SurvHTE-Bench: A Benchmark for Heterogeneous Treatment Effect Estimation in Survival Analysis

[1] Shahriar Noroozizadeh*, [1] Xiaobin Shen*,
[2] Jeremy C. Weiss, [1] George H. Chen

[1] Carnegie Mellon University
[2] National Institutes of Health
* Equal contribution, listed alphabetically.

Accepted at ICLR 2026

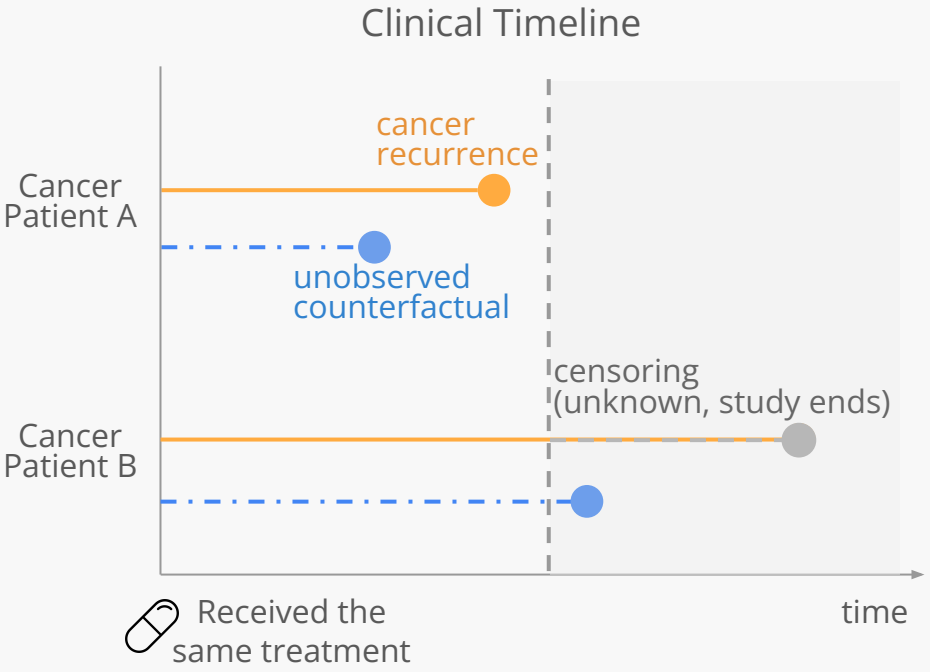
April 2026

Carnegie Mellon University

NIH

Paper (arXiv)

Heterogeneous Treatment Effects (HTE) Matter.
Survival Data Makes It Harder.



Unobserved counterfactuals

+

Confounding

+

Censoring

Classic causal challenge

Survival-specific challenge

A Growing Field. But No Shared Benchmark.

Growing method landscape

Incompatible evaluation setups

✖ No fair cross-paper comparison

⇩

SurvHTE-Bench

Examples from recent papers

Incompatible evaluation settings

Recent Paper	Causal Setup	Survival Model	Censoring Rate	Estimand
Bo et al. (2024)	RCT only	Weibull	<30%	Survival Probability
Curth et al. (2021)	Confounding, informative censoring	Discrete model	Varies	RMST, Survival Probability
Cui et al. (2023)	Ignorability holds	Cox, Poisson	~50%	RMST
Meir et al. (2025)	Ignorability holds	AFT, Cox, Poisson	30%-70%	RMST

SurvHTE-Bench: A Unified Benchmark Framework

52 Diverse Datasets

40
Synthetic
Controlled stress tests & assumption violations

10
Semi-Synthetic
Real covariates + simulated outcomes

2
Real-World
Known ground truth + real-world stress tests

Causal Estimands

RMST
(restricted mean survival time)
Survival probability
both with different horizons

53 Evaluated Methods

Outcome Imputation Methods
Imputed censored outcomes + standard CATE learners

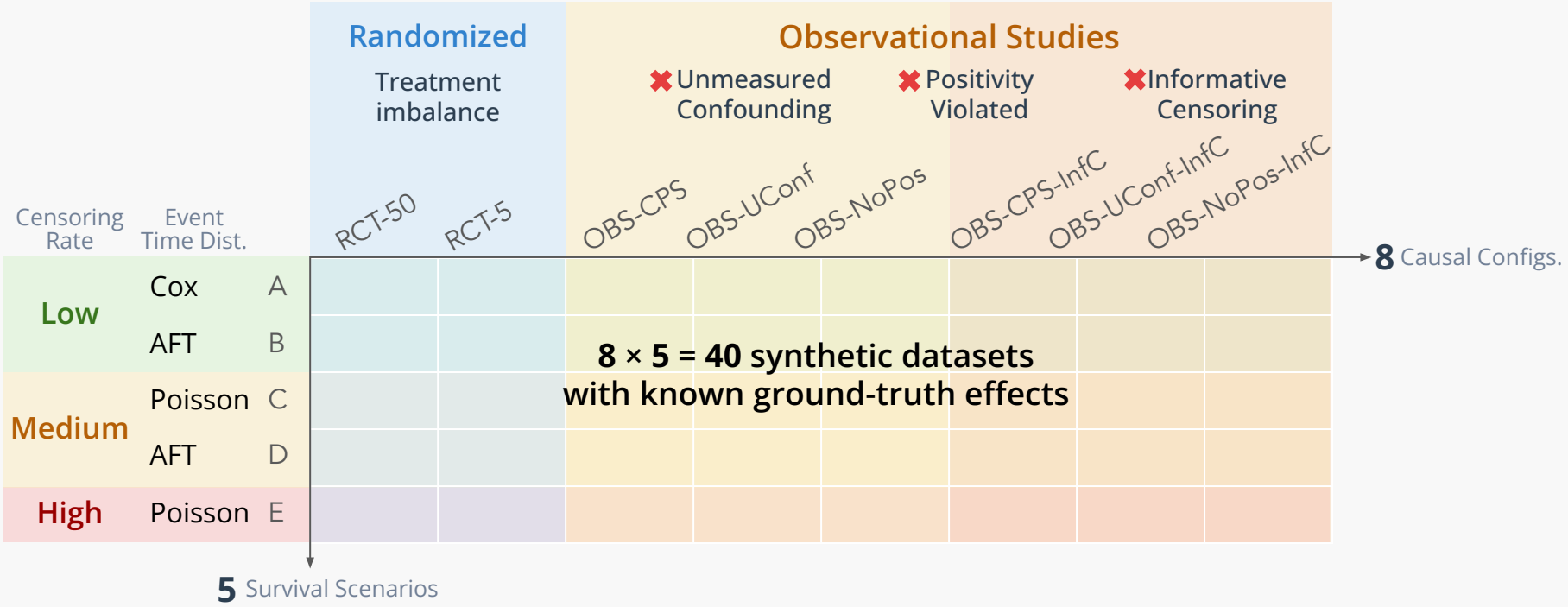
Direct-Survival CATE methods
SurvITE, Causal Survival Forests

Survival Meta-Learners
S-, T-, Matching + survival models

Evaluation Metrics

CATE RMSE
(root mean squared error vs. known ground truth)
Borda count rankings
Auxiliary: ATE bias, method-specific diagnostics

Synthetic Benchmark Design. Controlled Stress Tests.



No Single Method Dominates.

Bridging to Reality: Stress-Testing the Methods

52 Diverse Datasets

40	10	2
Synthetic	**Semi-Synthetic**	**Real-World**
Controlled stress tests & assumption violations	Real covariates + simulated outcomes	Known ground truth + real-world stress tests

Causal Estimands

RMST
(restricted mean survival time)
Survival probability
both with different horizons

53 Evaluated Methods

➔ **10 semi-synthetic datasets: (1 ACTG HIV, 9 MIMIC-IV)**

***Retained Control*:** Outcomes simulated to have ground-truth CATE and treatment assignment known

Real-World Challenge: Methods evaluated against high-dimensional, highly correlated real clinical covariates

Evaluation Metrics

CATE RMSE
(root mean squared error vs. known ground truth)

Auxiliary: ATE bias, method-specific diagnostics

Bridging to Reality: Stress-Testing the Methods

52 Diverse Datasets

40	10	2
Synthetic	**Semi-Synthetic**	**Real-World**
Controlled stress tests & assumption violations	Real covariates + simulated outcomes	Known ground truth + real-world stress tests

Causal Estimands

RMST
(restricted mean survival time)
Survival probability
both with different horizons

53 Evaluated Methods

➔ **2 real-world datasets: (ACTG HIV, Twins Dataset)**

The Unknowable: True counterfactuals are missing; no ground-truth CATE exists.

Evaluation Strategy 1 (ACTG HIV): Inject artificial censoring to measure estimator stability across varying dropout rates.

Evaluation Strategy 2 (Twins Dataset): Leverage the twin as a strong proxy for the unobserved counterfactual to benchmark accuracy.

Evaluation Metrics

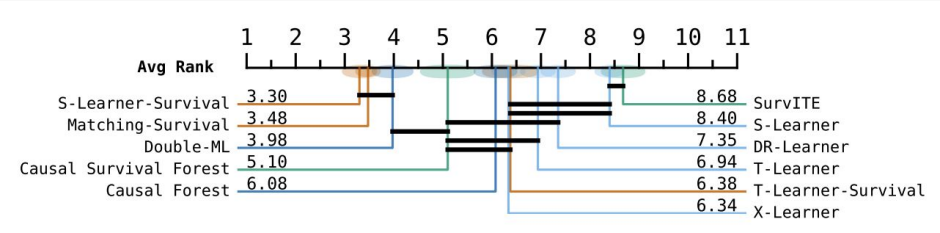
CATE RMSE
(root mean squared error vs. known ground truth)
Stability across censoring rates

Auxiliary: Proxy ATE bias, Proxy CATE RMSE

What SurvHTE-BENCH Reveals

Fair comparison

No single method dominates across settings



Reproducible benchmark

A shared measuring stick for survival HTE

Code

github.com/Shahriarnz14/SurvHTE-Bench

Extensible by design

Easy to extend with new methods & datasets

Datasets

huggingface.co/datasets/snoroozi/SurvHTE-Bench

Funding: DIR of NLM, NIH; NSF Career Award; NSERC Doctoral Fellowship; ORISE, CMU TCS Presidential Fellowship