Clinical value of predicting individual treatment effects for intensive BP therapy

A machine learning experiment to estimate treatment effects from randomized trial data

Tony Duan, Pranav Rajpurkar, Dillon Laird, Andrew Y. Ng, Sanjay Basu

May 10, 2019

NBER Machine Learning in Health Care Conference, Cambridge MA

Table of contents

1. Background

2. Methods

3. Evaluation

4. Results

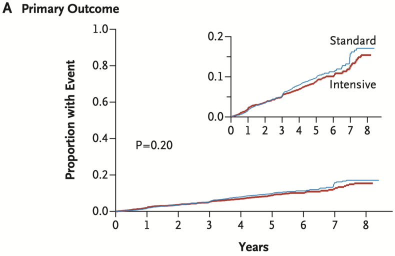
Background

Background

How should we prescribe BP medications to prevent CVD events?

- Intensive: target SBP < 120 mmHg
- Standard: target SBP < 140 mmHg

ACCORD RCT with 4733 participants: $HR = 0.88$ (95% CI 0.73-1.06) [1].



No. at Risk									
Intensive	2362	2273	2182	2117	1770	1080	298	175	80
Standard	2371	2274	2196	2120	1793	1127	358	195	108

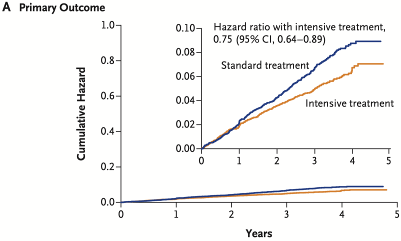
[1] [The ACCORD Study Group, 2010]

Background

How should we prescribe BP medications to prevent CVD events?

- Intensive: target SBP < 120 mmHg
- Standard: target SBP < 140 mmHg

SPRINT RCT with 9361 participants: $HR = 0.75$ (95% CI 0.64-0.89) [2].



A Primary Outcome

No. at Risk					
Standard treatment	4683	4437	4228	2829	721
Intensive treatment	4678	4436	4256	2900	779

[2] [The SPRINT Research Group, 2015]

Background

Who benefits more or less from intensive blood pressure therapy?

We assume the potential outcomes framework.

- feature vector $X^{(i)} \in \mathbb{R}^p$
- response $Y^{(i)} \in \{0, 1\}$
- treatment $W^{(i)} \in \{0, 1\}$

Traditionally, RCTs measure the average treatment effect (ATE).

$$\tau = \mathbb{E}[Y(1) - Y(0)]$$

We want to estimate conditional average treatment effects (CATE).

$$\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$$

Epidemiology perspective: $-\tau(x)$ is the absolute risk reduction (ARR).

Background

We want to estimate conditional average treatment effects (CATE).

$$\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$$

Most causal inference techniques assume unconfoundedness (ignorability).

$$\{Y(0), Y(1)\} \perp\!\!\!\perp W|X$$

In the RCT setting, we can make an even stronger assumption.

$$\{Y(0), Y(1)\} \perp\!\!\!\perp W$$

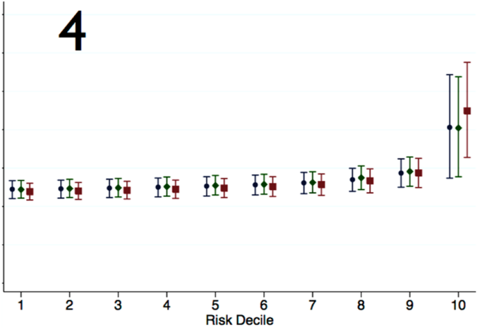
With propensity scores $\Pr[W = 0] = \Pr[W = 1] = \frac{1}{2}$.

Background

Traditional approaches to assessing heterogeneity in treatment effects have been subgroup analyses, partitioned by baseline risk.

It's typically assumed that effects are proportional to baseline risk [3].

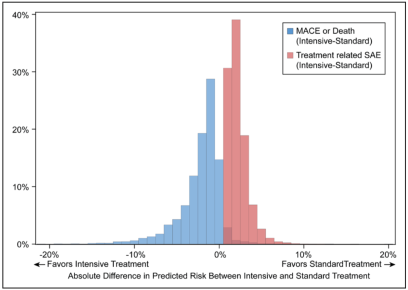
For example, higher benefit from treatment if CVD risk is high.



[3] [Burke et al., 2014]

Background

Prior work has used data from SPRINT and ACCORD to develop Cox/logistic regression models for ARR benefit/harm, showing significant heterogeneity in predicted effects [4].



But these make strong assumptions around linearity/proportional hazards.

[4] [Patel et al., 2017]

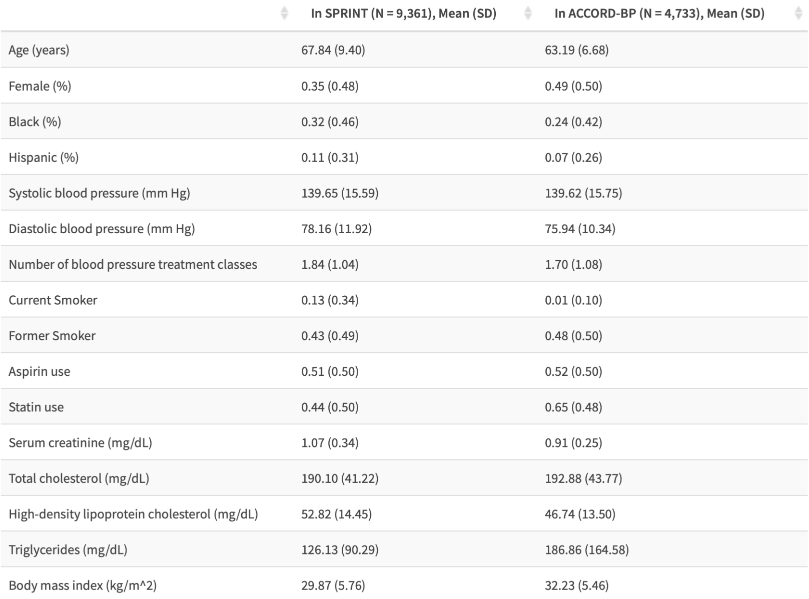
Background

Can machine learning methods improve estimates of treatment effects?

Will they reveal effects that are proportional to baseline risk?

Methods

Data



	In SPRINT (N = 9,361), Mean (SD)	In ACCORD-BP (N = 4,733), Mean (SD)
Age (years)	67.84 (9.40)	63.19 (6.68)
Female (%)	0.35 (0.48)	0.49 (0.50)
Black (%)	0.32 (0.46)	0.24 (0.42)
Hispanic (%)	0.11 (0.31)	0.07 (0.26)
Systolic blood pressure (mm Hg)	139.65 (15.59)	139.62 (15.75)
Diastolic blood pressure (mm Hg)	78.16 (11.92)	75.94 (10.34)
Number of blood pressure treatment classes	1.84 (1.04)	1.70 (1.08)
Current Smoker	0.13 (0.34)	0.01 (0.10)
Former Smoker	0.43 (0.49)	0.48 (0.50)
Aspirin use	0.51 (0.50)	0.52 (0.50)
Statin use	0.44 (0.50)	0.65 (0.48)
Serum creatinine (mg/dL)	1.07 (0.34)	0.91 (0.25)
Total cholesterol (mg/dL)	190.10 (41.22)	192.88 (43.77)
High-density lipoprotein cholesterol (mg/dL)	52.82 (14.45)	46.74 (13.50)
Triglycerides (mg/dL)	126.13 (90.29)	186.86 (164.58)
Body mass index (kg/m^2)	29.87 (5.76)	32.23 (5.46)

Data

RCT data is time-to-event, so we need to account for censoring to predict treatment effects for binarized outcomes at 3 years.

- feature vector $X^{(i)} \in \mathbb{R}^p$
- time to censoring or event $T^{(i)} \in \mathbb{R}^+$
- censoring indicator $C^{(i)} \in \{0, 1\}$
- treatment $W^{(i)} \in \{0, 1\}$

How do we define binarized outcomes in the presence of censoring?

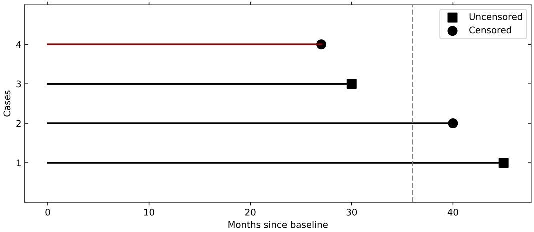
[In our dataset, relatively low censorship rate $\sim 23\%$.]

Data

We use inverse probability of censoring weighting; weight data points by [5]

$$\omega(T^{(i)}, C^{(i)}, X^{(i)}) = \underbrace{\frac{1\{T^{(i)} \geq 3\}}{\hat{p}(C \geq 3 | X = X^{(i)})}}_{\text{[case 1, case 2]}} + \underbrace{\frac{1\{T^{(i)} \leq 3, C^{(i)} = 0\}}{\hat{p}(C \geq T^{(i)} | X = X^{(i)})}}_{\text{[case 3]}}.$$

We use Cox regression to estimate the censoring distribution $\hat{p}(C|X)$.



[5] [Vock et al., 2016]

S-Learner and T-Learner

Consider meta-learning methods for predicting treatment effects.

S-Learner ("single")

Use machine learning to learn $p(Y|X,W)$, then predict

$$\tau(x) = \hat{p}(Y|X,W=1) - \hat{p}(Y|X,W=0)$$

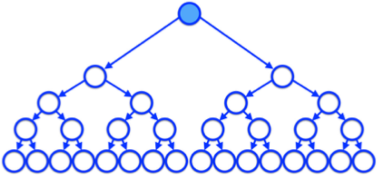
T-Learner ("two")

Use machine learning to learn two separate models:
$p(Y|X,W=0)$ and $p(Y|X,W=1)$, then predict

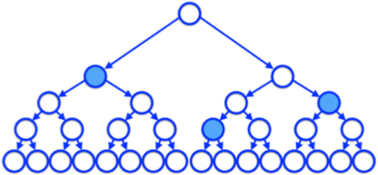
$$\tau(x) = \hat{p}(Y|X,W=1) - \hat{p}(Y|X,W=0)$$

S-Learner and T-Learner

In the context of random forests, the difference is that the T-learner forces the first split to be on W, whereas the S-learner treats W like any other covariate [6].



T-learner



S-learner

[6] [Künzel et al., 2019]

X-Learner

X-Learner

Three-step process:

1. Estimate response surfaces conditional on treatment.

$$\mu_0(x) = \mathbb{E}[Y(0)|X_{W=0}] \qquad \mu_1(x) = \mathbb{E}[Y(1)|X_{W=1}]$$

2. Impute treatment effects for each participant, using the model corresponding to the un-observed outcome.

$$\tau_1^{(i)} = Y_{W=1}^{(i)} - \hat{\mu}_0(X_{W=1}^{(i)}) \qquad \tau_0^{(i)} = \hat{\mu}_1(X_{W=0}^{(i)}) - Y_{W=0}^{(i)}$$

3. Fit models for $\hat{\tau}_1(x)$ and $\hat{\tau}_0(x)$ using imputed treatment effects, then

$$\hat{\tau}(x) = e(x)\hat{\tau}_0(x) + (1 - e(x))\hat{\tau}_1(x),$$

where $e(x)$ is the estimated propensity, in this case $e(x) = \frac{1}{2}$.

Experiment

Conventional: S-Learner logistic regression that predicts 3-year CVD event using treatment as an indicator variable, as well as all interaction terms between treatment and covariates.

1. Regress $Y \sim X + W + WX$.
2. Predict $\hat{\tau}(X) = \hat{p}(Y|X, W = 1) - \hat{p}(Y|X, W = 0)$.

Machine learning: X-learner with random forest base learners.

Evaluation

C-statistic for benefit

C-statistic. Proportion of all pairs with discordant outcomes, in which the $Y = 1$ event was assigned a higher probability than the $Y = 0$ event.

C-for-benefit. Proportion of all matched pairs with unequal observed benefit, in which the patient pair receiving greater treatment benefit was predicted to do so. Match each pair of patients to have one $(W = 0, W = 1)$ and identical predicted ARR.

> *The c-for-benefit thus represents the probability that from two randomly chosen matched patient pairs with unequal observed benefit, the pair with greater observed benefit also has a higher predicted benefit* [7]*.*

[7] [van Klaveren et al., 2018]

C-statistic for benefit

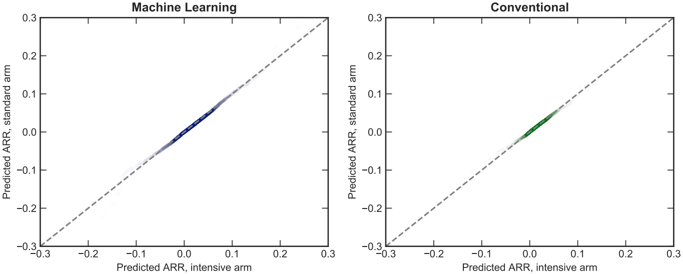


Figure 1: Matching of patient pairs across treatment arms. Ideally the pairs lie exactly along the diagonal, which is observed in this case due to the RCT nature of the data.

Decision value of RMST

Restricted mean survival time is defined as,

$$\text{RMST} = \mathbb{E}[\min(T, 3)] = \int_0^3 p(T > t)dt.$$

We estimate the RMST under the policy implied by the ARR model (i.e. treat those with ARR > 0, do not treat those with ARR ≤ 0).

This is known as *off-policy policy evaluation*.

1. Fit $\hat{p}(T \geq 3 | \hat{\tau}(x) \leq 0, W = 1), \hat{p}(T \geq 3 | \hat{\tau}(x) > 0, W = 0)$ via KM.
2. Estimate $\text{R\hat{M}ST}_{\hat{\tau}(x) \leq 0}$ and $\text{R\hat{M}ST}_{\hat{\tau}(x) > 0}$ via integration of KM.
3. $\text{R\hat{M}ST} = \mathbb{E}\left(1\{\hat{\tau}(x) \leq 0\}\text{R\hat{M}ST}_{\hat{\tau}(x) \leq 0} + 1\{\tau(x) > 0\}\text{R\hat{M}ST}_{\hat{\tau}(x) > 0}\right)$

Has been advocated for as a less biased and more interpretable method for model selection when predicting treatment effects [8].

[8] [Schuler and Shah, 2018]

Calibration

Risk estimation: Partition individuals into quantiles of predicted risk, and compare empirical risk to predicted.

ARR estimation: Partition individuals into quantiles of predicted ARR, and compare empirical risk reduction to predicted.

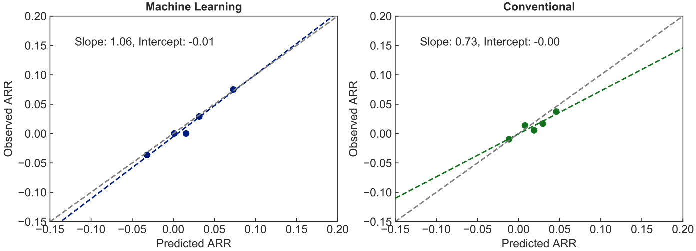


Figure 2: We compare predicted against observed absolute risk reduction at quintiles of predicted absolute risk reduction. Kaplan-Meier estimates are used to account for censoring.

Results

Predicted absolute risk reductions

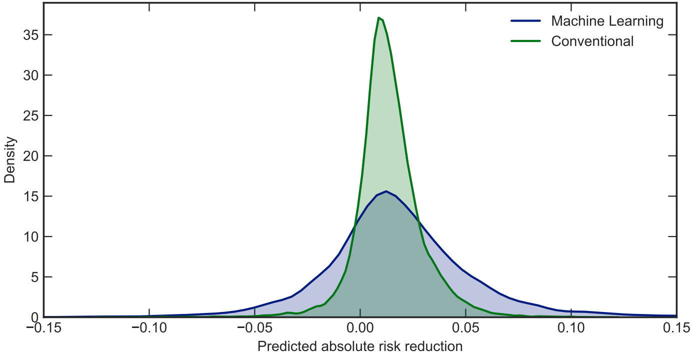


Figure 3: Distributions of predicted absolute risk reduction show greater heterogeneity in machine learning predictions compared to a conventional method.

Results

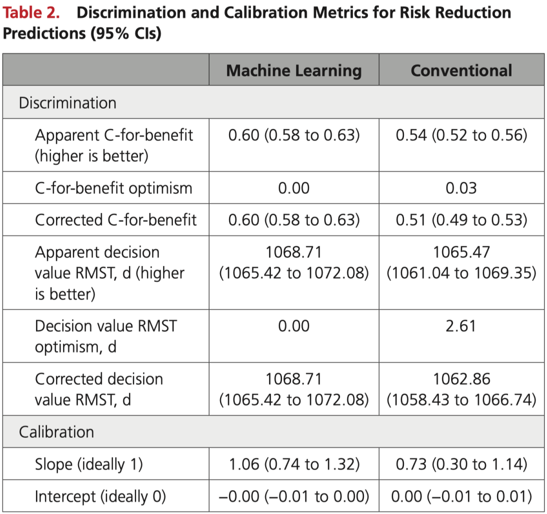


Table 2. Discrimination and Calibration Metrics for Risk Reduction Predictions (95% CIs)

	Machine Learning	Conventional
Discrimination		
Apparent C-for-benefit (higher is better)	0.60 (0.58 to 0.63)	0.54 (0.52 to 0.56)
C-for-benefit optimism	0.00	0.03
Corrected C-for-benefit	0.60 (0.58 to 0.63)	0.51 (0.49 to 0.53)
Apparent decision value RMST, d (higher is better)	1068.71 (1065.42 to 1072.08)	1065.47 (1061.04 to 1069.35)
Decision value RMST optimism, d	0.00	2.61
Corrected decision value RMST, d	1068.71 (1065.42 to 1072.08)	1062.86 (1058.43 to 1066.74)
Calibration		
Slope (ideally 1)	1.06 (0.74 to 1.32)	0.73 (0.30 to 1.14)
Intercept (ideally 0)	–0.00 (–0.01 to 0.00)	0.00 (–0.01 to 0.01)

Baseline policy of intensive treatment for SPRINT and standard treatment for ACCORD yields 3-year decision RMST of 1061.24 days (95% CI: 1057.37 - 1064.10).

Clinical relevance

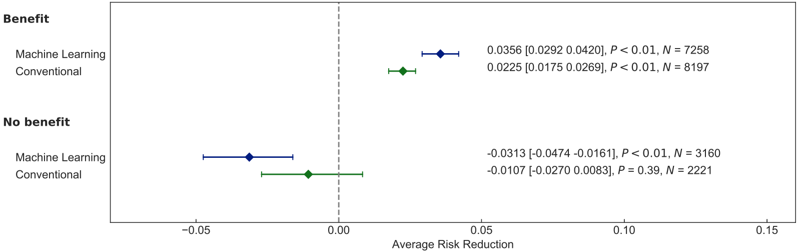


Figure 4: Bootstrapped observed risk reduction when partitioning individuals into buckets of treatment (ARR > 0) and control (ARR ≤ 0). ML produces more a discriminative decision boundary.

Clinical relevance

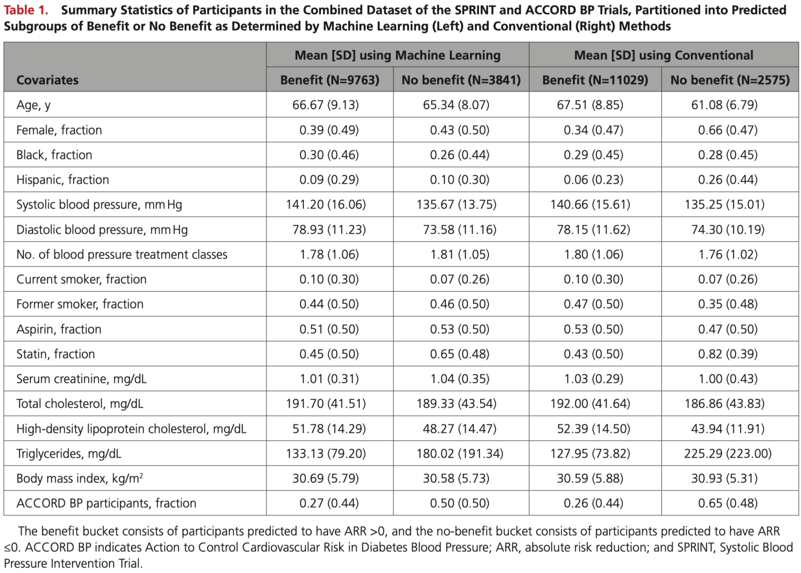


Table 1. Summary Statistics of Participants in the Combined Dataset of the SPRINT and ACCORD BP Trials, Partitioned into Predicted Subgroups of Benefit or No Benefit as Determined by Machine Learning (Left) and Conventional (Right) Methods

	Mean [SD] using Machine Learning		Mean [SD] using Conventional	
Covariates	**Benefit (N=9763)**	**No benefit (N=3841)**	**Benefit (N=11029)**	**No benefit (N=2575)**
Age, y	66.67 (9.13)	65.34 (8.07)	67.51 (8.85)	61.08 (6.79)
Female, fraction	0.39 (0.49)	0.43 (0.50)	0.34 (0.47)	0.66 (0.47)
Black, fraction	0.30 (0.46)	0.26 (0.44)	0.29 (0.45)	0.28 (0.45)
Hispanic, fraction	0.09 (0.29)	0.10 (0.30)	0.06 (0.23)	0.26 (0.44)
Systolic blood pressure, mm Hg	141.20 (16.06)	135.67 (13.75)	140.66 (15.61)	135.25 (15.01)
Diastolic blood pressure, mm Hg	78.93 (11.23)	73.58 (11.16)	78.15 (11.62)	74.30 (10.19)
No. of blood pressure treatment classes	1.78 (1.06)	1.81 (1.05)	1.80 (1.06)	1.76 (1.02)
Current smoker, fraction	0.10 (0.30)	0.07 (0.26)	0.10 (0.30)	0.07 (0.26)
Former smoker, fraction	0.44 (0.50)	0.46 (0.50)	0.47 (0.50)	0.35 (0.48)
Aspirin, fraction	0.51 (0.50)	0.53 (0.50)	0.53 (0.50)	0.47 (0.50)
Statin, fraction	0.45 (0.50)	0.65 (0.48)	0.43 (0.50)	0.82 (0.39)
Serum creatinine, mg/dL	1.01 (0.31)	1.04 (0.35)	1.03 (0.29)	1.00 (0.43)
Total cholesterol, mg/dL	191.70 (41.51)	189.33 (43.54)	192.00 (41.64)	186.86 (43.83)
High-density lipoprotein cholesterol, mg/dL	51.78 (14.29)	48.27 (14.47)	52.39 (14.50)	43.94 (11.91)
Triglycerides, mg/dL	133.13 (79.20)	180.02 (191.34)	127.95 (73.82)	225.29 (223.00)
Body mass index, kg/m^2	30.69 (5.79)	30.58 (5.73)	30.59 (5.88)	30.93 (5.31)
ACCORD BP participants, fraction	0.27 (0.44)	0.50 (0.50)	0.26 (0.44)	0.65 (0.48)

The benefit bucket consists of participants predicted to have ARR >0, and the no-benefit bucket consists of participants predicted to have ARR ≤0. ACCORD BP indicates Action to Control Cardiovascular Risk in Diabetes Blood Pressure; ARR, absolute risk reduction; and SPRINT, Systolic Blood Pressure Intervention Trial.

Heterogeneity beyond baseline risk

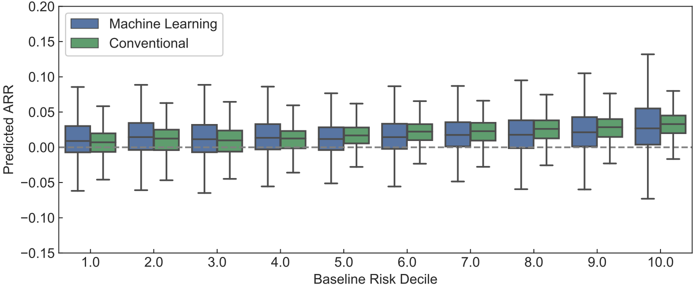


Figure 5: Predictions with the ML method exhibit more heterogeneity, not necessarily proportional to baseline risk. Baseline risk predictions made using the ACC/AHA ASCVD risk calculator.

Interpretability

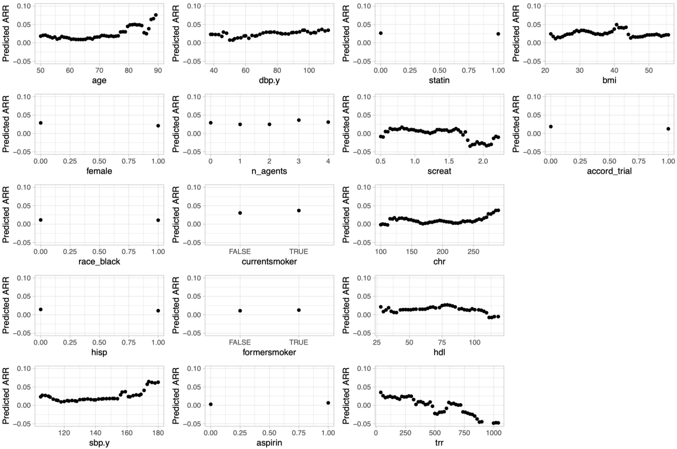
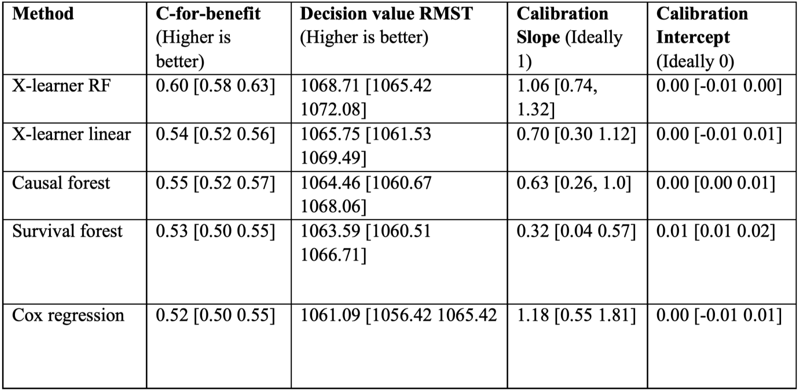


Figure 6: Partial dependence plots show non-linear dependencies between estimated treatment effect and covariates.

Sensitivity analysis

We found that the X-learner with RF beat out alternative ML methods, as well as a Cox regression baseline.



Method	**C-for-benefit** (Higher is better)	**Decision value RMST** (Higher is better)	**Calibration Slope** (Ideally 1)	**Calibration Intercept** (Ideally 0)
X-learner RF	0.60 [0.58 0.63]	1068.71 [1065.42 1072.08]	1.06 [0.74, 1.32]	0.00 [-0.01 0.00]
X-learner linear	0.54 [0.52 0.56]	1065.75 [1061.53 1069.49]	0.70 [0.30 1.12]	0.00 [-0.01 0.01]
Causal forest	0.55 [0.52 0.57]	1064.46 [1060.67 1068.06]	0.63 [0.26, 1.0]	0.00 [0.00 0.01]
Survival forest	0.53 [0.50 0.55]	1063.59 [1060.51 1066.71]	0.32 [0.04 0.57]	0.01 [0.01 0.02]
Cox regression	0.52 [0.50 0.55]	1061.09 [1056.42 1065.42	1.18 [0.55 1.81]	0.00 [-0.01 0.01]

References i

Burke, J. F., Hayward, R. A., Nelson, J. P., and Kent, D. M. (2014).
Using Internally Developed Risk Models to Assess Heterogeneity in Treatment Effects in Clinical Trials.
Circulation: Cardiovascular Quality and Outcomes, 7(1):163–169.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019).
Metalearners for estimating heterogeneous treatment effects using machine learning.
Proceedings of the National Academy of Sciences, page 201804597.

Patel, K. K., Arnold, S. V., Chan, P. S., Tang, Y., Pokharel, Y., Jones, P. G., and Spertus, J. A. (2017).
Personalizing the Intensity of Blood Pressure Control: Modeling the Heterogeneity of Risks and Benefits From SPRINT (Systolic Blood Pressure Intervention Trial).
Circulation: Cardiovascular Quality and Outcomes, 10(4):e003624.

References ii

- Schuler, A. and Shah, N. (2018).
  General-purpose validation and model selection when estimating individual treatment effects.
  Technical report.
  arXiv: 1804.05146.
- The ACCORD Study Group (2010).
  Effects of Intensive Blood-Pressure Control in Type 2 Diabetes Mellitus.
  New England Journal of Medicine, 362(17):1575–1585.
- The SPRINT Research Group (2015).
  A Randomized Trial of Intensive versus Standard Blood-Pressure Control.
  New England Journal of Medicine, 373(22):2103–2116.

References iii

- van Klaveren, D., Steyerberg, E. W., Serruys, P. W., and Kent, D. M. (2018). **The proposed ‘concordance-statistic for benefit’ provided a useful metric when modeling heterogeneous treatment effects.** *Journal of Clinical Epidemiology*, 94:59–68.

- Vock, D. M., Wolfson, J., Bandyopadhyay, S., Adomavicius, G., Johnson, P. E., Vazquez-Benitez, G., and O’Connor, P. J. (2016). **Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting.** *Journal of Biomedical Informatics*, 61:119–131.