Metalearners for Heterogeneous Treatment Effects on Survival Outcomes in Experiments

Yizhe Xu[1], Nikolaos Ignatiadis[2], Erik Sverdrup[3], Scott Fleming[1], Stefan Wager,[3], Nigam Shah[1]

[1] Stanford Center for Biomedical Informatics Research Institute, Stanford University; [2] Department of Statistics, Stanford University;
[3] Graduate School of Business, Stanford University; Correspondence: yizhex@stanford.edu

Background

Estimation of conditional average treatment effects (CATEs) plays a central role in modern medicine for informing treatment decision-making at a patient level. Metalearners estimate CATEs in an effective and flexible way by repurposing predictive machine learning models towards causal estimation. In this work, we provide concrete guidance for their application in the context of treatment heterogeneity estimation in randomized controlled trials with survival outcomes. The guidance we provide is supported by a comprehensive simulation study in which we vary the complexity of the underlying baseline risk and CATE functions, the magnitude of the heterogeneity in treatment, the censoring mechanism, and the balance in treatment assignment. Building upon our findings, we reanalyze the Systolic Blood Pressure Intervention Trial (SPRINT) and the Action to Control Cardiovascular Risk in Diabetes (ACCORD) trial: While recent literature suggests the existence of heterogeneous effects of intensive blood pressure treatment with multiple treatment effect modifiers, we argue that many of these modifiers may be spurious discoveries. This paper is accompanied by `survlearners`, an R package that provides well-documented implementations of the CATE estimation strategies described in this work.

Problem Setup and Assumptions

We discuss the problem of CATE estimation in Randomized controlled trials (RCTs) with right-censored data. Patient baseline covariates are denoted by $X_i \in \mathbb{R}^d$ and binary treatment $W_i \in \{0,1\}$. Let $T_i(w)$ be the potential survival time from subject i under treatment w, and C_i be the censoring time. $U_i = \min(T_i, C_i)$ is the observed follow-up time and $\Delta_i = \mathbb{I}(T_i < C_i)$ is the event indicator.

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x],$$

where $Y_i(w) = \mathbb{I}\{T_i(w) > t_0\}$ is the indicator of survival beyond time t_0.

Assumptions:

Consistency: The observed T_i is almost always the same as the potential outcome under w, i.e., $T_i = T_i(w)$

RCT: The treatment assignment is randomized, i.e., $W_i \perp \{X_i, T_i(1), T_i(0)\}$ and $\mathbb{P}(W_i = 1) = e$ with $0 < e < 1$.

Noninformative censoring: Censoring is independent of survival time given X_i and W_i, i.e., $C_i \perp T_i | X_i, W_i$

Positivity: There exists subjects who had events before the time horizon t_0, i.e., $\mathbb{P}(C_i > t_0 \mid X_i, W_i\} > \eta, \eta > 0$

Five State-of-the-art Metalearners

	Risk model	Censoring model	CATE model
S	$\widehat{\mu}(\cdot) = \mathcal{M}(Y \sim [X,W];\ \mathcal{O})$	not applicable	$\widehat{\tau}(x) = \widehat{\mu}([x,1]) - \widehat{\mu}([x,0])$
T	$\widehat{\mu}_{(1)}(\cdot) = \mathcal{M}(Y \sim X;\ \mathcal{O}_1)$ $\widehat{\mu}_{(0)}(\cdot) = \mathcal{M}(Y \sim X;\ \mathcal{O}_0)$		$\widehat{\tau}(x) = \widehat{\mu}_{(1)}(x) - \widehat{\mu}_{(0)}(x)$
X		$\widehat{S^C}(\cdot) = \mathcal{M}^{\text{oob}}(\mathbb{1}\{C \geq u\} \sim [X,W]; \mathcal{O})$ $\widehat{K} = 1/\widehat{S^C}(\min\{U, t_0\}, X, W)$	$Y^{*,X} = (1-W)(\widehat{\mu}_{(1)}(X) - Y) + W(Y - \widehat{\mu}_{(0)}(X))$ $\widehat{\tau}_{(1)}(x) = \mathcal{M}(Y^{*,X} \sim X;\ \mathcal{O}_1 \cap \mathcal{O}_{\text{comp}}, \widehat{K})$ $\widehat{\tau}_{(0)}(x) = \mathcal{M}(Y^{*,X} \sim X;\ \mathcal{O}_0 \cap \mathcal{O}_{\text{comp}}, \widehat{K})$ $\widehat{\tau}(x) = (1-e)\widehat{\tau}_{(1)}(x) + e\widehat{\tau}_{(0)}(x)$
R	$\widehat{\mu}_{(1)}(\cdot) = \mathcal{M}^{\text{oob}}(Y \sim X;\ \mathcal{O}_1)$ $\widehat{\mu}_{(0)}(\cdot) = \mathcal{M}^{\text{oob}}(Y \sim X;\ \mathcal{O}_0)$		$\widehat{m}(x) = e\widehat{\mu}_{(1)}(x) + (1-e)\widehat{\mu}_{(0)}(x)$ $Y^{*,R} = (Y - \widehat{m}(x))/(W-e)$ $\widehat{\tau}(x) = \mathcal{M}(Y^{*,R} \sim X;\ \mathcal{O}_{\text{comp}}, \widehat{K} \cdot (W-e)^2)$
M	not applicable		$Y^{*,M} = WY/e + (1-W)Y/(1-e)$ $\widehat{\tau}(x) = \mathcal{M}(Y^{*,M} \sim X;\ \mathcal{O}_{\text{comp}}, \widehat{K})$

TABLE 1: Overview of metalearners. S-learner: Modeling risk ($\hat{\mu}$) as a function of X_i, W_i, and their interactions; T-learner: Risk modeling stratified by treatment; M-learner: Metalearning by directly modeling treatment heterogeneity ($\hat{\tau}$) with inverse probability censoring weights ($\widehat{K}$); X- and R-learners: Modeling both risk and treatment heterogeneity.

Main Takeaways from the Simulation Study

	S-Learner	T-Learner	X-Learner	R-Learner	M-Learner
Requirements on Predictive Models					
A. Risk models					
a. Risk models under both treatment arms, $\mu_{(1)}$ and $\mu_{(0)}$, can be well-estimated	●	●	●	○	◌
	and	and	or		
b. When $e \ll 0.5$, the risk model $\mu_{(0)}$ can be well-estimated and vice versa	●	●	●	○	◌
B. CATE models can be well estimated	◌	◌	●	●	●
C. Censoring models can be well estimated	◌	◌	●	●	●
Recommendations on Metalearners					
A. Easy use of off-the-shelf predictive models	○	●	●	●	●
B. Allow directly imposing structural assumptions on CATE	○	○	●	●	●
C. Highly unstable CATE estimates	○	○	○	○	●

● Required / Yes ○ Not required / No ◌ Not applicable

Considerations

A. Risk models
- a. S- and T-learners: Requires sufficient numbers of treated and untreated subjects – bias-variance tradeoff
- b. X-learner: Only the risk model under the arm with the larger sample size matters under an unbalanced treatment assignment
- c. X- and R-learners: Robust to high variance of risk estimates

B. CATE models
- a. Often are simpler than risk models
- b. Prefer to be parsimonious for better interpretations
- c. S- and T-learners: May result in complex estimated CATE models due to small HTE or complicated risk functions

C. Censoring models
- a. All metalearners need to account for censoring
- b. Need to model censoring as a function of covariates under heterogeneous censoring

Conclusions

Our work provides guidance on ***when*** and ***how*** to apply each metalearner for time-to-event outcomes in light of the specific characteristics of a dataset, and we created the R package **survlearners** as an off-the-shelf tool to facilitate the implementation of our recommendations.

Funding Support: Supported by R01 HL144555 from the National Heart, Lung, and Blood Institute (NHLBI).

Simulation Results

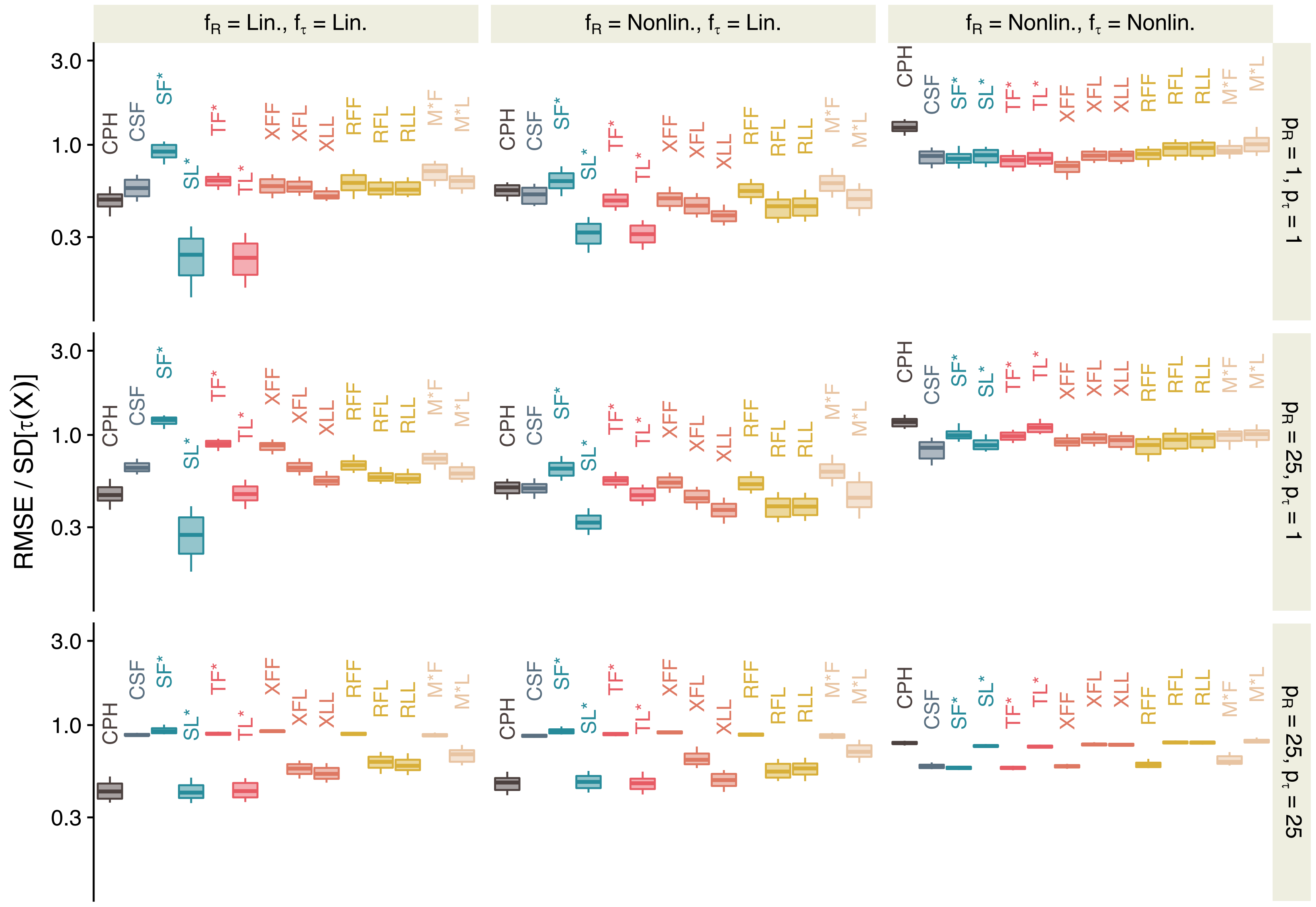

FIGURE 1: Rescaled root mean squared errors of metalearners. The function forms (f_R and f_τ) vary between linear (Lin) and nonlinear (Nonlin) across columns, and the numbers of predictors (P_R and P_τ) vary across rows. We use 3-letter acronyms for each method, wherein the first letter corresponds to the meta-learner, the second to the risk model, and the third to the CATE model.

Case Study on SPRINT and ACCORD-BP

- SPRINT and ACCORD-BP are two large randomized controlled trials (RCTs) that compare the effectiveness of intensive blood pressure therapy (< 120 mm Hg).
- Prior works found significant treatment heterogeneity and effect modifiers, but we argue that they are probably false discoveries
- We conduct a sharp null analysis on both RCTs and show some metalearners (e.g., T-learner) yield large CATEs when they are zero, by design
- We also perform a CATE analysis based on our learning from the simulation study to show a lack of evidence of treatment heterogeneity.

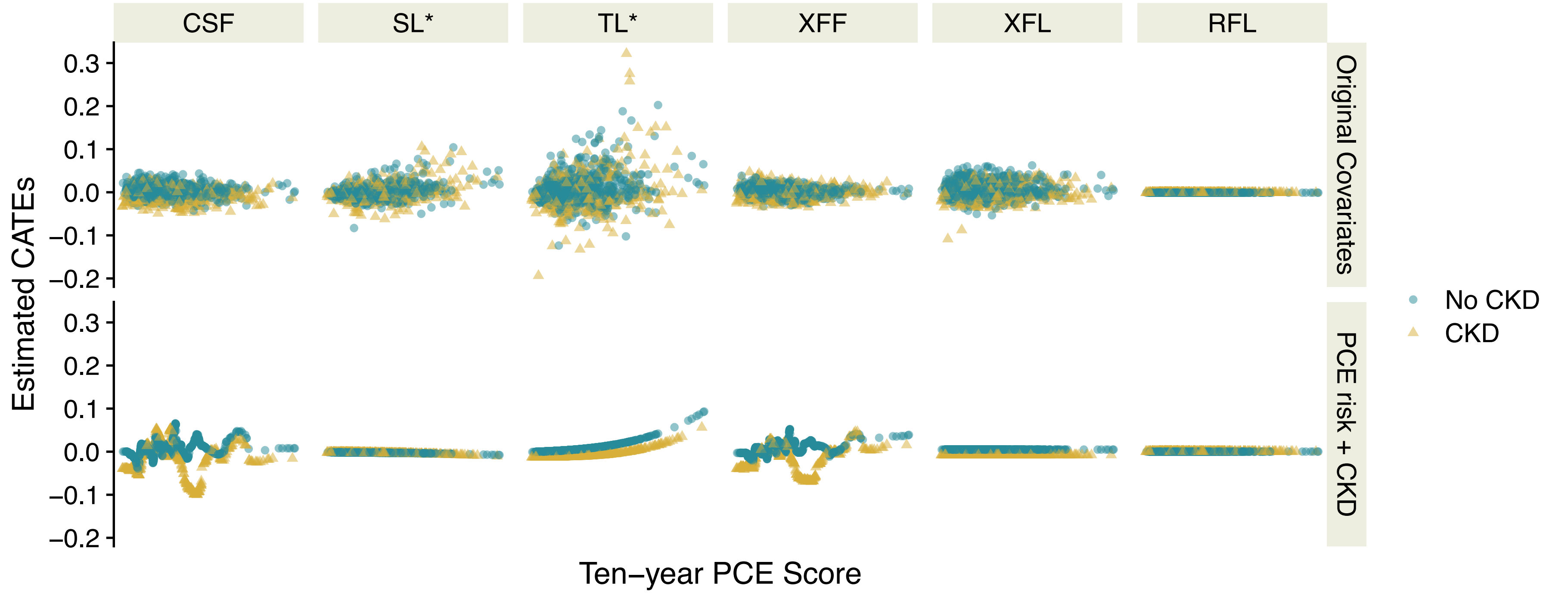

FIGURE 2: Sharp null analysis on untreated participants in SPRINT. An artificial randomized treatment assignment is created. The censoring weights are estimated using a survival forest model. The analysis was replicated with the original covariates in SPRINT (Row 1) and the estimated ten-year CVD risk (using pooled cohort equation) and subclinical CKD as the predictors (Row 2).