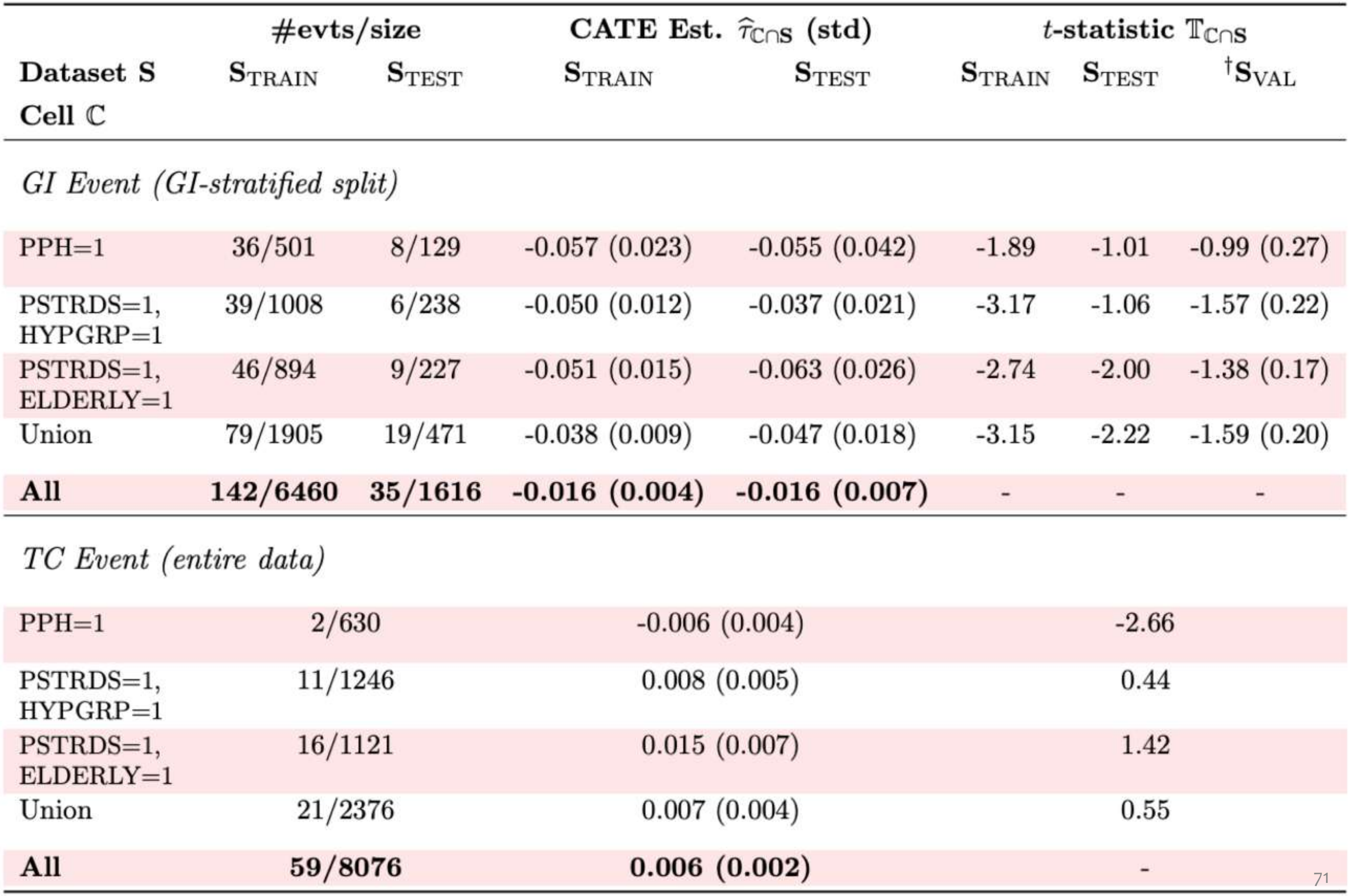


$$\widetilde{\mathbf{G}}_{\mathfrak{q}} = \{x \in \mathcal{X} | \mathbf{M}(x) \in (-\infty, \mathfrak{m}_{\mathfrak{q}}]\}$$

Bottom quantile based subgroup $\widetilde{\mathbf{G}}_{\mathfrak{q}}$	$\mathbb{T}_{\widetilde{\mathbf{G}}_{\mathfrak{q}}}$
$\mathfrak{q} = 0.1$	-1.32 (0.20)
$\mathfrak{q} = 0.2$	**-1.58** (0.19)
$\mathfrak{q} = 0.3$	-1.47 (0.16)
$\mathfrak{q} = 0.4$	-1.02 (0.12)
$\mathfrak{q} = 0.5$	-0.81 (0.12)

(average across 12 folds of 3 random CV splits)

Performance on Test Set

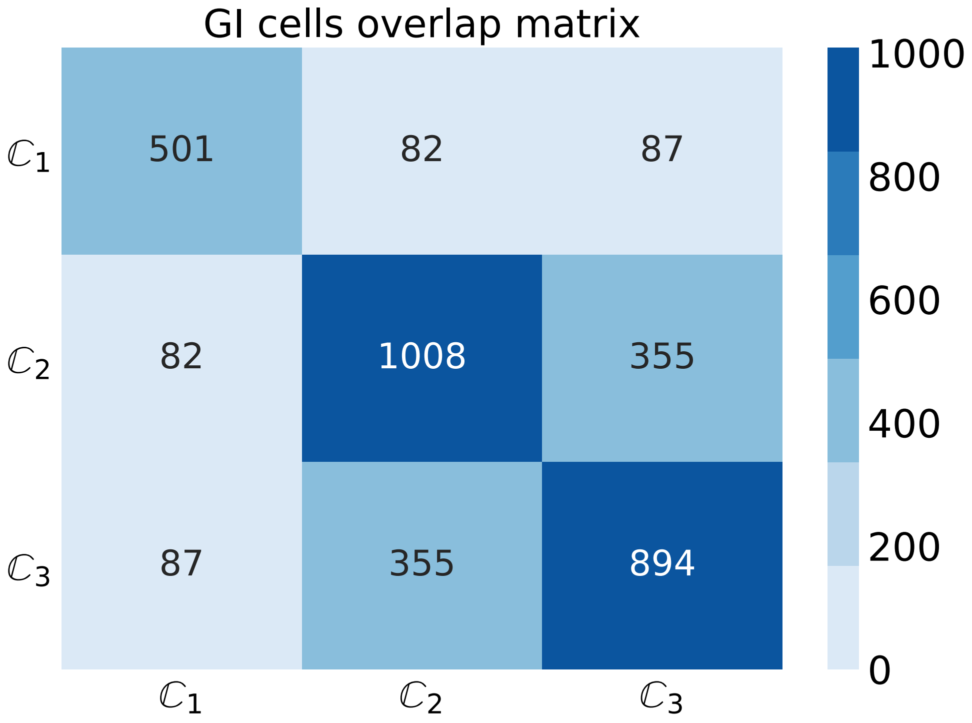


	#evts/size		CATE Est. $\widehat{\tau}_{\mathbb{C}\cap\mathbf{S}}$ (std)		t-statistic $\mathbb{T}_{\mathbb{C}\cap\mathbf{S}}$		
Dataset S **Cell** $\mathbb{C}$	$\mathbf{S}_{\text{TRAIN}}$	$\mathbf{S}_{\text{TEST}}$	$\mathbf{S}_{\text{TRAIN}}$	$\mathbf{S}_{\text{TEST}}$	$\mathbf{S}_{\text{TRAIN}}$	$\mathbf{S}_{\text{TEST}}$	$^{\dagger}\mathbf{S}_{\text{VAL}}$
GI Event (GI-stratified split)							
PPH=1	36/501	8/129	-0.057 (0.023)	-0.055 (0.042)	-1.89	-1.01	-0.99 (0.27)
PSTRDS=1, HYPGRP=1	39/1008	6/238	-0.050 (0.012)	-0.037 (0.021)	-3.17	-1.06	-1.57 (0.22)
PSTRDS=1, ELDERLY=1	46/894	9/227	-0.051 (0.015)	-0.063 (0.026)	-2.74	-2.00	-1.38 (0.17)
Union	79/1905	19/471	-0.038 (0.009)	-0.047 (0.018)	-3.15	-2.22	-1.59 (0.20)
All	**142/6460**	**35/1616**	**-0.016 (0.004)**	**-0.016 (0.007)**	-	-	-
TC Event (entire data)							
PPH=1	2/630		-0.006 (0.004)		-2.66		
PSTRDS=1, HYPGRP=1	11/1246		0.008 (0.005)		0.44		
PSTRDS=1, ELDERLY=1	16/1121		0.015 (0.007)		1.42		
Union	21/2376		0.007 (0.004)		0.55		
All	**59/8076**		**0.006 (0.002)**		-		

Cell search results:

Top stable cells for high negative CATE for GI Event

GI cells overlap matrix

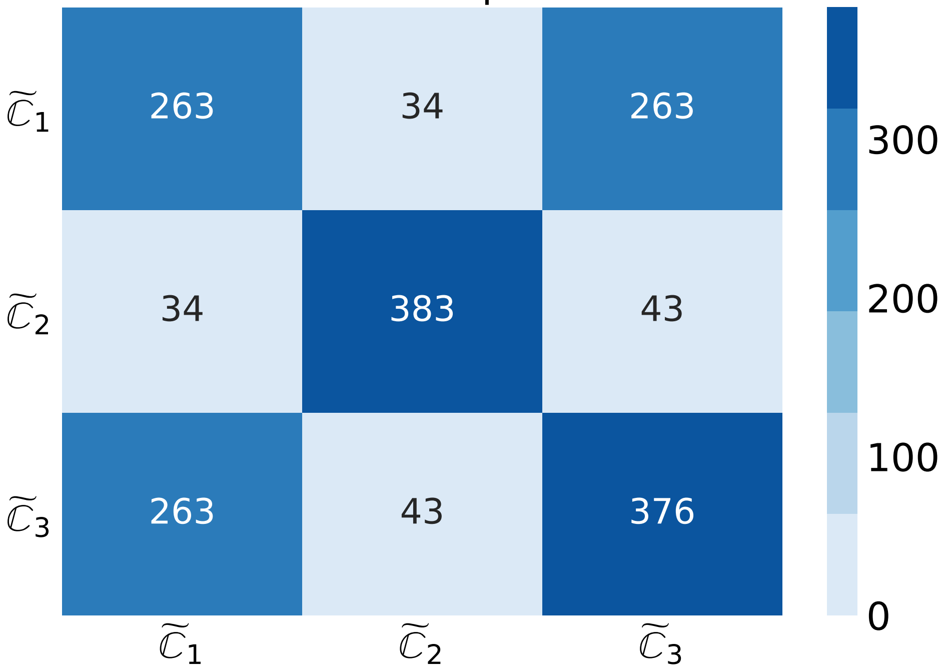


	$\mathbb{C}_1$	$\mathbb{C}_2$	$\mathbb{C}_3$
$\mathbb{C}_1$	501	82	87
$\mathbb{C}_2$	82	1008	355
$\mathbb{C}_3$	87	355	894

$\mathbb{C}_1$ = Patients with prior history of GI events
$\mathbb{C}_2$ = Patients with prior usage of steroids, and history of hypertension
$\mathbb{C}_3$ = Elderly patients with prior usage of steroids

Cell search results:

Top stable cells for high positive CATE for TC Event

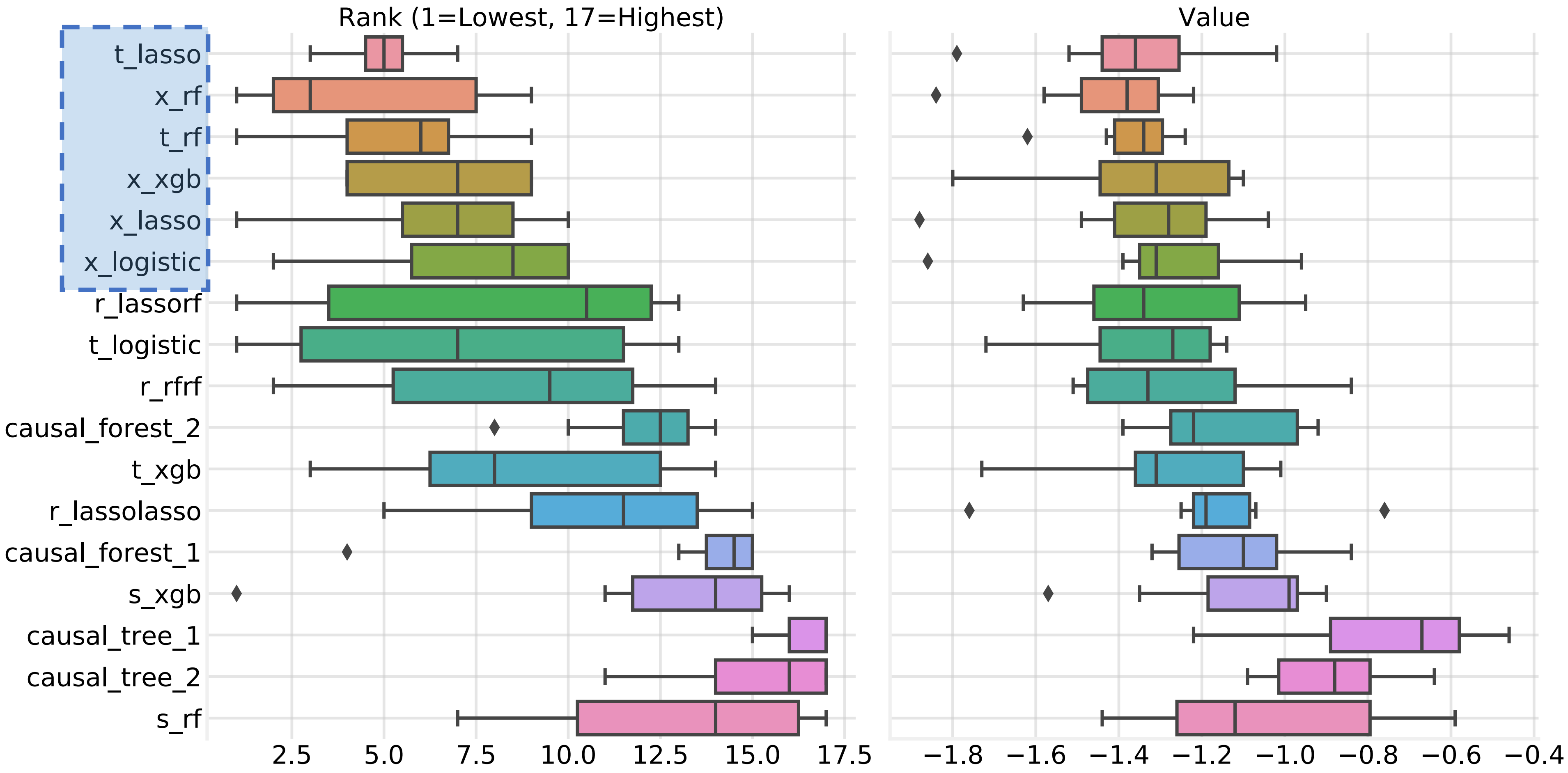


TC cells overlap matrix

	$\widetilde{\mathbb{C}}_1$	$\widetilde{\mathbb{C}}_2$	$\widetilde{\mathbb{C}}_3$
$\widetilde{\mathbb{C}}_1$	263	34	263
$\widetilde{\mathbb{C}}_2$	34	383	43
$\widetilde{\mathbb{C}}_3$	263	43	376

$\widetilde{\mathbb{C}}_1 =$ Patients with aspirin indicated
$\widetilde{\mathbb{C}}_2 =$ Elderly male patients
$\widetilde{\mathbb{C}}_3 =$ Patients with history of atherosclerotic cardiovascular disease

P + S check:
Ranking of CATE models


Rank (1=Lowest, 17=Highest)
Value
t_lasso
x_rf
t_rf
x_xgb
x_lasso
x_logistic
r_lassorf
t_logistic
r_rfrf
causal_forest_2
t_xgb
r_lassolasso
causal_forest_1
s_xgb
causal_tree_1
causal_tree_2
s_rf