SURVHTE-BENCH: A BENCHMARK FOR HETEROGENEOUS TREATMENT EFFECT ESTIMATION IN SURVIVAL ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating heterogeneous treatment effects (HTEs) from right-censored survival data is critical in high-stakes applications such as precision medicine and individualized policy-making. Yet, the survival analysis setting poses unique challenges for HTE estimation due to censoring, unobserved counterfactuals, and complex identification assumptions. Despite recent advances, from causal survival forests to survival meta-learners and outcome imputation approaches, evaluation practices remain fragmented and inconsistent. We introduce SURVHTE-BENCH, the first comprehensive benchmark for HTE estimation with censored outcomes. The benchmark spans (i) a modular suite of synthetic datasets with known ground truth, systematically varying causal assumptions and survival dynamics, (ii) semi-synthetic datasets that pair real-world covariates with simulated treatments and outcomes, and (iii) real-world datasets from a twin study (with known ground truth) and from an HIV clinical trial. Across synthetic, semi-synthetic, and real-world settings, we provide the first rigorous comparison of survival HTE methods under diverse conditions and realistic assumption violations. SURVHTE-BENCH establishes a foundation for fair, reproducible, and extensible evaluation of causal survival methods. The data and code of our benchmark are anonymously available at: `https://anonymous.4open.science/r/SurvHTE-Benchmark-206B`.

1 INTRODUCTION

In many causal inference applications where we aim to quantify how well a treatment works, estimating *heterogeneous treatment effects* (HTEs) could be more useful than only estimating population-level *average treatment effects* (ATEs), building on the intuition that the same treatment can vary in effectiveness when given to different individuals. In survival analysis with right-censored outcomes (common in clinical trials and electronic health records), estimating HTEs can be especially challenging. In addition to the standard difficulties of causal inference (unobserved counterfactuals, confounding), the analyst must account for censoring, where the event of interest is only observed for a subset of subjects. These features complicate identification and estimation, yet they are central in high-stakes applications such as precision medicine and individualized policy-making (Zhu & Gallego, 2020; Chapfuwa et al., 2021; Curth et al., 2021).

Recent years have seen a growing set of causal survival methods, including deep representation-learning approaches (Chapfuwa et al., 2021; Curth et al., 2021), causal survival forests (Cui et al., 2023), survival meta-learners (Bo et al., 2024; Noroozizadeh et al., 2025), and outcome imputation methods (Xu et al., 2024; Meir et al., 2025). Despite methodological advancement, no standardized benchmark exists, limiting reproducibility and fair comparisons. Most studies rely on bespoke simulations or limited real datasets with unknown ground truth, with differing levels of censoring, survival distributions, and causal assumptions. As a result, comparisons are not standardized, robustness of different proposed methods is unclear, and progress is difficult to measure.

We introduce SURVHTE-BENCH, the first comprehensive benchmark for HTE estimation in right-censored survival data. Our contributions are as follows:

- **Method unification:** We categorize existing survival HTE methods (and natural extensions of existing such methods that technically have not previously been published) into three broad families: outcome imputation methods, direct-survival models, and survival meta-learners. We provide a

modular implementation of 52 methods among these three families. This is the first systematic framework unifying survival HTE methods that facilitates reproducibility and extensibility.

- **Synthetic benchmark design:** We present a curated suite of 40 synthetic datasets spanning eight causal configurations (with different combinations of randomization, unobserved confounding, overlap violation, informative censoring) crossed with five survival scenarios (with different survival and censoring distributions), yielding controlled settings with known ground-truth HTEs under realistic assumption violations.
- **Semi-synthetic and real data:** We also include 6 semi-synthetic datasets from existing literature (real covariates with simulated treatments and outcomes) that aim to be more realistic compared to purely synthetic datasets while still having ground truth on HTEs. We further include 2 widely studied real datasets: the Twins dataset that has known ground truth (Almond et al., 2005) (i.e., per twin, one has the treatment and the other does not, so that we observe both counterfactual outcomes), and the HIV clinical trial dataset without known ground truth (Hammer et al., 1996).
- **Comprehensive evaluation:** We compare representative estimators across all settings. Our results show that no single method dominates: performance depends on causal assumptions, censoring, and survival dynamics. Notably, S-learners among survival meta-learners demonstrate robustness under severe violations and high censoring.

While prior work has explored subsets of these design choices (e.g., Cui et al. (2023); Meir et al. (2025)), SURVHTE-BENCH is the first to systematically evaluate survival HTE methods under assumption violations, diverse survival models, and across synthetic, semi-synthetic, and real data. We focus on binary treatments and static covariates with right-censored outcomes, as even this basic setting lacks a standardized benchmark. More complex extensions (time-varying treatments, longitudinal covariates, and instrumental variables) are beyond our present scope.

2 BACKGROUND AND RELATED WORK

We briefly review the problem setup, identification assumptions, existing evaluation practices, and the three families of survival HTE estimators.

Problem setup. For each unit (data point) i, we observe covariates $X_i \in \mathcal{X}$, a binary treatment $W_i \in \{0, 1\}$, and a possibly censored event time $\widetilde{T}_i = \min(T_i, C_i)$ with event indicator $\delta_i = \mathbb{1}\{T_i \leq C_i\}$, where δ_i is 1 if the event of interest happened (in which case T_i is the event time) or 0 if the outcome is censored (in which case T_i is the censoring time). Using the standard potential outcomes framework, $T_i(w)$ denotes the potential event time under treatment $w \in \{0, 1\}$ with $T_i = T_i(W_i)$. We assume that the tuple $(X_i, W_i, T_i(0), T_i(1), C_i)$ is i.i.d. across different i.

We aim to estimate the *conditional average treatment effect* (CATE) with respect to a transformation of the event time $y(\cdot)$:

$$\tau(x) := \mathbb{E}\big[y\big(T_i(1)\big) - y\big(T_i(0)\big)|X_i = x\big], \tag{1}$$

where $y(\cdot)$ encodes the survival estimand of interest, and the expectation is taken over the randomness of the two potential outcomes. For example, if we want the survival estimand to be the restricted mean survival time (RMST) up to a user-specified time horizon $h > 0$, then we would set $y(t) := \min\{t, h\}$. Other choices for estimands are also possible (e.g., median survival time, survival probability at a fixed time). In this paper, we focus on RMST, which is interpretable, robust under censoring, and widely adopted (Shen et al., 2018; Curth et al., 2021; Cui et al., 2023), while noting that our benchmark design allows extensions to other estimands.

Identification assumptions. Identification of $\tau(x)$ relies on the following assumptions (Cui et al., 2023) (and in our benchmark, we vary whether these get violated):

- (A1) *Consistency*: $T_i = T_i(W_i)$ almost surely.
- (A2) *Ignorability*: $\{T_i(0), T_i(1)\} \perp W_i \mid X_i$.
- (A3) *Positivity*: $\eta_e \leq \mathbb{P}(W_i = 1|X_i = x) \leq 1 - \eta_e$ for some $\eta_e > 0$.
- (A4) *Ignorable censoring*: $T_i \perp\!\!\!\perp C_i \mid X_i, W_i$.
- (A5) *Censoring positivity*: For horizon h, $\mathbb{P}(C_i < h|X_i, W_i) \leq 1 - \eta_C$ for some $0 < \eta_C \leq 1$.

Violations are common: unmeasured prognostic factors break ignorability, treatment guidelines break positivity, and drop-out linked to prognosis induces informative censoring. A central goal of SURVHTE-BENCH is to measure how estimators behave under such violations.

Existing evaluation practice. Because only one potential outcome is observed per unit, validation typically relies on author-specific simulations. Prior studies vary assumptions in narrow ways: e.g.,

censoring up to 30% (Bo et al., 2024) or heavy censoring but assuming ignorability (Meir et al., 2025). Consequently, results are not comparable across papers, and estimator robustness under simultaneous assumption violations remains unclear. To date, no public benchmark exists with known individual-level ground truth with varying levels of assumption violations and survival distributions.

Overview of existing survival HTE estimators. We group existing methods into three families:

- *Outcome imputation methods* (Xu et al., 2024; Meir et al., 2025): Replace censored times with imputed survival times (e.g., IPCW-based reweighting introduced in Qi et al. (2023)). Then apply standard CATE estimators such as causal forests (Athey et al., 2019), double ML (Chernozhukov et al., 2018), or meta-learners including S(ingle)-, T(wo)-, X(cross)-, D(oubly)R(obust)-learners (Athey & Imbens, 2015; Künzel et al., 2019; Kennedy, 2023).[1]
- *Direct-survival CATE models:* Extend causal inference directly to time-to-event outcomes, e.g., targeted learning (Van der Laan & Rose, 2011), targeted maximum likelihood estimation (Stitelman & van der Laan, 2010; Stitelman et al., 2011), tree-based estimators (Zhang et al., 2017), SurvITE (Curth et al., 2021), Bayesian approaches (Henderson et al., 2020), or causal survival forests (Cui et al., 2023).
- *Survival meta-learners* (Xu et al., 2023; Bo et al., 2024; Noroozizadeh et al., 2025): Adapt S(ingle)-, T(wo)-, or matching-learners to survival outcomes by using survival models such as random survival forests or deep survival models.

While these approaches appear in disparate papers, we are the first to categorize them into these three families, and we implement 52 methods within these families in a unified, modular framework.

3 SURVHTE-BENCH

SURVHTE-BENCH probes how survival CATE estimators behave when assumptions (A1)–(A5) hold and when they are either mildly or severely violated. As real data with ground-truth CATEs are scarce, the bulk of our benchmark relies on synthetic datasets. We also include semi-synthetic data (real covariates with simulated treatments and outcomes) and two real-world datasets. As already stated in Section 2, in this paper we focus on the case where the target estimand is RMST up to a user-specified time horizon h (other estimands are possible but we defer this to future work).

Synthetic data. We construct a modular suite of 40 synthetic datasets that systematically vary across two orthogonal axes: (1) causal configuration: treatment mechanism, positivity, confounding, censoring mechanism; (2) survival scenario: event-time distribution and censoring rate. Crossing 8 causal configurations with 5 survival scenarios yields $8 \times 5 = 40$ synthetic datasets, each with binary, time-fixed treatment, five independently sampled covariates each distributed as Uniform$(0, 1)$, and up to 50,000 units. For each unit i, we generate both $T_i(0)$ and $T_i(1)$, ensuring that ground-truth CATEs are always known.

Table 1: Causal configurations of synthetic datasets. `RCT` = randomized controlled trial; `OBS` = observational study; `50(5)` = 50%(5%) treatment rate; `CPS`= correct specified propensity score (ignorability satisfied); `UConf` = unobserved confounding (ignorability violated); `NoPos` = lack of positivity; `InfC` = informative censoring (ignorable censoring violated). ✓= assumption holds, ✗= violated.

Causal Configs.	RCT	Ignorability	Positivity	Ignorable Censoring
`RCT-50`	✓	✓	✓	✓
`RCT-5`	✓	✓	✓	✓
`OBS-CPS`	✗	✓	✓	✓
`OBS-UConf`	✗	✗	✓	✓
`OBS-NoPos`	✗	✓	✗	✓
`OBS-CPS-InfC`	✗	✓	✓	✗
`OBS-UConf-InfC`	✗	✗	✓	✗
`OBS-NoPos-InfC`	✗	✓	✗	✗

The 8 **causal configurations** (Table 1) include randomized controlled trials (`RCT-50`, `RCT-5`) and observational studies with correctly specified propensity scores (i.e., these are known during training) with all confounders observed in estimation (`OBS-CPS`), unobserved confounding (`OBS-UConf`), or lack of positivity (`OBS-NoPos`). Each observational setting has variants with suffix "`-InfC`", where ignorable censoring is replaced by informative censoring, where censoring times depend stochastically on event times. These violations reflect common real-world challenges: unmeasured risk factors in treatment decisions (violating ignorability), treatment imbalance in observational studies (violating positivity), and dropout mechanisms correlated with health outcomes (violating ignorable censoring). We do not model interference (consistency

[1] Standard CATE estimators do not handle censoring. By imputing censored times with survival times as a preprocessing step, we make it appear as if there is no censoring, so standard CATE estimators can be applied.

violations) or censoring-positivity violations, which require specialized designs beyond our scope. Additional variations, such as informative censoring with the censoring time driven by unobserved factors, are included in the Appendix I to illustrate the extensibility of our modular setup.

The 5 **survival scenarios** (Table 2) include Cox proportional hazards (low censoring), accelerated failure time (AFT) models (low and high censoring), and Poisson hazards (medium and high censoring). These distributions cover proportional hazards (Cox) and non-proportional hazards (AFT[2], Poisson), with censoring levels ranging from under 30% to over 70%. This variety reflects practical challenges like high censoring common in EHR cohorts, accelerated processes in oncology, and discrete hazard approximations in epidemiology. Within each survival scenario, coefficients are tuned so that event times are comparable across different causal configurations. Full generation formulas and summary statistics (e.g., censoring rate, treatment rate, ATE) for each dataset are in Appendix A.

Table 2: Survival scenarios of synthetic datasets. "Low" <30%, "Med" 30-70%, "High" >70% censoring. AFT = accelerated failure time.

Survival Scenario	Survival Time Distribution	Censoring Rate
A	Cox	Low
B	AFT	Low
C	Poisson	Med
D	AFT	High
E	Poisson	High

Evaluation metrics. Per dataset, averaged over 10 random splits, we report:

- CATE root mean square error (RMSE): $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\tau}(X_i) - \tau(X_i))^2}$.
- ATE bias: $\frac{1}{n}\sum_{i=1}^{n}\hat{\tau}(X_i) - \Delta$, where Δ is the true ATE from the population and can be approximated using the average CATE from a very large sample (i.e., from 50,000 simulated samples).
- Auxiliary imputation accuracy: mean absolute error (MAE) between imputed and true event times.
- Auxiliary regression/survival fit: MAE for regression-based learners, AUC for propensity score models, and the time-dependent C-index (Antolini et al., 2005) for survival models.

Survival CATE methods implemented. We evaluate the three broad families of survival CATE methods (52 variants total; see Appendix C for the full list, Appendix D for methodological details):

- *Outcome imputation methods*: meta learners (S-, T-, X-, DR-Learners) paired with base regression learners (lasso, random forest, XGBoost), plus double ML and causal forest, each combined with the three imputations (Pseudo-obs, Margin, and IPCW-T (Qi et al., 2023), see Appendix B for details). In total, we implement 42 variants.
- *Direct-survival CATE models*: We include the canonical causal survival forest.
- *Survival meta-learners*: S-, T-, and matching-learners paired with survival learners (Random Survival Forest (Ishwaran et al., 2008), DeepSurv (Katzman et al., 2018), and DeepHit (Lee et al., 2018)), for a total of $3 \times 3 = 9$ variants.

Note that some implemented methods are straightforward extensions of existing ideas despite previously not being published. For example, (Qi et al., 2023) suggested ways of replacing censoring times with imputed survival times for the purposes of model evaluation, but their imputation strategies naturally can be coupled with standard CATE learners to obtain survival CATE estimators. Similarly, pairing meta-learners with different base learners (e.g., lasso regression, XGBoost, or DeepSurv) yields natural yet previously unpublished variants.

Semi-synthetic data. We include 6 semi-synthetic datasets from prior work, pairing real covariates (ACTG HIV trial, MIMIC-IV ICU records) with simulated treatments and outcomes, covering moderate to extreme censoring regimes. These datasets preserve realistic feature distributions while retaining ground-truth CATEs. Details are in Section 4.2.

Real data. Finally, we incorporate two real datasets, one with ground truth (for which we can use the same evaluation metrics as with synthetic data) and one without ground truth but with a low censoring rate (for which we compare how models perform on the original dataset vs on the dataset with artificially introduced censoring). These provide opportunities to evaluate how methods behave under real covariate and outcome structures. Details are in Section 4.3.

4 Benchmarking Results

We now present benchmark results across synthetic, semi-synthetic, and real data, spanning controlled violations of causal assumptions to realistic covariate structures.

[2]The AFT noise distribution we use (that is additive in log survival time) is Gaussian so that the resulting model does *not* satisfy the proportional hazards assumption (which would require the noise to be Gumbel).

4.1 Synthetic Experiment Results and Analyses

We begin with synthetic datasets, where we evaluate 52 estimator variants across the 40 synthetic datasets (Section 3), systematically spanning varying causal configurations and survival scenarios. This controlled setting enables us to probe estimator robustness under systematic violations of identification assumptions. Our analyses aim to address four questions: **(Q1)** Which estimators perform best overall in terms of CATE RMSE and ATE bias? **(Q2)** How do violations of causal assumptions (ignorability, positivity, ignorable censoring) affect performance? **(Q3)** How does the censoring rate influence estimation quality? **(Q4)** How do component choices (imputation algorithms and base learners) affect final CATE accuracy?

Evaluation protocol. Per synthetic dataset, we conduct experiments with a random selection of 5,000, 2,500, and 2,500 points for training, validation, and testing samples, repeated over 10 random splits. The validation set is used for selecting the best variant within each method family, while test sets are reserved strictly for evaluation. Additional convergence analyses with varying training set sizes are in Appendix F.7. Across all experiments, the horizon parameter h is set to the maximum observed time in each dataset, which is a common practice that allows for consistent estimation of the restricted mean survival time over the entire observed period. Further experimental details, including hyperparameters, are in Appendix E.

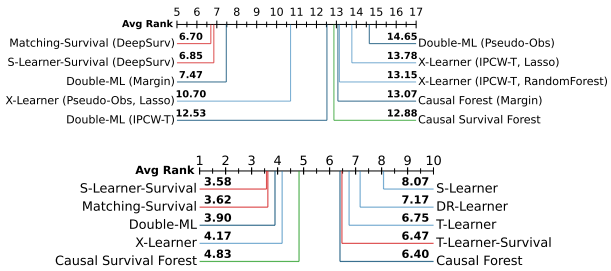


Figure 1: (top) Borda count rankings of the top 10 estimator variants (out of 52 total), based on CATE RMSE across 40 datasets and averaged over 10 repeats (lower is better). (bottom) Family-level rankings, where for each dataset the best method variant within each method family is chosen using validation performance and then ranked on the held-out test set.

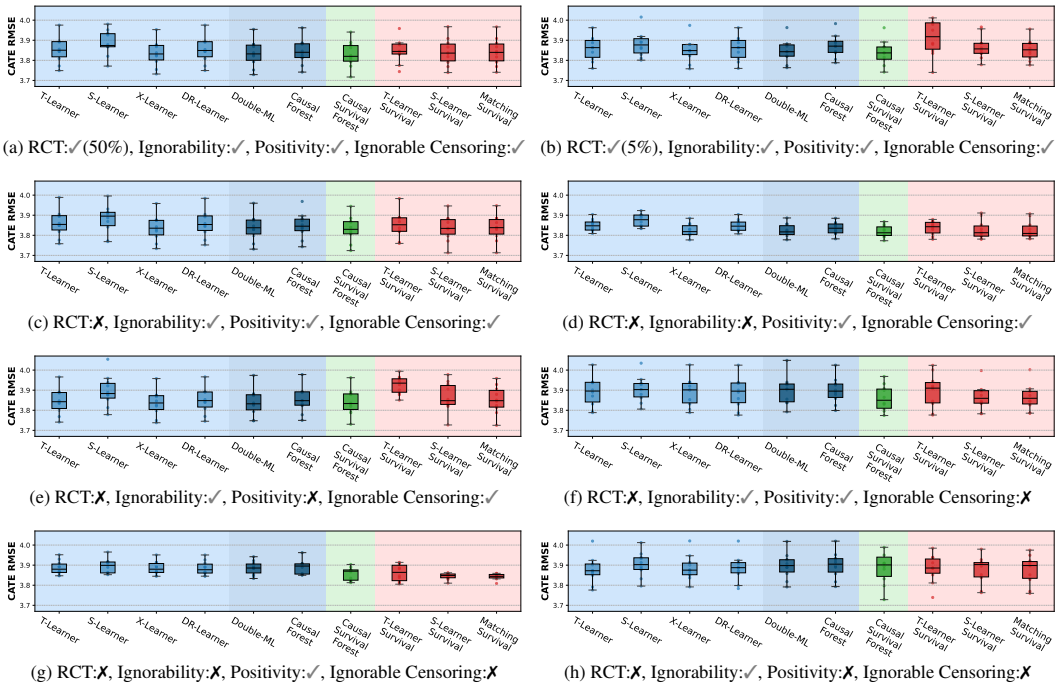
We present results using the following visualizations:

- **Borda count rankings.** To provide a clear summary across the diverse experimental settings, we adopt the Borda count method, which ranks methods by CATE RMSE in each dataset (lower is better) and then averages the ranks across datasets. This approach yields a single, interpretable score that reflects overall relative performance while accounting for variability across scenarios. Similar strategies have been used in other benchmarking studies (e.g., Han et al. 2022) to enable transparent comparisons across heterogeneous tasks. We report rankings at two levels: (i) individual estimator variants (52 total; Figure 1, top), and (ii) aggregated method families, where the best variant per family is selected on validation data (10 total; Figure 1, bottom). The latter mimics a practical deployment setting where practitioners would tune and select the strongest model within a family. More granular rankings stratified by survival scenario (Figure 5) and by causal configuration (Figure 6) are provided in Appendix F.3.
- **CATE RMSE.** We report absolute CATE RMSE across 10 repeats, grouped by survival scenario, with one panel per set of eight causal configurations. In the main paper, we show Scenario `C` as an illustrative example (Figure 2); results for the other scenarios are deferred to Appendix F.4.
- **ATE bias.** We report ATE bias results, computed analogously to CATE RMSE, in Appendix F.5. While the focus of this benchmark is on CATE estimation, these serve as a complementary check.

Additionally, in Appendix F.6, we report a series of auxiliary evaluations of key components, including imputation error (Appendix F.6.1) for imputation-based methods and regression model accuracy (Appendix F.6.2) or survival model performance (Appendix F.6.3) for meta-learners. These results establish how component-level performance relates to downstream CATE estimation.

Key findings. Overall, performance is strongly context-dependent. For example, in low-censoring randomized settings, outcome imputation methods such as X-Learner and Double-ML excel. As censoring intensifies or when assumptions are violated, survival meta-learners and Causal Survival Forests gain a clear advantage. Within method families, the choice of imputation algorithm or survival base model critically determines outcomes. We summarize detailed findings below.

Overall performance (Q1). Figure 1 (top) presents the Borda count rankings of the top-10 performing methods out of the 52 total configurations evaluated (full ranking in Appendix F.1). The



(a) RCT:✓(50%), Ignorability:✓, Positivity:✓, Ignorable Censoring:✓

(b) RCT:✓(5%), Ignorability:✓, Positivity:✓, Ignorable Censoring:✓

(c) RCT:✗, Ignorability:✓, Positivity:✓, Ignorable Censoring:✓

(d) RCT:✗, Ignorability:✗, Positivity:✓, Ignorable Censoring:✓

(e) RCT:✗, Ignorability:✓, Positivity:✗, Ignorable Censoring:✓

(f) RCT:✗, Ignorability:✓, Positivity:✓, Ignorable Censoring:✗

(g) RCT:✗, Ignorability:✗, Positivity:✓, Ignorable Censoring:✗

(h) RCT:✗, Ignorability:✓, Positivity:✗, Ignorable Censoring:✗

Figure 2: CATE RMSE in Scenario C across 10 experimental repeats.

highest-performing estimators are survival meta-learners built on DeepSurv, with Matching-Survival (6.70 out of 52) and S-Learner-Survival (6.85) leading, followed by Double-ML with Margin imputation (7.47). Among outcome imputation approaches, Margin appears most frequently in the top performers, though Pseudo-obs and IPCW-T are also represented.

At the method family level (Figure 2 for Scenario C and Figure 1 (bottom) across all causal configurations and survival scenarios), we see how each approach performs when optimally configured. At this level, S-Learner-Survival (3.58 out of 10) and Matching-Survival (3.62) maintain their advantage, followed by Double-ML (3.90) and X-Learner (4.17).

Violations of causal assumptions (Q2). Performance shifts substantially depending on assumption violations (Figure 6). In randomized balanced trials (RCT-50), outcome imputation methods dominate (X-Learner, 3.20; Double-ML, 3.40), but with imbalanced treatment (RCT-5), Double-ML remains strong (2.20) while T-Learner-Survival, which relies on fitting base models on the treated units, drops to last place (8.60) due to sparse treated units.

Under ignorability violation (OBS-UConf, Figure 6,d), Double-ML is the only competitive imputation method, whereas survival meta-learners and Causal Survival Forest retain relatively stable performance. Examining ATE bias (Figures 12-16,d), we see that across all scenarios, survival meta-learners and Causal Survival Forest methods maintain relatively consistent bias levels despite ignorability violations, whereas survival meta-learners often exhibit a slightly increased bias.

Under positivity violation (OBS-NoPos, Figure 6,e), we see the more sophisticated outcome imputation approaches like X-Learner and Double-ML maintain strong performance and outperform survival meta-learners. However, when positivity violations occur alongside other violations (Figure 6,h), survival meta-learners regain their advantage, demonstrating their robustness to multiple simultaneous violations. Causal Survival Forest sees a large drop in its ranking, suggesting its limited robustness to regions of covariate space with deterministic treatment assignment.

Under informative censoring (InfC, Figure 2,f-h), survival meta-learners and causal survival forest continue to outperform outcome imputation approaches. However, all methods show degraded performance compared to their ignorable censoring counterparts with higher CATE RMSE variability.

Impact of censoring rate (Q3). For the impact of censoring rate and survival time distribution (Figure 5), in low-censoring Scenario A, Double-ML and X-Learner lead the rankings, but as censoring

increases through Scenarios B to E, survival meta-learners and causal survival forest progressively move to the top. By Scenario D (high censoring), S-Learner-Survival (1.62) and Matching-Survival (2.38) dramatically outperform all other approaches. This pattern suggests that direct survival modeling provides increasing advantages as censoring rates rise, likely due to better handling of the uncertainty in heavily censored data compared to outcome imputation approaches.

Separately, in Appendix F.5, we show ATE bias across different datasets. We observe apparent divergence of the estimated ATE from the true ATE in Scenarios D and E (Figure 15, 16), where the censoring rate is very high. Especially when the true underlying event time follows an AFT distribution (Scenario D), almost all estimators failed under all different causal configurations, suggesting the challenging task of treatment effect estimation under a high censoring rate.

Component effects on CATE estimation (Q4). Auxiliary evaluations in Appendix F.6 demonstrate that both imputation accuracy and base learner performance influence downstream CATE estimation. Among outcome imputation methods, Margin consistently achieves the lowest imputation error and degrades the least under heavy censoring (Appendix F.6.1), which translates into Margin-based variants appearing more frequently among the top-ranked estimators (Figure 1). For survival meta-learners, higher concordance indices of DeepSurv across survival scenarios (Appendix F.6.3) explain why DeepSurv-based configurations dominate overall rankings.

4.2 Semi-synthetic data results

To bridge the gap between controlled synthetic experiments and real-world complexity, we evaluate methods on semi-synthetic datasets that pair real covariate distributions with simulated treatments and outcomes. This approach addresses a critical limitation of purely synthetic data—the potential lack of representativeness in covariate structures—while maintaining ground-truth CATEs for rigorous evaluation. These datasets preserve real-world covariate correlations, mixed data types, and high dimensionality while enabling controlled evaluation against known treatment effects.

Dataset construction. We construct 6 semi-synthetic datasets with real covariates from two sources:

- **ACTG semi-synthetic**: Based on 23 covariates from the ACTG HIV clinical trial (Hammer et al., 1996), with treatment and event times simulated following Chapfuwa et al. (2021). This dataset captures moderate censoring (51%) with realistic treatment imbalance.
- **MIMIC semi-synthetic**: Derived from 36 covariates in the MIMIC-IV ICU database (Johnson et al., 2023), with treatment and outcomes generated following Meir et al. (2025). We create five variants with censoring rates from 53% to 88%, simulating the range from moderate to extreme censoring common in longitudinal EHR studies.

Table 3 presents CATE RMSE results across our semi-synthetic datasets, revealing how realistic covariate structures modulate the core performance patterns observed in synthetic experiments.

Table 3: CATE RMSE on semi-synthetic datasets across 10 experimental repeats. Best two methods per dataset are **bolded**.

Method Family (censoring rate)	**ACTG** (51%)	**MIMIC-i** (88%)	**MIMIC-ii** (82%)	**MIMIC-iii** (74%)	**MIMIC-iv** (66%)	**MIMIC-v** (53%)
Outcome Imputation Methods						
T-Learner	11.257 ± 0.239	7.964 ± 0.046	**7.912 ± 0.046**	7.915 ± 0.043	7.912 ± 0.043	7.908 ± 0.043
S-Learner	11.300 ± 0.221	7.977 ± 0.044	7.968 ± 0.047	7.956 ± 0.050	7.959 ± 0.046	7.958 ± 0.048
X-Learner	**11.072 ± 0.196**	7.964 ± 0.046	**7.912 ± 0.046**	7.915 ± 0.043	7.912 ± 0.043	7.908 ± 0.043
DR-Learner	11.334 ± 0.225	7.964 ± 0.046	7.912 ± 0.047	**7.911 ± 0.043**	7.911 ± 0.043	7.909 ± 0.043
Double-ML	**10.651 ± 0.239**	**7.954 ± 0.047**	7.936 ± 0.045	7.919 ± 0.044	7.917 ± 0.046	**7.891 ± 0.050**
Causal Forest	11.154 ± 0.175	7.967 ± 0.045	7.949 ± 0.044	7.934 ± 0.043	7.931 ± 0.047	7.909 ± 0.044
Direct-Survival Methods						
Causal Survival Forest	11.674 ± 0.169	7.963 ± 0.057	7.942 ± 0.039	7.929 ± 0.037	7.911 ± 0.051	7.893 ± 0.042
Survival Meta-Learners						
T-Learner Survival	11.428 ± 0.160	8.007 ± 0.075	7.980 ± 0.233	7.911 ± 0.054	**7.902 ± 0.042**	7.902 ± 0.046
S-Learner Survival	11.713 ± 0.237	**7.921 ± 0.044**	7.912 ± 0.052	**7.900 ± 0.045**	**7.901 ± 0.046**	**7.897 ± 0.042**
Matching Survival	12.523 ± 0.289	7.949 ± 0.043	7.935 ± 0.053	7.920 ± 0.047	7.921 ± 0.046	7.912 ± 0.042

Validation and extension of synthetic findings. The semi-synthetic results confirm our synthetic benchmark's core insights while revealing additional structure-dependent nuances. Double-ML's dominance on ACTG data (10.65 RMSE) validates our synthetic benchmark finding that sophisticated causal methods excel in moderate-dimensional settings with controlled confounding. Similarly, S-Learner Survival's consistent top-tier performance across MIMIC variants (appearing as best or second-best on four of five datasets) also agrees with our synthetic benchmark finding that survival meta-learners provide robust performance under challenging censoring conditions.

Enhanced understanding of censoring sensitivity. The MIMIC censoring rate range (53% to 88%) provides granular validation of synthetic censoring effects while revealing method-specific stability patterns not observable in synthetic experiments. S-Learner Survival maintains stability across this range (RMSE range: 7.897-7.921), while T-Learner Survival exhibits instability at extreme censoring ($\pm$0.233 standard deviation at 82% censoring). This extends synthetic findings by showing that censoring tolerance varies not just between method families but within them, providing more precise guidance for high-censoring scenarios.

Convergence effects in realistic high-dimensional settings. The MIMIC results reveal a novel finding absent from synthetic experiments: performance convergence in high-dimensional, realistic covariate spaces. All methods cluster within a narrow RMSE range (7.89-8.01), contrasting with ACTG's broader spread (10.65-12.52). This convergence suggests that while synthetic experiments correctly identify relative method strengths, realistic covariate correlations and mixed data types may compress performance differences, making method selection dependent on secondary factors like stability, interpretability, and computational efficiency.

Implications for method selection. The results of our synthetic benchmarks provided foundational insights that can generalize to realistic settings, while our semi-synthetic evaluation reveals additional practical considerations. For method selection: (1) In moderate-dimensional settings with balanced censoring, sophisticated causal methods such as Double-ML offer clear advantages. (2) In high-dimensional, heavily censored settings typical of EHR studies, survival meta-learners provide the best combination of performance and stability. (3) The choice of method should explicitly consider dataset dimensionality and covariate complexity, not just censoring rates and sample sizes.

In Appendix G.2, we provide a more detailed analysis of the semi-synthetic results.

4.3 Benchmarking on Real Data

We also evaluate the three families of survival CATE estimators on two real-world datasets, one with known ground truth and one without.

Twin data. The Twins dataset (Almond et al., 2005; Curth et al., 2021) includes twin births from 1989-1991, with birth weight being heavier in the twin as treated and time to mortality as outcome. With known outcomes for both twins, this dataset provides ground truth for CATE evaluation. After replicating the same random treatment assignment strategy and the censoring time assignment following Curth et al. (2021), the treatment rate and censoring rate for the dataset are 68.1% and 84.8% respectively across 11,400 twin pairs. Since most of the mortality events occur within 30 days, we use $h = 30$ days during estimation. Figure 3 shows S- and DR-Learners (with imputation) and S-Learner-Survival exhibit lower CATE RMSE (7.2 days). T-Learner-Survival and Causal Forest with imputation exhibit the worse performance, consistent with their overall worse ranking from the benchmarking on our synthetic datasets (Figure 1). Surprisingly, Double ML with imputation exhibits the worst performance on the twin data, which is different from the overall ranking, suggesting potential unique patterns in this dataset. In Appendix H, we also show the result with $h = 180$ days; the conclusions are similar.

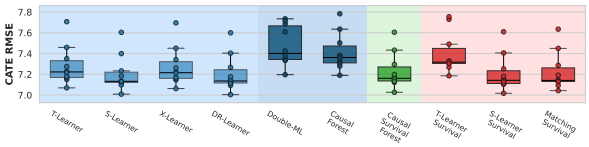


Figure 3: CATE RMSE for twin birth data using different estimator families with $h = 30$. Box plots show the distribution of error across 10 experimental runs.

HIV clinical trial. The ACTG 175 dataset (Hammer et al., 1996) compared four antiretroviral treatments in 2,139 HIV-infected patients. Following Meir et al. (2025), we convert time to months with h=30 months (13.7% baseline censoring) and introduce artificial censoring to test robustness (increasing to >90% censoring). More details on data and processing can be found in Appendix H.

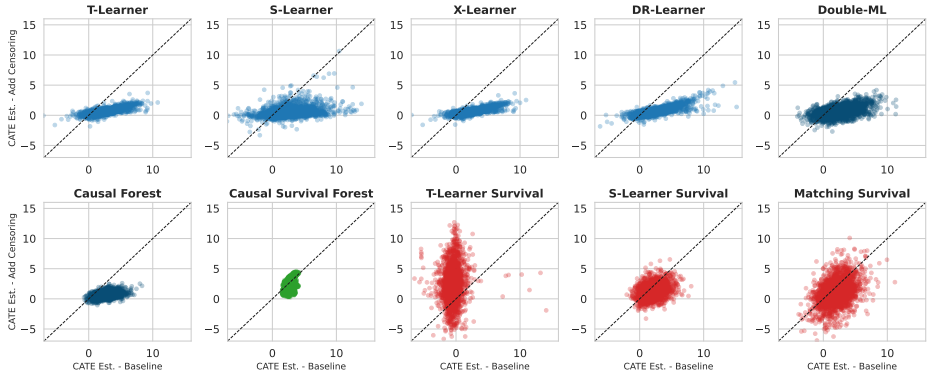


Figure 4: CATE estimation comparison between baseline and high-censoring conditions under ZDV vs. ZDV+ddI treatments. Each point represents an individual patient, with the dashed diagonal line indicating perfect consistency between baseline CATE estimation and that with the additional censoring injected.

Figure 4 compares CATE estimates between baseline and high-censoring conditions for the ZDV vs. ZDV+ddI comparison (results for other treatment comparisons are in Appendix H). Each point represents an individual patient, with the 45-degree dashed line indicating perfect consistency between conditions. We observe distinct behavioral patterns: Causal Survival Forest (green) produces estimates that cluster tightly around their original values; outcome imputation methods (blue) show higher variation in baseline estimates but concentrated predictions under high censoring; survival meta-learners (red) display substantial deviations from the 45-degree line, indicating sensitivity to censoring conditions. As ground truth is unknown, we cannot determine which approach is more accurate, but these patterns reveal fundamental differences in how estimators respond to increased censoring. For example, survival meta learner (the red scatter plots), especially the T- and matching-learners, exhibit instability under increased censoring settings (large variance in y-axis values).

5 Discussion

SurvHTE-Bench provides the first comprehensive and extensible platform for systematically benchmarking heterogeneous treatment effect estimators under right-censored survival settings. By spanning synthetic, semi-synthetic, and real datasets, the benchmark enables both controlled stress-testing of estimators under systematic assumption violations and validation in realistic clinical-like settings. Our empirical evaluations reveal strengths and weaknesses across estimator families.

While we have attempted to make our benchmark reasonably comprehensive, various limitations remain. First, while the synthetic datasets include numerous scenarios representing common real-world violations, they do not encompass all possible complexities, particularly varying degrees of severity in assumption violations. The binary nature of our violations (either present or absent) may not capture the nuanced continuum of partial violations. Second, extending the evaluation to include additional estimands such as survival probabilities at fixed horizons would further enrich the benchmark's scope. Our focus on restricted mean survival time represents just one of several clinically relevant estimands for treatment effect estimation. Additionally, we limit our analysis to static, binary treatments with fixed baseline covariates, excluding scenarios involving time-varying treatments, instrumental variables, and dynamic covariate structures.

Future work could expand SurvHTE-Bench in several directions. Incorporating a wider variety of direct causal estimation methods, such as g-computation approaches and SurvITE (Curth et al., 2021) specifically designed for survival outcomes, would provide a more comprehensive evaluation landscape, especially because Causal Survival Forest proved to be competitive but showed vulnerability to certain assumption violations like positivity. Exploring more complex data-generating mechanisms that better mimic the heterogeneity and longitudinal nature of real-world clinical data represents another promising direction. Finally, extending the benchmark to support multi-valued or continuous treatments would address important practical scenarios encountered in precision medicine and policy optimization.

ETHICS STATEMENT

SURVHTE-BENCH has significant positive potential for improving personalized medicine and clinical decision-making by enabling systematic evaluation of survival analysis methods under realistic assumption violations. By providing standardized benchmarks and practical guidance on when different estimators excel or fail, our work could accelerate the development of more reliable causal inference methods for high-stakes healthcare applications, ultimately supporting better patient outcomes through more informed treatment selection.

At the same time, our benchmark carries potential risks if misapplied. Practitioners may misinterpret benchmark results or place undue confidence in algorithmic decision-making, which could reduce necessary human oversight in clinical contexts. Moreover, although our study is methodological and does not involve human subjects directly, differences in estimator performance across demographic groups could exacerbate existing healthcare disparities if ignored. We therefore stress that our benchmark should not be used as a substitute for rigorous domain-specific validation, fairness assessment, or clinical trial evidence.

All datasets used in this work are either publicly available synthetic or semi-synthetic datasets, or real-world datasets with proper access provisions (e.g., credentialed approval for MIMIC-IV). No personally identifiable information was used, and all data handling complies with the terms of use of the original sources. We encourage future applications of SURVHTE-BENCH to incorporate fairness audits, domain-specific validation, and appropriate safeguards to ensure responsible deployment.

REPRODUCIBILITY STATEMENT

We provide complete resources to reproduce our results across synthetic, semi-synthetic, and real-data settings. (1) *Synthetic data:* The benchmark design and evaluation protocol are described in the main text (Sections 3 and 4.1), including the 8 causal configurations and 5 survival scenarios (40 datasets total). Extended generation formulas and per-dataset summaries are in Appendix A; imputation procedures in Appendix B; the full list of implemented estimators in Appendix C; causal method overviews in Appendix D; training details and hyperparameter grids in Appendix E; and additional synthetic results/analyses in Appendix F. (2) *Semi-synthetic data:* Setup, statistics, and full results appear in Appendix G with summary discussion in Section 4.2. (3) *Real data:* Processing details and additional analyses are provided in Appendix H; see also Section 4.3. We further study additional censoring mechanisms in Appendix I.

Code and instructions. The full codebase used for all experiments is available at the anonymized repository: `https://anonymous.4open.science/r/SurvHTE-Benchmark-206B`, with scripts and READMEs to reproduce all figures and tables from raw inputs.

Datasets. In the same anonymized repository, we include: (i) the complete synthetic suite (40 datasets from the 8×5 design); (ii) the semi-synthetic datasets, comprising the *ACTG (semi-synthetic)* dataset; and (iii) real-data materials for *Twins* and *ACTG 175*. For the semi-synthetic MIMIC resources, because MIMIC-IV requires credentialed access, we provide code to generate these datasets rather than redistributing raw MIMIC data. The MIMIC-IV dataset itself is hosted on PhysioNet at https://physionet.org/content/mimiciv/3.1/ and is publicly available to researchers upon credentialed approval. All other datasets listed above are included in the supplementary package in preprocessed or generated form, together with scripts to reproduce all splits and metrics.

In addition to enabling replication of our reported results, we intend SURVHTE-BENCH to serve as community infrastructure for the evaluation of survival HTE methods. The benchmark is designed to be modular and extensible, allowing researchers to incorporate new estimators or datasets while preserving comparability. This ensures not only reproducibility of our experiments but also a lasting resource for the community, providing a standardized basis for measuring progress in survival causal inference, a resource that has been missing until now, as well as in related areas of machine learning.

REFERENCES

Douglas Almond, Kenneth Y. Chay, and David S. Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, 2005.

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in medicine*, 24(24):3927–3944, 2005.

Susan Athey and Guido W Imbens. Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050(5):1–26, 2015.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178, 2019.

Na Bo, Yue Wei, Lang Zeng, Chaeryon Kang, and Ying Ding. A Meta-Learner Framework to Estimate Individualized Treatment Effects for Survival Outcomes. *Journal of Data Science*, 22 (4):505–523, 2024.

Paidamoyo Chapfuwa, Serge Assaad, Shuxi Zeng, Michael J. Pencina, Lawrence Carin, and Ricardo Henao. Enabling counterfactual survival analysis with balanced representations. In *Conference on Health, Inference, and Learning*, pp. 133–145, 2021.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Yifan Cui, Michael R Kosorok, Erik Sverdrup, Stefan Wager, and Ruoqing Zhu. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):179–211, 2023.

Alicia Curth, Changhee Lee, and Mihaela van der Schaar. SurvITE: learning heterogeneous treatment effects from time-to-event data. In *Advances in Neural Information Processing Systems*, 2021.

Scott M. Hammer, David A. Katzenstein, Michael D. Hughes, Holly Gundacker, Robert T. Schooley, Richard H. Haubrich, W. Keith Henry, Michael M. Lederman, John P. Phair, Manette Niu, Martin S. Hirsch, and Thomas C. Merigan. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 335(15):1081–1090, 1996.

Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. *Advances in neural information processing systems*, 35:32142–32159, 2022.

Nicholas C Henderson, Thomas A Louis, Gary L Rosner, and Ravi Varadhan. Individualized treatment effects with censored data via fully nonparametric bayesian accelerated failure time models. *Biostatistics*, 21(1):50–68, 2020.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841 – 860, 2008.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. DeepSurv: Personalized Treatment Recommender System Using A Cox Proportional Hazards Deep Neural Network. *BMC Medical Research Methodology*, 18:1–12, 2018.

Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

Tomer Meir, Uri Shalit, and Malka Gorfine. Heterogeneous Treatment Effect in Time-to-Event Outcomes: Harnessing Censored Data with Recursively Imputed Trees. *arXiv preprint arXiv:2502.01575*, 2025.

Shahriar Noroozizadeh, Pim Welle, Jeremy C Weiss, and George H Chen. The impact of medication non-adherence on adverse outcomes: Evidence from schizophrenia patients via survival analysis. In *Conference on Health, Inference, and Learning*, pp. 573–609. PMLR, 2025.

Shi-ang Qi, Neeraj Kumar, Mahtab Farrokh, Weijie Sun, Li-Hao Kuan, Rajesh Ranganath, Ricardo Henao, and Russell Greiner. An effective meaningful way to evaluate survival models. In *International Conference on Machine Learning*, 2023.

Jincheng Shen, Lu Wang, Stephanie Daignault, Daniel E Spratt, Todd M Morgan, and Jeremy MG Taylor. Estimating the optimal personalized treatment strategy based on selected variables to prolong survival via random survival forest with weighted bootstrap. *Journal of biopharmaceutical statistics*, 28(2):362–381, 2018.

Ori M Stitelman and Mark J van der Laan. Collaborative targeted maximum likelihood for time to event data. *The International Journal of Biostatistics*, 6(1), 2010.

Ori M Stitelman, C William Wester, Victor De Gruttola, and Mark J van der Laan. Targeted maximum likelihood estimation of effect modification parameters in survival analysis. *The International Journal of Biostatistics*, 7(1), 2011.

Mark J Van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011.

Shenbo Xu, Raluca Cobzaru, Stan N Finkelstein, Roy E Welsch, Kenney Ng, and Zach Shahn. Estimating heterogeneous treatment effects on survival outcomes using counterfactual censoring unbiased transformations. *arXiv preprint arXiv:2401.11263*, 2024.

Yizhe Xu, Nikolaos Ignatiadis, Erik Sverdrup, Scott Fleming, Stefan Wager, and Nigam Shah. Treatment heterogeneity with survival outcomes. In *Handbook of Matching and Weighting Adjustments for Causal Inference*, pp. 445–482. Chapman and Hall/CRC, 2023.

Weijia Zhang, Thuc Duy Le, Lin Liu, Zhi-Hua Zhou, and Jiuyong Li. Mining heterogeneous causal effects for personalized cancer treatment. *Bioinformatics*, 33(15):2372–2378, 2017.

Jie Zhu and Blanca Gallego. Targeted estimation of heterogeneous treatment effect in observational survival analysis. *Journal of Biomedical Informatics*, 107, 2020.

APPENDIX

This appendix provides supplementary material supporting the main text, including detailed descriptions of data generation processes, methodological explanations, experimental setups, and additional results. We begin by describing the mathematical formulations used to create our synthetic datasets, followed by detailed explanations of imputation methods and causal inference techniques. We then provide comprehensive information about model training procedures and hyperparameter settings for reproducibility. The appendix concludes with extensive additional experimental results on synthetic, semi-synthetic, and real-world datasets.

We would like to declare **the use of Large Language Models (LLMs)** in this work. LLMs were used as general-purpose assistive tools. Specifically, they supported parts of the writing process (editing, formatting, and polishing text) without contributing to the core methodology, scientific rigor, or originality of the research. In addition, LLMs were used to assist with improving visualization code for figures, documenting the code, and minor refactoring. No part of the conceptualization, design, or execution of the research relied on LLMs.

Appendix A: Additional Details of the Synthetic Datasets. This section provides the mathematical formulations for generating covariates, treatment assignments, event times, and censoring times across different scenarios. It describes how the synthetic datasets systematically vary across causal configurations and survival scenarios, including details on covariate generation, treatment assignment mechanisms, event time generation, censoring time generation, and observed data construction.

Appendix B: Imputation Methods Details. This section explains three surrogate imputation strategies for estimating true event time in right-censored survival data: Margin Imputation, IPCW-T Imputation, and Pseudo-Observation Imputation. It provides mathematical formulations for each method and discusses their respective advantages and limitations.

Appendix C: List of CATE Estimators in SURVHTE Benchmark. This section details the 52 different conditional average treatment effect (CATE) estimator variants evaluated in the benchmark, including outcome imputation methods, direct-survival CATE models, and survival meta-learners, with a breakdown of how these variants are constructed.

Appendix D: Detailed Overview of Causal Inference Methods. This section provides comprehensive explanations of various causal inference methods, including meta-learners (T-learner, S-learner, X-learner, DR-learner), Double ML, Causal Forest, Causal Survival Forest, and Survival Meta-Learners, discussing their implementation in survival contexts.

Appendix E: Model Training Details and Hyperparameters. This section covers the hyperparameter grids, model selection procedures, and computational costs associated with each method class evaluated in the benchmark, providing details on the experimental setup for reproducibility.

Appendix F: Additional Experimental Results for Synthetic Dataset. This section presents comprehensive experimental results on synthetic datasets, including full rankings of models, performance across different survival scenarios and causal configurations, detailed CATE RMSE and ATE bias plots, evaluation of auxiliary components, and convergence behavior under varying training set sizes.

Appendix G: Semi-Synthetic Datasets. This section includes data setup and detailed analysis of semi-synthetic datasets derived from ACTG 175 and MIMIC-IV, including covariate statistics, censoring rate range, and comprehensive performance results across methods.

Appendix H: Real-World Datasets. This section provides detailed descriptions of data preprocessing and additional experimental results for the Twins dataset and the ACTG 175 HIV clinical trial dataset, including CATE RMSE results with different time horizons and comparisons of CATE estimates between baseline and high-censoring conditions.

Appendix I: Informative Censoring via Unobserved Confounding. An additional dataset illustrating how censoring can depend on latent variables, violating the ignorable censoring assumption. Demonstrates the extensibility of SURVHTE-BENCH to incorporate alternative censoring mechanisms beyond the 8×5 design.

A Additional Details of the Synthetic Datasets

Our synthetic datasets systematically vary across two orthogonal dimensions: causal configurations (treatment assignment mechanisms and assumption violations) and survival scenarios (event-time distributions and censoring mechanisms). This section provides the mathematical formulations for generating covariates, treatment assignments, event times, and censoring times across all scenarios. For event time and censoring time distribution, we adapt the generation process from Meir et al. (2025) and make some adjustments, with, for example, different censoring mechanisms under informative censoring settings; for treatment assignment in observational study settings, we adapt the propensity score from Cui et al. (2023). For simplicity, we omit the unit index i in this section.

A.1 Covariate Generation

Following Cui et al. (2023); Meir et al. (2025), for all scenarios, we generate five baseline covariates independently from uniform distributions:

$$X_m \sim \text{Uniform}(0,1), \quad m = 1,2,3,4,5$$

Additionally, we generate two latent confounders $U_1, U_2 \sim \text{Uniform}(0,1)$ that are used when testing violations of the ignorability assumption.

A.2 Treatment Assignment Mechanisms

The treatment assignment mechanism W varies according to the causal configuration:

Randomized Controlled Trials (`RCT-50`, `RCT-5`): Treatment is assigned randomly with probability p:

$$W \sim \text{Bernoulli}(p)$$

where $p = 0.5$ for `RCT-50` and $p = 0.05$ for `RCT-5`.

Observational Studies (`OBS-`): Treatment assignment depends on covariates through a propensity score mechanism:

$$e(X) = \frac{1 + \text{Beta}(X_1; 2, 4)}{4} \quad \texttt{(OBS-CPS)}$$

$$e(X,U) = \frac{1 + \text{Beta}(0.3X_1 + 0.7U_1; 2, 4)}{4} \quad \texttt{(OBS-UConf)}$$

$$e(X) = \begin{cases} 1 & \text{if } X_1 > 0.8 \\ 0 & \text{if } X_1 < 0.2 \\ 0.5 & \text{otherwise} \end{cases} \quad \texttt{(OBS-NoPos)}$$

where $\text{Beta}(x; a, b)$ denotes the Beta probability density function with parameters a and b evaluated at x. For all observational configurations, $W \sim \text{Bernoulli}(e(\cdot))$.

A.3 Event Time Generation

Event times $T(w)$ under treatment $w \in \{0,1\}$ are generated according to five different survival scenarios:

Scenario `A` (Cox Model): Event times follow a Cox proportional hazards model with Weibull baseline hazard:

$$\begin{aligned} \lambda_T(t|W,X) &= h_0(t) \cdot \exp(\beta^T Z) \\ &= 0.5t^{-0.5} \cdot \exp[X_1 + (-0.5 + X_2) \cdot W + \epsilon] \end{aligned}$$

where $h_0(t) = 0.5t^{-0.5}$ is the Weibull baseline hazard with shape parameter $k = 0.5$ and scale parameter $\lambda_0 = 1.0$, and $\epsilon = 0.5(U_1 - X_2)$ if unobserved confounding is present, and $\epsilon = 0$ otherwise. Event times are generated via inverse transform sampling from the corresponding survival function.

Scenario `B` (AFT Model): Event times follow an Accelerated Failure Time (AFT) model:

$$\begin{aligned} \log T(w) = &-1.85 - 0.8 \cdot \mathbb{1}(X_1 < 0.5) + 0.7\sqrt{X_2} + 0.2X_3 \\ &+ [0.7 - 0.4 \cdot \mathbb{1}(X_1 < 0.5) - 0.4\sqrt{X_2}] \cdot w + \epsilon + \eta \end{aligned}$$

where $\eta \sim \mathcal{N}(0,1)$ and ϵ is defined as in Scenario A.

Scenario C (Poisson Model): Event times follow a Poisson distribution:

$$\lambda(w) = X_2^2 + X_3 + 6 + 2(\sqrt{X_1} - 0.3) \cdot w + \epsilon$$
$$T(w) \sim \text{Poisson}(\lambda(w))$$

Scenario D (AFT Model): Event times follow an AFT model with parameters adjusted for higher censoring:

$$\log T(w) = 0.3 - 0.5 \cdot \mathbb{1}(X_1 < 0.5) + 0.5\sqrt{X_2} + 0.2X_3 + [1 - 0.8 \cdot \mathbb{1}(X_1 < 0.5) - 0.8\sqrt{X_2}] \cdot w + \epsilon + \eta$$

Scenario E (Poisson Model): Event times follow a Poisson distribution with adjusted parameters:

$$\lambda(w) = X_2^2 + X_3 + 7 + 2(\sqrt{X_1} - 0.3) \cdot w + \epsilon$$
$$T(w) \sim \text{Poisson}(\lambda(w))$$

A.4 Censoring Time Generation

Censoring times C are generated differently across scenarios and depend on whether informative censoring is present:

Ignorable Censoring (Non-InfC Scenarios):

Scenario A: $C \sim \text{Uniform}(0, 3)$

Scenario B: $\lambda_C(t|W,X) = h_{0C}(t) \cdot \exp(\gamma^T Z)$
$= 2.0t^{1.0} \cdot \exp[\mu]$
where $\mu = -1.75 - 0.5\sqrt{X_2} + 0.2X_3 + [1.15 + 0.5 \cdot \mathbb{1}(X_1 < 0.5) - 0.3\sqrt{X_2}] \cdot W$

Scenario C: $C = \begin{cases} \infty & \text{with probability } 0.6 \\ 1 + \mathbb{1}(X_4 < 0.5) & \text{with probability } 0.4 \end{cases}$

Scenario D: $\lambda_C(t|W,X) = h_{0C}(t) \cdot \exp(\gamma^T Z)$
$= 2.0t^{1.0} \cdot \exp[\nu]$
where $\nu = -0.9 + 2\sqrt{X_2} + 2X_3 + [1.15 + 0.5 \cdot \mathbb{1}(X_1 < 0.5) - 0.3\sqrt{X_2}] \cdot W$

Scenario E: $C \sim \text{Poisson}(3 + \log(1 + \exp(2X_2 + X_3)))$

For scenarios B and D, $h_{0C}(t) = 2.0t^{1.0}$ is the Weibull baseline hazard for censoring with shape parameter $k = 2.0$ and scale parameter $\lambda_0 = 1.0$. Censoring times are generated via inverse transform sampling from the corresponding survival function.

Informative Censoring (-InfC Scenarios): When testing violations of ignorable censoring assumptions, we replace the above mechanisms with:

$$C_i \sim \text{Exponential}(\text{rate} = \lambda_0 + \alpha \cdot T_i) \tag{2}$$

where $\lambda_0 = 1.0$ and $\alpha = 0.1$ are baseline parameters that create dependence between censoring and event times.

While in the main benchmark we induce informative censoring by making censoring times dependent on event times, this is not the only way to violate the ignorable censoring assumption. To demonstrate the extensibility of our modular design and for completeness, we additionally include in Appendix I a setting where informative censoring arises through unobserved confounding.

A.5 Observed Data Construction

The observed survival data consists of:

$$\tilde{T} = \min(T, C) \quad \text{(observed time)}$$
$$\delta = \mathbb{1}(T \leq C) \quad \text{(event indicator)}$$

where $T = T(W)$ represents the factual event time under the observed treatment assignment.

The combination of these five survival scenarios with eight causal configurations yields our comprehensive benchmark of 40 synthetic datasets, each designed to test estimator performance under specific combinations of survival dynamics and causal assumption violations.

Table 4: Summary of event time and censoring time generation across survival scenarios

Scenario	Event Time Distribution	Censoring Mechanism	Censoring Rate
A	Cox (Weibull baseline, $k = 0.5$)	Uniform$(0, 3)$	Low ($< 30\%$)
B	AFT (Log-normal)	Cox (Weibull baseline, $k = 2.0$)	Low ($< 30\%$)
C	Poisson	Piecewise uniform	Medium (30-70%)
D	AFT (Log-normal)	Cox (Weibull baseline, $k = 2.0$)	High ($> 70\%$)
E	Poisson	Poisson	High ($> 70\%$)

Table 5: Censoring rate of synthetic datasets (50,000 samples). Notice that the censoring rates are different from Table 2 under informative censoring due to changes in the censoring distribution.

	Survival Scenarios				
Causal Configurations	A	B	C	D	E
RCT-50	0.203	0.073	0.392	0.913	0.794
RCT-5	0.200	0.036	0.390	0.881	0.770
OBS-CPS	0.201	0.066	0.393	0.914	0.789
OBS-UConf	0.201	0.073	0.392	0.918	0.795
OBS-NoPos	0.203	0.082	0.393	0.912	0.803
OBS-CPS-InfC	0.116	0.052	0.885	0.366	0.926
OBS-UConf-InfC	0.116	0.054	0.888	0.381	0.929
OBS-NoPos-InfC	0.116	0.058	0.891	0.403	0.932

Table 6: Treatment rate of synthetic datasets (50,000 samples).

	Survival Scenarios				
Causal Configurations	A	B	C	D	E
RCT-50	0.502	0.502	0.502	0.502	0.502
RCT-5	0.049	0.049	0.049	0.049	0.049
OBS-CPS	0.503	0.503	0.503	0.503	0.503
OBS-UConf	0.539	0.539	0.539	0.539	0.539
OBS-NoPos	0.500	0.500	0.500	0.500	0.500
OBS-CPS-InfC	0.503	0.503	0.503	0.503	0.503
OBS-UConf-InfC	0.539	0.539	0.539	0.539	0.539
OBS-NoPos-InfC	0.500	0.500	0.500	0.500	0.500

Table 7: Average treatment effect (ATE) of synthetic datasets (50,000 samples).

	Survival Scenarios				
Causal Configurations	A	B	C	D	E
RCT-50	0.163	0.125	0.750	0.724	0.754
RCT-5	0.163	0.125	0.750	0.724	0.754
OBS-CPS	0.163	0.125	0.750	0.724	0.754
OBS-UConf	0.004	0.132	0.740	0.831	0.740
OBS-NoPos	0.163	0.125	0.750	0.724	0.754
OBS-CPS-InfC	0.163	0.125	0.750	0.724	0.754
OBS-UConf-InfC	0.004	0.132	0.740	0.831	0.740
OBS-NoPos-InfC	0.163	0.125	0.750	0.724	0.754

B IMPUTATION METHODS DETAILS

We follow Qi et al. (2023) to implement three surrogate imputation strategies for estimating the true event time T in right-censored survival data. Let $Y = \min(T, C)$ be the observed time, with censoring indicator $\delta = \mathbb{1}\{T \leq C\}$. Let $S_{\text{KM}(\mathcal{D})}(t)$ denote the Kaplan-Meier estimate of the survival function using the dataset $\mathcal{D}$, and N the number of subjects. The three methods below are used to impute a surrogate outcome $\tilde{T}_i$ for censored subject i observed at time t_i.

1. Margin Imputation: This method assigns a "best guess" value to each censored subject using the nonparametric Kaplan–Meier estimator. This surrogate value, called the *margin time*, can be interpreted as the conditional expectation of the event time given that the event occurs after the censoring time. For a subject censored at time t_i, the margin-imputed event time is computed as:

$$\tilde{T}_i^{\text{margin}} = \mathbb{E}[T_i \mid T_i > t_i] = t_i + \frac{\int_{t_i}^{\infty} S_{\text{KM}(\mathcal{D})}(t)\, dt}{S_{\text{KM}(\mathcal{D})}(t_i)} \tag{3}$$

where $S_{\text{KM}(\mathcal{D})}(t)$ is the Kaplan–Meier survival estimate derived from the training dataset.

The reliability of this imputation depends on the censoring time. For example, if a subject is censored very early (e.g., at time 0), the margin time is highly uncertain due to the lack of observed data beyond that point. In contrast, if a subject is censored near the maximum observed follow-up, the margin time is more likely to be close to the true event time.

2. IPCW-T Imputation: This method imputes a surrogate event time for censored subjects based on the observed outcomes of subsequent uncensored individuals. Specifically, for a subject censored at time t_i, the imputed value is calculated as the average event time of all uncensored subjects with observed times after t_i:

$$\tilde{T}_i^{\text{IPCW}} = \frac{\sum_{j=1}^{N} \mathbb{1}\{t_i < t_j\} \cdot \mathbb{1}\{\delta_j = 1\} \cdot t_j}{\sum_{j=1}^{N} \mathbb{1}\{t_i < t_j\} \cdot \mathbb{1}\{\delta_j = 1\}} \tag{4}$$

This imputes the event time for subject i by averaging the observed event times of those uncensored subjects who experienced the event after t_i. The method is motivated by the idea that these subsequent subjects provide empirical evidence about the possible timing of the unobserved event.

However, a limitation of this approach is that it fails to provide an imputation when there are no uncensored subjects observed after t_i. In such cases, the denominator of the expression becomes zero, and the method is unable to approximate the event time. In Qi et al. (2023), subjects for whom this occurs are excluded from evaluation, whereas in our setup we used the original observed time as the imputed time.

3. Pseudo-Observation Imputation: This method imputes the event time using pseudo-observations, which estimate the contribution of each subject to an overall unbiased estimator of the event time distribution. Let $\hat{\theta}$ be an estimator of the mean event time based on right-censored data, and let $\hat{\theta}^{-i}$ denote the same estimator computed with the i-th subject removed from the dataset. Then, the pseudo-observation for subject i is defined as:

$$\tilde{T}_i^{\text{pseudo}} = e_{\text{Pseudo-Obs}}(t_i, \mathcal{D}) = N \cdot \hat{\theta} - (N-1) \cdot \hat{\theta}^{-i} \tag{5}$$

This quantity can be interpreted as the individual contribution of subject i to the overall estimate $\hat{\theta}$. In practice, both $\hat{\theta}$ and $\hat{\theta}^{-i}$ can be computed using the mean of the Kaplan–Meier survival curve:

$$\hat{\theta} = \mathbb{E}_t[S_{\text{KM}(\mathcal{D})}(t)], \quad \hat{\theta}^{-i} = \mathbb{E}_t[S_{\text{KM}(\mathcal{D}\setminus\{i\})}(t)]$$

Once the pseudo-observations $\tilde{T}_i^{\text{pseudo}}$ are computed for all censored subjects, they are substituted in place of the true event times for evaluation or modeling.

Although pseudo-observations are not exact conditional expectations, they can approximate $\mathbb{E}[T_i \mid X_i]$ under certain assumptions. In particular, when censoring is independent of covariates and the

sample size is large, pseudo-observations behave asymptotically like i.i.d. draws from the true conditional expectation:

$$\mathbb{E}[\tilde{T}_i^{\text{pseudo}} \mid X_i] \approx \mathbb{E}[T_i \mid X_i]$$

This makes the pseudo-observation method a principled, nonparametric approach for imputing censored survival times, particularly when estimating global quantities like the mean event time.

These imputation strategies enable us to transform the survival outcome into a fully observed target variable, allowing the application of standard regression-based methods in causal effect estimation. To ensure meaningful estimates, it is important that each imputed event time for a censored subject is guaranteed to be greater than or equal to the censoring time–reflecting the fact that the true event must occur after the last time it was observed. In our implementation, we manually enforce this constraint by setting the imputed value to the observed censoring time whenever the imputation procedure yields a value less than t_i.

C List of CATE Estimators in SurvHTE Benchmark

As mentioned in Section 3, in our benchmark, we evaluate three families of survival CATE methods, totaling 52 variants. We list the number of variants for each type of CATE estimator in Table 8.

For the outcome imputation methods, we first apply one of three imputation strategies (Pseudo-observation, Margin, or IPCW-T) (Qi et al., 2023) to handle the censored data, transforming the survival problem into a standard regression task. After imputation, we use these imputed outcomes with four different meta-learner frameworks (S-, T-, X-, and DR-Learners), each implemented with three different base regression models (Lasso Regression, Random Forest, and XGBoost), resulting in $3 \times 4 \times 3 = 36$ different variants. Additionally, we pair each imputation method with two specialized causal inference methods: Causal Forest (Athey et al., 2019) and Double ML (Chernozhukov et al., 2018), which adds $3 \times 1 + 3 \times 1 = 6$ more variants, for a total of 42 outcome imputation method variants.

For direct-survival CATE models, we include the Causal Survival Forest (CSF) (Cui et al., 2023) as a standalone method, which is specifically designed to handle right-censored data without requiring separate imputation steps.

For survival meta-learners, we implement three types of meta-learning frameworks that have been extended to handle censored data directly: S-learner, T-learner, and matching-learner (Noroozizadeh et al., 2025). Each of these frameworks is combined with three different base survival models (Random Survival Forest (Ishwaran et al., 2008), DeepSurv (Katzman et al., 2018), and DeepHit (Lee et al., 2018)) that estimate the underlying survival functions, resulting in $3 \times 3 = 9$ survival meta-learner variants.

In total, our benchmark evaluates $42 + 1 + 9 = 52$ different method configurations across the 40 synthetic datasets and the two real-world datasets described in Section 3.

Table 8: Breakdown of benchmarked survival-CATE estimator variants used in our experiments. Each row corresponds to a specific combination of method class, imputation strategy (if applicable), base learner(s), and CATE learner(s). Cells with numbers in parentheses indicate how many variants are contributed by the method(s) listed in that cell. The final column reports the total number of method variants constructed using that combination.

Method Class	**Imputation** *(No. options)*	**Base Learner** *(No. options)*	**CATE Learner** *(No. options)*	**No. Variants**
Outcome Imputation Method	Pseudo-obs, Margin, IPCW-T (3)	Lasso Regression, Random Forest, XGBoost (3)	Meta-Learners (S-, T-, X-, DR-) (4)	**36**
		—	Causal Forest (1)	**3**
		—	Double ML (1)	**3**
Direct-Survival CATE Models	—	—	Causal Survival Forest (1)	**1**
Survival Meta-Learners	—	Random Survival Forest, DeepSurv, DeepHit (3)	Survival Meta-Learners (S-, T-, Matching-) (3)	**9**
Total				**52**

D Detailed Overview of Causal Inference Methods

This section provides a comprehensive explanation of the causal inference methods evaluated in our benchmark. We begin with outcome imputation methods that transform censored survival data into standard regression problems, followed by direct-survival CATE models specifically designed for right-censored data, and finally survival meta-learners that adapt standard meta-learner frameworks to handle censoring. For each method, we present the theoretical foundation, algorithmic procedure, and specific implementation considerations in the survival analysis context. Our exposition focuses on highlighting the unique characteristics that make each approach suitable for different survival and causal inference scenarios, with particular attention to how these methods handle the challenges posed by censoring and treatment effect heterogeneity.

D.1 Outcome Imputation Method

Meta-learners represent a flexible framework for estimating conditional average treatment effects (CATEs) by decomposing the causal inference problem into standard supervised learning tasks. The key advantage of meta-learners is that they allow practitioners to leverage any out-of-the-box machine learning algorithm as a "base learner" while maintaining principled approaches to causal effect estimation. This modularity makes meta-learners particularly attractive in practice, as they can incorporate state-of-the-art ML methods (e.g., random forests, gradient boosting, neural networks) without requiring specialized causal inference implementations. For detailed explanations on meta-learners, one can refer to Künzel et al. (2019); Kennedy (2023). We provide a simplified overview below and largely refer to the documentation of the `econml` package.

T-learner (Künzel et al., 2019). The T-Learner (Two-Learner) adopts the most straightforward approach by fitting separate outcome models for treated and control groups. Given binary treatment $W \in \{0, 1\}$, features X, and outcome Y, the T-Learner:

1. Splits the data by treatment assignment: (X^0, Y^0) for controls and (X^1, Y^1) for treated units
2. Trains separate outcome models (i.e. predicting the outcome Y using features X):

$$\text{For control units: } \hat{\mu}_0 = M_0(Y^0 \sim X^0)$$
$$\text{For treated units: } \hat{\mu}_1 = M_1(Y^1 \sim X^1)$$

3. Estimates CATE as:

$$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$

where M_0 and M_1 can be any regression algorithm. The T-Learner is conceptually simple but can suffer from high variance when treatment groups have different sizes or when the outcome models extrapolate poorly to regions with limited overlap.

S-learner (Künzel et al., 2019). The S-Learner (Single-Learner) takes a unified modeling approach by including treatment assignment as an additional feature. The procedure involves:

1. Training a single model using all available data:

$$\hat{\mu} = M(Y \sim (X, W))$$

2. Estimating CATE as:

$$\hat{\tau}(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$$

This approach leverages all available data for training and can be more sample-efficient than the T-Learner. However, it relies heavily on the base learner's ability to capture treatment-feature interactions, and may perform poorly when these interactions are complex or when the treatment effect is small relative to the baseline outcome.

X-learner (Künzel et al., 2019). The X-Learner represents a more sophisticated approach that combines ideas from both T-Learner and inverse propensity weighting. The algorithm proceeds in multiple stages:

1. Fit initial outcome models:

$$\hat{\mu}_0 = M_1(Y^0 \sim X^0)$$
$$\hat{\mu}_1 = M_2(Y^1 \sim X^1)$$

2. Compute imputed treatment effects:

$$\text{For treated units: } \hat{D}^1 = Y^1 - \hat{\mu}_0(X^1)$$
$$\text{For control units: } \hat{D}^0 = \hat{\mu}_1(X^0) - Y^0$$

3. Model treatment effects:

$$\hat{\tau}_0 = M_3(\hat{D}^0 \sim X^0)$$
$$\hat{\tau}_1 = M_4(\hat{D}^1 \sim X^1)$$

4. Combine estimates using propensity scores:

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

where $g(x)$ is the estimation for propensity score $P(W = 1|X = x)$ and is typically fitted using logistic regression. The X-Learner is particularly effective when treatment groups have different sizes or when treatment effects are heterogeneous, as it explicitly models treatment effect variation and uses propensity weighting for optimal combination.

DR-learner (Kennedy, 2023). The DR-Learner (Doubly Robust Learner) extends the doubly robust framework to meta-learning by combining outcome modeling with propensity score estimation. The approach constructs doubly robust scores that remain consistent if either the outcome model or propensity model is correctly specified. It includes the following steps:

1. Fit outcome modeling for each treatment

$$\hat{\mu}_0 = M_1(Y^0 \sim X^0)$$
$$\hat{\mu}_1 = M_2(Y^1 \sim X^1)$$

2. Construct propensity score modeling

$$\hat{g} = M_g(W \sim X)$$

3. Construct doubly robust outcomes:

$$\hat{Y}_0^{DR} = \hat{\mu}_0(X) + \frac{(Y - \hat{\mu}_0(X))}{\hat{g}(X)} \cdot \mathbb{1}\{W = 0\}$$
$$\hat{Y}_1^{DR} = \hat{\mu}_1(X) + \frac{(Y - \hat{\mu}_1(X))}{\hat{g}(X)} \cdot \mathbb{1}\{W = 1\}$$

4. Final CATE estimation: $\hat{\tau}(x) = \hat{Y}_1^{DR} - \hat{Y}_0^{DR}$

The DR-Learner provides theoretical robustness guarantees and often performs well in practice, particularly when either outcome or treatment assignment can be accurately modeled.

Double ML (Chernozhukov et al., 2018). Double Machine Learning (Double ML or DML) represents a principled framework for estimating heterogeneous treatment effects when confounders are high-dimensional or when their relationships with treatment and outcome cannot be adequately captured by parametric models. The key insight of DML is to decompose the causal inference problem into two predictive tasks that can be solved using arbitrary machine learning algorithms while maintaining favorable statistical properties. Specifically, DML assumes the following structural relationships:

- $Y = \theta(X) \cdot W + g(X, Z) + \epsilon \quad \text{with} \quad \mathbb{E}[\epsilon|X, Z] = 0$
- $W = f(X, Z) + \eta \quad \text{with} \quad \mathbb{E}[\eta|X, Z] = 0$
- $\mathbb{E}[\eta \cdot \epsilon|X, Z] = 0$

where Y is the outcome, W is the treatment, X are the features of interest for heterogeneity, Z are confounding variables, and $\theta(X)$ is the conditional average treatment effect we aim to estimate. The method proceeds by first estimating two nuisance functions:

- Outcome regression: $q(X, Z) = \mathbb{E}[Y|X, Z]$
- Treatment regression: $f(X, Z) = \mathbb{E}[W|X, Z]$

These nuisance functions can be estimated using any machine learning algorithm capable of regression (for continuous treatments) or classification (for binary treatments). Popular choices include random forests, gradient boosting, neural networks, or regularized linear models. After obtaining estimates $\hat{q}$ and $\hat{f}$, DML constructs residualized outcomes and treatments:

$$\tilde{Y} = Y - \hat{q}(X, Z)$$
$$\tilde{W} = W - \hat{f}(X, Z)$$

The final step estimates $\theta(X)$ by regressing $\tilde{Y}$ on $\tilde{W}$ and X

$$\hat{\theta} = \arg\min_{\theta} \mathbb{E}_n[(\tilde{Y} - \theta(X) \cdot \tilde{W})^2]$$

Causal Forest (Athey et al., 2019). Causal Forest extends the random forest methodology to directly estimate heterogeneous treatment effects in a non-parametric, data-adaptive manner. Unlike meta-learners that rely on global models, Causal Forest estimates treatment effects locally by learning similarity metrics in the feature space and weighting observations accordingly. Causal Forest builds upon the same structural assumptions as DML but estimates $\theta(x)$ locally for each target point x. The method constructs a forest where each tree is grown using a **causal splitting criterion** that maximizes treatment effect heterogeneity rather than prediction accuracy. For a target point x, the treatment effect is estimated by solving:

$$\hat{\theta}(x) = \arg\min_{\theta} \sum_{i=1}^{n} K_x(X_i) \cdot (\tilde{Y}_i - \theta \cdot \tilde{W}_i)^2$$

where $K_x(X_i)$ represents the similarity between points x and X_i as determined by how frequently they fall in the same leaf across the forest, and $\tilde{Y}$, $\tilde{W}$ are residuals from nuisance function estimates.

Implementation in Survival Context In our benchmark, meta-learners, double ML, and causal forest are applied to survival outcomes through outcome imputation methods. We first apply imputation techniques (Pseudo-obs, Margin, or IPCW-T, see Appendix B for details) to convert censored survival times into continuous outcomes, then apply the meta-learners described above with various base regression algorithms (Lasso Regression, Random Forest, XGBoost). This two-stage approach allows leveraging the rich ecosystem of causal inference methods developed for continuous outcomes while handling the complexities of censored data.

D.2 Direct-Survival CATE Models

Causal Survival Forest (CSF) (Cui et al., 2023) extends the causal forest methodology directly to right-censored survival data by incorporating doubly robust estimating equations from survival analysis. Unlike meta-learners that require outcome imputation, CSF handles censored observations natively while maintaining the adaptive partitioning advantages of tree-based methods. CSF builds upon the causal forest framework of Athey et al. (2019) but adapts the splitting criterion and estimation procedure for survival outcomes. For a detailed explanation of the method, please refer to the original paper by Cui et al. (2023). We provide an overview of the estimation procedures as follows:

1. **Nuisance estimation**: Using cross-fitting, estimate nuisance components including:
   - Propensity scores: $\hat{e}(x) = P(W = 1|X = x)$
   - Outcome regression: $\hat{m}(x) = E[y(T)|X = x]$
   - Censoring survival function: $\hat{S}_w^C(s|x) = P(C \geq s|W = w, X = x)$
   - Conditional expectations: $\hat{Q}_w(s|x) = E[y(T)|X = x, W = w, T \wedge h > s]$

   where $y(T)$ is a transformation applied on the event time T, the same as defined in Eq.1.
2. **Forest construction**: Build a forest where each tree uses a causal splitting criterion that maximizes treatment effect heterogeneity. The splitting rule targets variation in the doubly robust scores rather than prediction accuracy.
3. **Local estimation**: For a target point x, compute forest weights $\alpha(x)$ indicating similarity based on leaf co-occurrence across trees, then estimate the CATE by solving:

   $$\sum \alpha(x)\psi_{\hat{\tau}(x)}(X, y(U), U \wedge h, W, \Delta^h; \hat{e}, \hat{m}, \hat{S}_w^C, \hat{Q}_w) = 0$$

   where ψ is the doubly robust score function that adjusts for both treatment assignment and censoring.

D.3 SURVIVAL META-LEARNERS

T-Learner-Survival (Bo et al., 2024; Noroozizadeh et al., 2025). The T-Learner can be adapted to right-censored survival data by fitting separate survival models for each treatment group. Let $W \in \{0, 1\}$ denote the treatment indicator, X be the covariate vector, and T the observed survival time with censoring indicator δ, and h the maximum follow-up time.

1. **Split data by treatment:** Partition the dataset into (X^0, T^0, δ^0) for $W = 0$ and (X^1, T^1, δ^1) for $W = 1$.
2. **Train separate survival models:** Fit a survival model (e.g., Random Survival Forest, DeepSurv, DeepHit) to each group:
$$\widehat{S}_0(u|x) = \text{Survival model fitted on } (X^0, T^0, \delta^0)$$
$$\widehat{S}_1(u|x) = \text{Survival model fitted on } (X^1, T^1, \delta^1)$$
3. **Estimate Restricted Mean Survival Time (RMST):** Compute RMST for each treatment as:
$$\widehat{\mu}_0(x) = \int_0^h \widehat{S}_0(u|x)du, \quad \widehat{\mu}_1(x) = \int_0^h \widehat{S}_1(u|x)du$$
4. **Estimate CATE:** For any x, estimate treatment effect:
$$\widehat{\tau}_{\text{T-learner}}(x) = \widehat{\mu}_1(x) - \widehat{\mu}_0(x)$$

S-Learner-Survival (Bo et al., 2024; Noroozizadeh et al., 2025). The S-Learner adapts by training a single survival model over all data with treatment as a covariate.

1. **Fit survival model:** Train a survival model over the full dataset using (X, W) as inputs:
$$\widehat{S}(u|x, w) = \text{Survival model fitted on } ((X, W), T, \delta)$$
2. **Estimate Restricted Mean Survival Time (RMST):** Compute RMST under both treatment conditions:
$$\widehat{\mu}(x, 0) = \int_0^h \widehat{S}(u|x, 0)du, \quad \widehat{\mu}(x, 1) = \int_0^h \widehat{S}(u|x, 1)du$$
3. **Estimate CATE:**
$$\widehat{\tau}_{\text{S-learner}}(x) = \widehat{\mu}(x, 1) - \widehat{\mu}(x, 0)$$

Matching-Survival (Noroozizadeh et al., 2025). The Matching-Learner estimates the CATE by imputing the counterfactual Restricted Mean Survival Time (RMST) using matched data points from the opposite treatment group.

1. **Estimate factual RMST:** Fit a survival model on the full dataset and compute:
$$\widehat{\mu}_{W_i}(X_i) = \int_0^h \widehat{S}(u|X_i, W_i)du$$
2. **Find matches:** For each individual i, identify K nearest neighbors $J_K(i)$ from the opposite treatment group $(1 - W_i)$.
3. **Estimate counterfactual RMST:** Average factual RMSTs of matched neighbors:
$$\widehat{\mu}_{1-W_i}(X_i) = \frac{1}{K} \sum_{j \in J_K(i)} \widehat{\mu}_{W_j}(X_j)$$
4. **Estimate CATE:** Compute CATE for each unit:
$$\widehat{\tau}_{\text{matching}}(X_i) = \big(\widehat{\mu}_{W_i}(X_i) - \widehat{\mu}_{1-W_i}(X_i)\big) \cdot \big(2W_i - 1\big)$$

This approach makes minimal modeling assumptions beyond nearest-neighbor similarity and is particularly helpful in settings with low overlap or where global models may be misspecified.

E Model Training Details and Hyperparameters on Benchmarking with Synthetic Data

To rigorously evaluate and compare the performance of causal inference models under controlled conditions, we conducted extensive benchmarking on synthetic datasets. Each synthetic dataset consisted of 50,000 samples generated under known data-generating processes explained in Appendix A. For each experimental repeat, we selected a subset of 5,000 samples for training, 2,500 for validation, and 2,500 for testing, using 10 distinct random seeds (experimental repeats) to ensure robustness. Hyperparameters for each model were tuned on the validation set to minimize the Conditional Average Treatment Effect Root Mean Squared Error (CATE-RMSE). Throughout this paper, final results are always reported on the held-out test set using the best-performing configuration. Appendix F.7 provides complementary experiments that analyze the convergence behavior of each method under varying training set sizes.

This appendix details the hyperparameter grids used for model selection, the specific survival and outcome models applied within each causal inference framework, and the average computational cost associated with each method class.

E.1 Hyperparameters for Outcome Imputation Methods

For methods based on outcome imputation, we employed standard regressors to estimate the conditional mean of the survival outcome given covariates and treatment assignment. We considered Lasso regression, Random Forest, and XGBoost as base models. Each was optimized using cross-validated grid search on the training set. The corresponding hyperparameter grids are listed in Table 9.

Table 9: Hyperparameter Grids for Outcome Imputation Regressors

Regressor	**Hyperparameter Grid**
Lasso	Alpha: {0.001, 0.01, 0.1, 1, 10}
Random Forest	Number of trees: {50, 100} Maximum depth: {3, 5, None}
XGBoost	Number of trees: {50, 100} Learning rate: {0.01, 0.1} Maximum depth: {3, 5}

E.2 Hyperparameters for Direct Survival CATE Models

For direct modeling of survival outcomes, we employed the Causal Survival Forest (CSF), which adapts the Causal Forest framework to handle right-censored data. We used the default hyperparameters from the original implementation. These are summarized in Table 10.

Table 10: Default Hyperparameters for Causal Survival Forest

Parameter	**Default Value**
Number of trees grown	2000
Fraction of data per tree	0.5
Variables tried per split	$\min(\lceil\sqrt{p}+20\rceil, p)$
Minimum samples in a leaf	5
Maximum imbalance of splits	0.05
Penalty for imbalance at split	0
Account for treatment and censoring in split stability	TRUE
Trees per subsample for confidence intervals	2

E.3 Hyperparameters for Survival Meta-Learners

For survival meta-learners–specifically T-Learner-Survival, S-Learner-Survival, and Matching-Learner-Survival–we used three different base survival models: Random Survival Forest (RSF), DeepSurv, and DeepHit. Each of these models was tuned using a predefined hyperparameter grid, listed in Table 11.

Table 11: Set of Hyperparameters for Survival Analysis Models

Model	Hyperparameter	Values
RSF	Number of estimators	{100, 250, 500}
	Minimum samples per split	{5, 10, 20}
	Minimum samples per leaf	{2, 5, 10}
DeepHit	Number of nodes per layer	{32, 64, 128, 256}
	Batch normalization	{True, False}
	Dropout rate	{0.0, 0.1, 0.2, 0.3}
	Learning rate	{0.001, 0.01, 0.05}
	Batch size	{128, 256, 512}
	Epochs	{200, 512, 1000}
	Alpha	{0.1, 0.2, 0.3, 0.5}
	Sigma	{0.05, 0.1, 0.2, 0.3}
DeepSurv	Number of nodes per layer	{32, 64, 128, 256}
	Batch normalization	{True, False}
	Dropout rate	{0.0, 0.1, 0.2, 0.3}
	Learning rate	{0.001, 0.01, 0.05}
	Batch size	{128, 256, 512}
	Epochs	{200, 512, 1000}

Hyperparameters were selected through empirical tuning informed by prior literature. For neural network-based models (DeepSurv, DeepHit), we used early stopping to mitigate overfitting. All experiments were made reproducible by setting random seeds. The best-performing hyperparameter configuration was selected using CATE-RMSE on the validation set, and all final results were obtained on the test set using these optimal configurations.

E.4 Computation Time of Causal Methods

We also measured the computational cost of each CATE estimation method in terms of average runtime per dataset and experimental repeat. Each runtime was recorded using Python's `time.time()` and averaged across 40 synthetic datasets and 10 random seeds. Table 12 presents the mean runtime (in seconds) and standard deviation (excluding the time required for imputation). As expected, neural network-based survival models incur substantially higher computational costs than classical or tree-based methods. All experiments were conducted on a machine equipped with an AMD Ryzen 9 5900X CPU, 128GB RAM, and an NVIDIA GeForce RTX 4090 GPU (CUDA version 12.2).

Table 12: Average computation time per dataset per experimental repeat for each causal method. Runtime is reported in seconds with standard deviation across runs.

Method Class	Method	Runtime (s)
Outcome Imputation Method: Meta-learners	T-learner	2.14 ± 1.38
	S-learner	1.84 ± 1.22
	X-learner	2.92 ± 2.42
	DR-learner	3.34 ± 1.88
Outcome Imputation Method: Forest / ML-based learners	Double ML	5.27 ± 0.40
	Causal Forest	5.75 ± 0.40
Direct-Survival CATE Models	Causal Survival Forest	0.78 ± 0.06
Survival Meta-Learners	T-learner Survival	31.31 ± 16.88
	S-learner Survival	22.99 ± 14.23
	Matching-Survival	49.40 ± 23.25

F Additional Experimental Results for Synthetic Dataset

This section provides comprehensive experimental results on our synthetic datasets, expanding on the key findings presented in the main text. We begin in Appendix F.1 with a full Borda ranking of all 52 model combinations, summarizing global performance across every causal configuration and survival scenario. In Appendix F.2, we explore how performance varies across different survival scenarios–illustrating the impact of censoring patterns and time-to-event distributions on method rankings. Appendix F.3 then delves into how violations of causal assumptions (treatment randomization, ignorability, positivity, and censoring mechanisms) reshape the ranking of models for effectiveness of each causal method.

Subsequent sections (F.4 and F.5) present detailed performance metrics—CATE RMSE and ATE bias, respectively—across all 8 causal configurations and 5 survival scenarios, with box plots capturing variability over 10 experiment repetitions. We also evaluate auxiliary components in F.6, including imputation methods and base learners (regression, survival, and propensity models), and in Appendix F.7 we examine convergence behavior under varying training set sizes. Together, these detailed results support the robustness, sensitivity, and practical trade-offs of each model family in a wide spectrum of data-generating and causal settings.

F.1 Full Ranking of Models

To compare the overall performance of the methods across all synthetic datasets, we computed a Borda ranking based on the average rank of each method's test set CATE RMSE (Table 13). The ranking procedure aggregates method performance across all combinations of causal configurations and survival scenarios. For each method, we first computed its RMSE on the test subset of the CATE predictions for each (causal configuration, survival scenario) pair. We then ranked all 52 methods (described in Appendix C) within each pair and calculated the average rank across these conditions. This average rank represents the method's Borda score and serves as a unified summary of its performance robustness in our synthetic data experiments.

Table 13: Borda Ranking of All Methods

Rank	Method	Score	Rank	Method	Score
1	(Matching-Survival, DeepSurv)	6.70	**27**	(X-Learner, Pseudo-Obs, XGB)	24.12
2	(S-Learner-Survival, DeepSurv)	6.85	**28**	(DR-Learner, IPCW-T, Lasso)	24.20
3	(Double-ML, Margin)	7.48	**29**	(Causal Forest, Pseudo-Obs)	24.98
4	(X-Learner, Pseudo-Obs, Lasso)	10.70	**30**	(X-Learner, Pseudo-Obs, RandomForest)	25.10
5	(Double-ML, IPCW-T)	12.52	**31**	(T-Learner, IPCW-T, Lasso)	25.58
6	(Causal Survival Forest)	12.88	**32**	(S-Learner, IPCW-T, RandomForest)	26.00
7	(Causal Forest, Margin)	13.08	**33**	(T-Learner, Margin, RandomForest)	26.75
8	(X-Learner, IPCW-T, RandomForest)	13.15	**34**	(S-Learner, Margin, Lasso)	29.00
9	(X-Learner, IPCW-T, Lasso)	13.78	**35**	(S-Learner, IPCW-T, Lasso)	29.00
10	(Double-ML, Pseudo-Obs)	14.65	**36**	(S-Learner, Pseudo-Obs, Lasso)	29.23
11	(S-Learner-Survival, RSF)	15.08	**37**	(T-Learner-Survival, DeepHit)	29.40
12	(X-Learner, IPCW-T, XGB)	17.03	**38**	(T-Learner-Survival, RSF)	30.25
13	(X-Learner, Margin, Lasso)	18.11	**39**	(T-Learner, IPCW-T, RandomForest)	30.55
14	(Causal Forest, IPCW-T)	18.33	**40**	(S-Learner, Pseudo-Obs, RandomForest)	31.85
15	(T-Learner-Survival, DeepSurv)	19.48	**41**	(S-Learner, Pseudo-Obs, XGB)	32.73
16	(Matching-Survival, DeepHit)	19.53	**42**	(DR-Learner, Margin, RandomForest)	35.28
17	(S-Learner, Margin, XGB)	19.68	**43**	(X-Learner, Margin, XGB)	36.33
18	(DR-Learner, Margin, Lasso)	20.28	**44**	(DR-Learner, IPCW-T, RandomForest)	36.65
19	(S-Learner-Survival, DeepHit)	20.68	**45**	(T-Learner, Margin, XGB)	38.85
20	(X-Learner, Margin, RandomForest)	21.00	**46**	(T-Learner, IPCW-T, XGB)	39.80
21	(T-Learner, Margin, Lasso)	21.68	**47**	(T-Learner, Pseudo-Obs, RandomForest)	42.10
22	(Matching-Survival, RSF)	21.85	**48**	(DR-Learner, Margin, XGB)	43.65
23	(DR-Learner, Pseudo-Obs, Lasso)	22.08	**49**	(DR-Learner, IPCW-T, XGB)	44.10
24	(S-Learner, Margin, RandomForest)	22.80	**50**	(DR-Learner, Pseudo-Obs, RandomForest)	46.15
25	(S-Learner, IPCW-T, XGB)	23.03	**51**	(T-Learner, Pseudo-Obs, XGB)	47.38
26	(T-Learner, Pseudo-Obs, Lasso)	23.85	**52**	(DR-Learner, Pseudo-Obs, XGB)	49.55

F.2 Ranking of Causal Methods for Different Survival Scenarios

In Figure 5, we present the Borda ranking of all causal model families across five different survival scenarios (A–E). For each scenario, the average rank of each method is computed over 8 distinct causal configurations, allowing us to assess robustness across varying underlying data-generating processes. The horizontal layout of each plot ranks methods from best (left, top to bottom) to worst (right, bottom to top), with rank values annotated next to each method for clarity.

These plots illustrate how model performance shifts as censoring rates and survival distributions vary. In Scenario A, which involves minimal censoring, outcome regression-based methods such as Double-ML and X-Learner dominate the rankings. However, as we move toward Scenarios D and E–both characterized by higher censoring–direct survival modeling approaches such as S-Learner-Survival, Matching-Survival, and Causal Survival Forest consistently rise to the top. This pattern suggests that survival-specific modeling is better suited to handle the uncertainty introduced by heavy censoring, outperforming outcome imputation strategies in such settings.

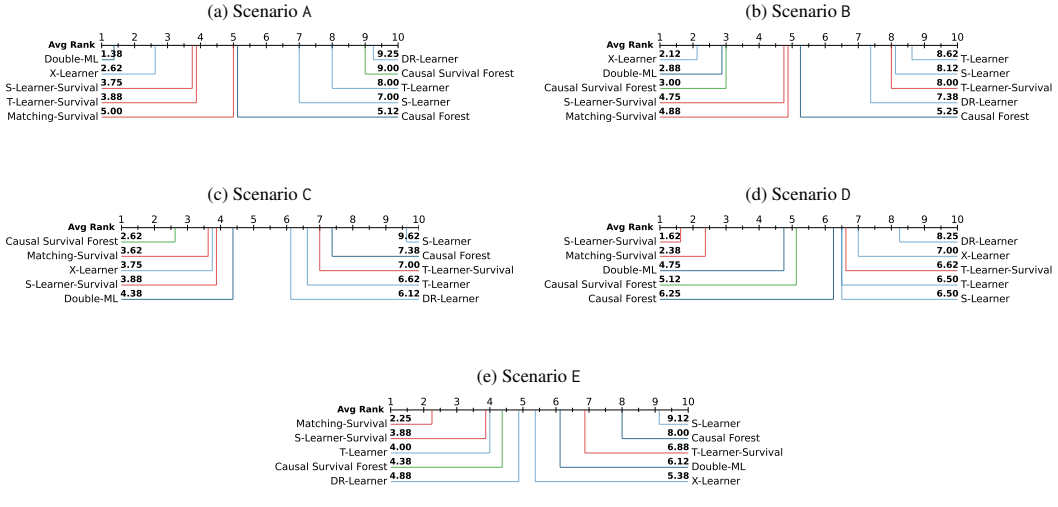


Figure 5: Average Ranking of Each Model for each Survival Scenario.

F.3 Ranking of Causal Methods for Different Causal Configurations

In Figure 6, we present the Borda ranking of causal model families across eight distinct causal configurations, each representing different combinations of assumptions related to treatment assignment (RCT vs. observational), ignorability, positivity, and censoring mechanisms. Within each configuration, the average rank of each method is computed over all survival scenarios, allowing us to isolate how assumption violations affect model performance independently of survival data characteristics.

Notably, outcome imputation approaches such as X-Learner and Double-ML perform best in randomized settings with balanced treatment (e.g., RCT-50%, panel a), but their performance deteriorates as we move to settings with imbalance, unmeasured confounding, or positivity violations. In contrast, survival-specific methods such as S-Learner-Survival, Matching-Survival, and Causal Survival Forest consistently rise in the rankings under these challenging conditions–particularly when multiple violations occur simultaneously (e.g., panel g). This trend suggests that survival meta-learners and direct modeling of the survival process offer increased robustness to violations of standard causal assumptions, especially in the presence of unmeasured confounding and informative censoring. Another finding here is that Causal Survival Forest maintains strong performance across many configurations, consistently ranking in the top half–particularly in settings involving unmeasured confounding or informative censoring. However, when the positivity assumption is violated (e.g., Figure 6e and h), its performance declines, suggesting limitations in modeling highly sparse regions of the covariate space with deterministic treatment assignment.

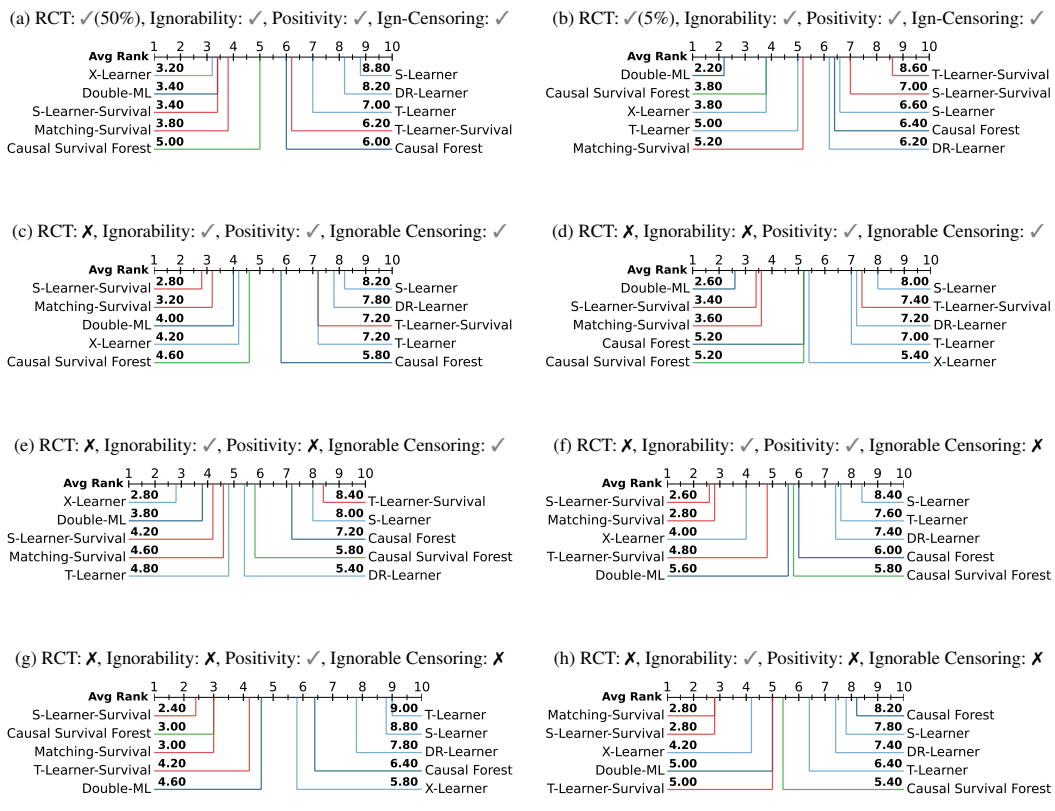
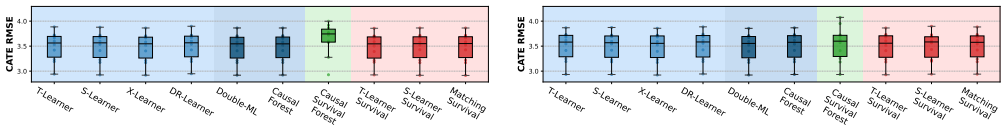


Figure 6: Average ranking of each model for each causal configuration.

F.4 FIGURE RESULTS - CATE RMSE

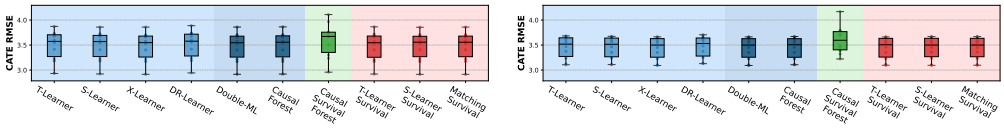
This section presents the complete CATE RMSE results for each family of causal inference methods across various survival analysis scenarios. For each scenario, we display performance under 8 distinct causal configurations, each varying in terms of treatment assignment (RCT vs. observational), ignorability, positivity, and censoring assumptions. These results highlight the robustness and sensitivity of different methods under varying degrees of assumption violations.

For each survival scenario and causal configuration, we selected the best hyperparameter setting and base model configuration for each causal method family based on validation set performance. The RMSE values shown in the figures reflect the performance of these selected models on the test set. The box plots are from the 10 independent experimental repeats to account for random seed variability.



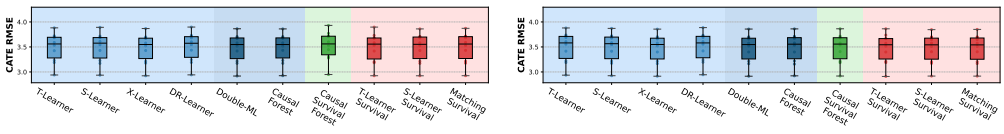
(a) RCT: ✓(50%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓

(b) RCT: ✓(5%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓



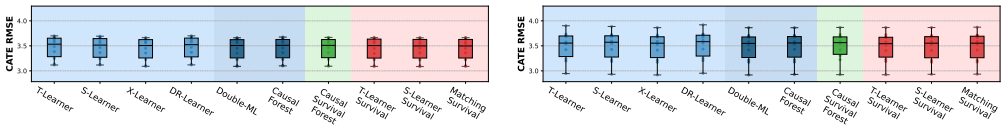
(c) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✓

(d) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✓



(e) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✓

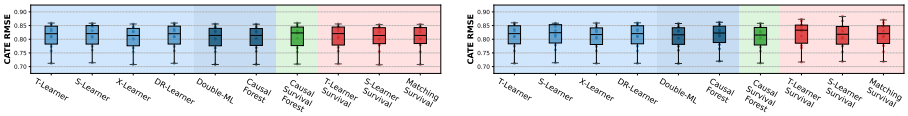
(f) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✗



(g) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✗

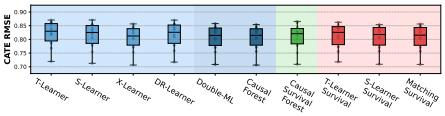
(h) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✗

Figure 7: CATE RMSE across different experiments in Scenario A.

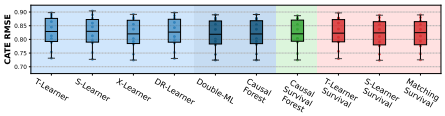


(a) RCT: ✓(50%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓

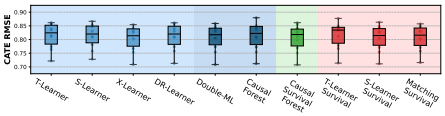
(b) RCT: ✓(5%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓



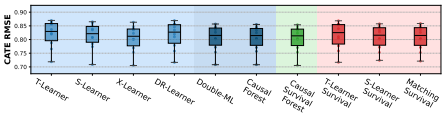
(c) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✓



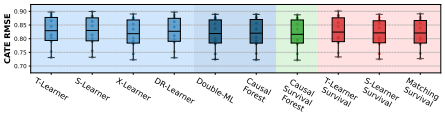
(d) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✓



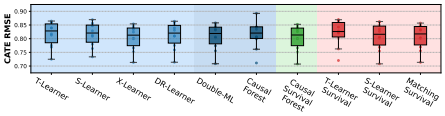
(e) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✓



(f) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✗

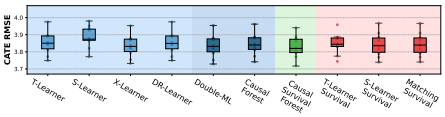


(g) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✗

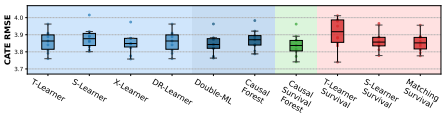


(h) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✗

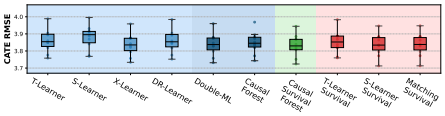
Figure 8: CATE RMSE across different experiments in Scenario B.



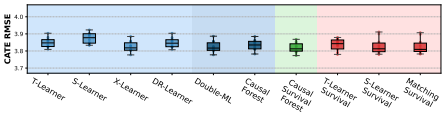
(a) RCT: ✓(50%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓



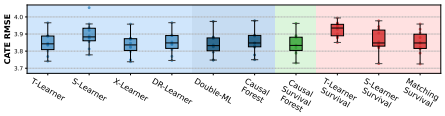
(b) RCT: ✓(5%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓



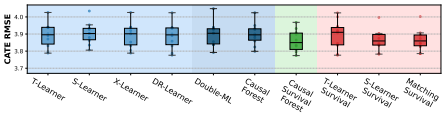
(c) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✓



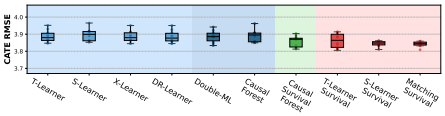
(d) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✓



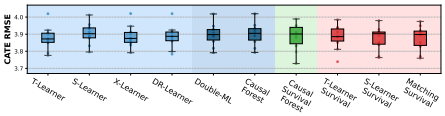
(e) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✓



(f) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✗

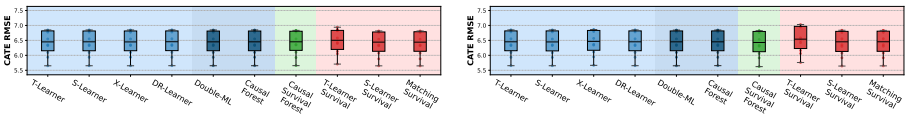


(g) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✗



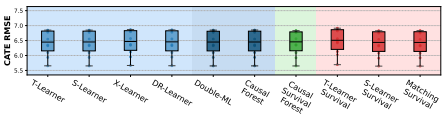
(h) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✗

Figure 9: CATE RMSE across different experiments in Scenario C.

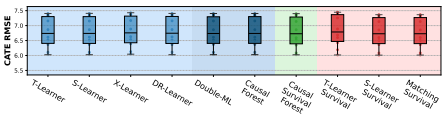


(a) RCT: ✓(50%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓

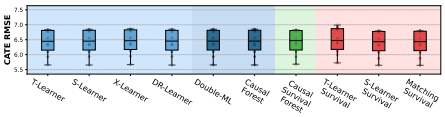
(b) RCT: ✓(5%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓



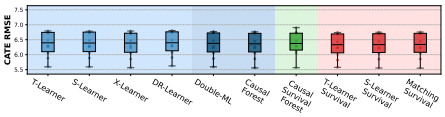
(c) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✓



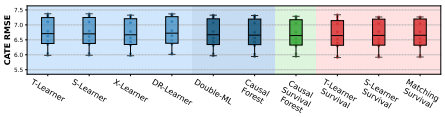
(d) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✓



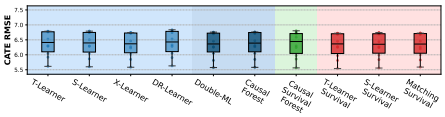
(e) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✓



(f) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✗

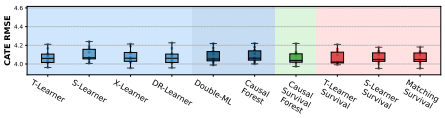


(g) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✗

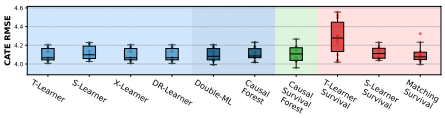


(h) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✗

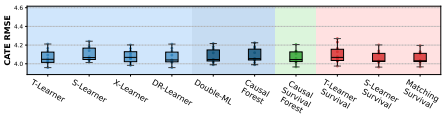
Figure 10: CATE RMSE across different experiments in Scenario D.



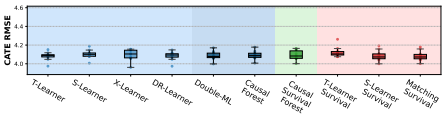
(a) RCT: ✓(50%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓



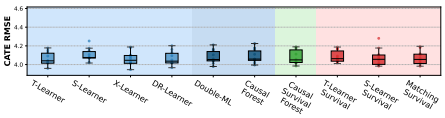
(b) RCT: ✓(5%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓



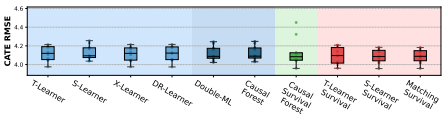
(c) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✓



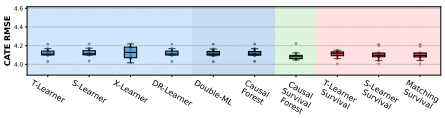
(d) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✓



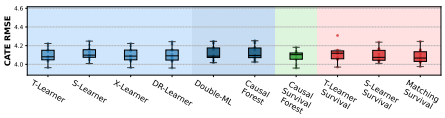
(e) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✓



(f) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✗



(g) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✗



(h) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✗

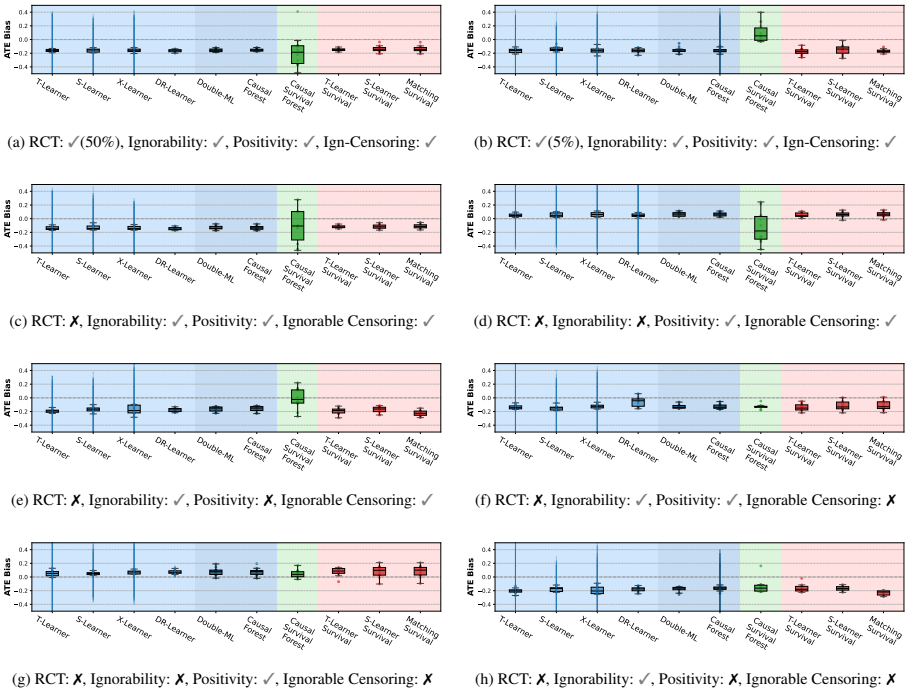
Figure 11: CATE RMSE across different experiments in Scenario E.

F.5 FIGURE RESULTS - ATE BIAS

This section presents the ATE bias results for each family of causal inference methods across various survival scenarios. As with the CATE RMSE results in Appendix F.4, we display performance under 8 distinct causal configurations per scenario, each varying in treatment assignment (RCT vs. observational), ignorability, positivity, and censoring assumptions.

For each survival scenario and causal configuration, the model shown corresponds to the best hyper-parameter setting and base model configuration selected based on CATE RMSE performance on the validation set – ATE bias was not used for model selection for consistent results with other sections. The reported ATE bias values are computed on the test set and defined as the difference between the *predicted ATE* from the test population and the *true ATE* in the full population.

Each box plot represents results from 10 independent experimental repeats to account for random seed variability. For meta-learners and double machine learning models, which by design can provide 95% confidence intervals for ATE estimates, we also include these intervals in the plots– adjusted accordingly to center around the ATE bias. These confidence intervals are obtained via 100 bootstrap samples and are notably wider than the variability observed across the 10 experimental repeats. The zero bias line is shown as a dashed reference line.



(a) RCT: ✓(50%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓

(b) RCT: ✓(5%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓

(c) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✓

(d) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✓

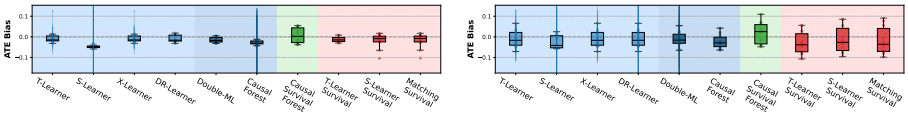
(e) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✓

(f) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✗

(g) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✗

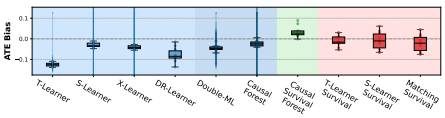
(h) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✗

Figure 12: ATE Bias across different experiments in Scenario A.

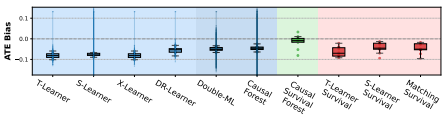


(a) RCT: ✓(50%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓

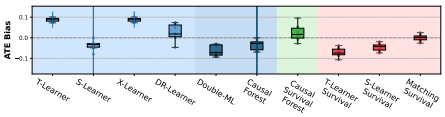
(b) RCT: ✓(5%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓



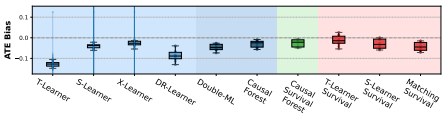
(c) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✓



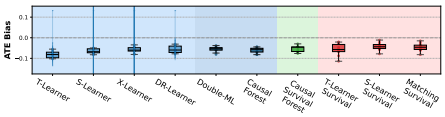
(d) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✓



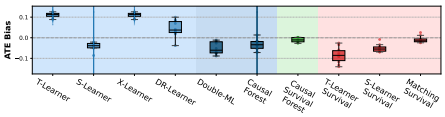
(e) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✓



(f) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✗

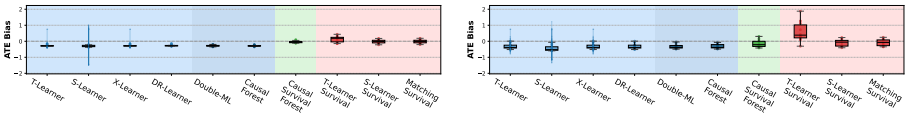


(g) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✗



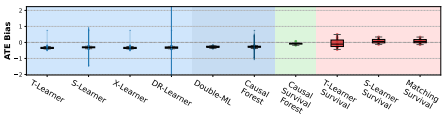
(h) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✗

Figure 13: ATE Bias across different experiments in Scenario B.

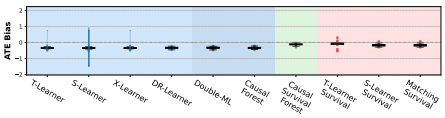


(a) RCT: ✓(50%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓

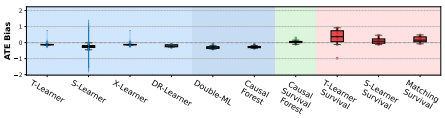
(b) RCT: ✓(5%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓



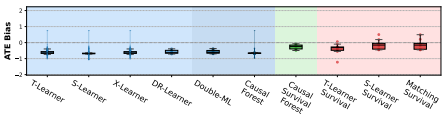
(c) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✓



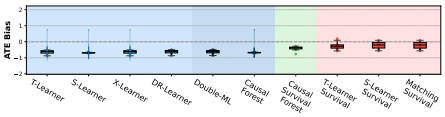
(d) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✓



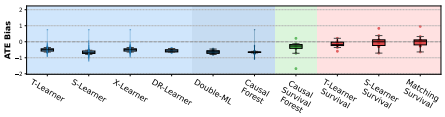
(e) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✓



(f) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✗

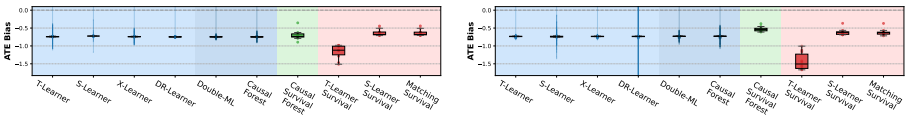


(g) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✗



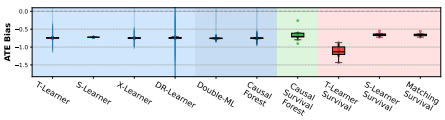
(h) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✗

Figure 14: ATE Bias across different experiments in Scenario C.

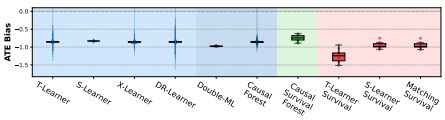


(a) RCT: ✓(50%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓

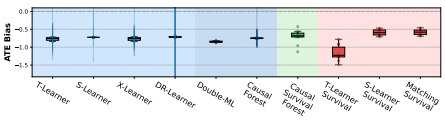
(b) RCT: ✓(5%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓



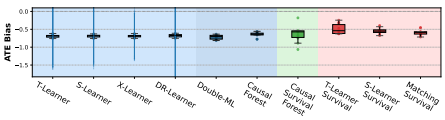
(c) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✓



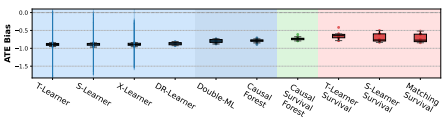
(d) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✓



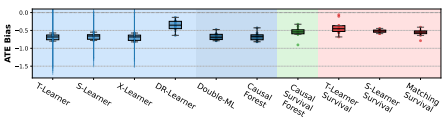
(e) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✓



(f) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✗

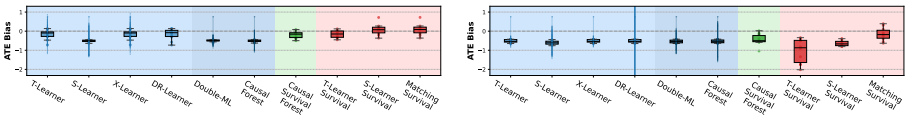


(g) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✗



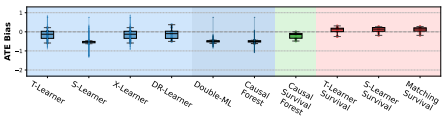
(h) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✗

Figure 15: ATE Bias across different experiments in Scenario D.

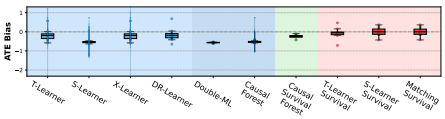


(a) RCT: ✓(50%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓

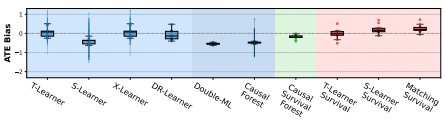
(b) RCT: ✓(5%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓



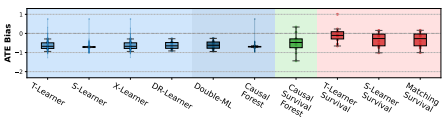
(c) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✓



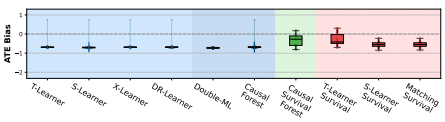
(d) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✓



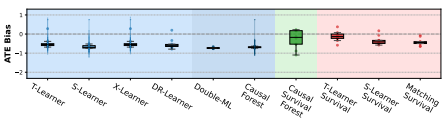
(e) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✓



(f) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✗



(g) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✗



(h) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✗

Figure 16: ATE Bias across different experiments in Scenario E.

F.6 Evaluation on Auxiliary Imputation and Base Learners

In this section, we report the performance of auxiliary imputation and base regression or survival learners on the test sets.

F.6.1 Imputation Evaluation

Table 14 reports the MAE of the three imputation methods (Pseudo-obs, Margin, IPCW-T) across eight causal configurations and five censoring scenarios on the test sets. Recall that Scenarios `A` and `B` have low censoring ($<30\%$), Scenario `C` medium (30–70%), and Scenarios `D` and `E` high ($>70\%$), except it is switched in `-InfC` causal configurations, as mentioned in Appendix A. We can tell that the imputation method Pseudo-obs is only competitive under minimal censoring and suffers from high variability. Margin imputation provides the best balance of accuracy and robustness, especially as censoring intensifies. IPCW-T imputation improves over Pseudo-obs in most cases, but generally underperforms relative to Margin in medium- and high-censor contexts.

F.6.2 Base Regression Learner Evaluation

See Table 15, 16, 17, 18 for MAE of prediction by the base regression learners for S-, T-, X-, DR-learners. The MAE is calculated by comparing a base learner's predicted event times and imputed event times by the imputation method (the latter is used as the "ground truth" for the base regression learners). Since there are three imputation methods used, we first take the average of MAE across three different imputation methods within each random split, then report the mean and standard deviation of the average MAE across 10 experimental repeats with different random splits.

See Table 19 for the AUC on the evaluation of the predicted propensity score of DR-learners.

F.6.3 Base Survival Learner Evaluation

See Table 20, 21, 22 for time-dependent concordance index on different base survival learners by the base survival learners for S-, T-, matching-learners. We report the mean and standard deviation across 10 experimental repeats with different random splits.

Table 14: Evaluation on imputation methods across different survival scenarios and causal configurations. MAE between the imputed and true event times on testing set is reported as mean ± std. over 10 experimental repeats. "Total Win" row counts the number of survival configurations × random split combinations (8 × 10 = 80) in which each method achieved the lowest MAE, and is calculated within each scenario. The same rule applies to all the tables below in Appendix F.6.

Survival Scenario	Causal Configuration	Imputation Method		
		Pseudo-obs	Margin	IPCW-T
A	RCT-50	**0.437±0.021**	0.446±0.025	0.470±0.027
	RCT-5	**0.378±0.027**	0.387±0.029	0.405±0.032
	OBS-CPS	**0.448±0.014**	0.459±0.014	0.481±0.015
	OBS-UConf	**0.423±0.026**	0.520±0.028	0.455±0.03
	OBS-NoPos	**0.411±0.023**	0.420±0.023	0.442±0.025
	OBS-CPS-InfC	0.390±0.020	**0.374±0.014**	0.388±0.014
	OBS-UConf-InfC	0.369±0.029	0.482±0.027	**0.362±0.028**
	OBS-NoPos-InfC	0.347±0.023	**0.336±0.024**	0.349±0.026
	Total Win	51	21	8
B	RCT-50	0.061±0.005	0.05±0.003	**0.048±0.004**
	RCT-5	0.027±0.003	0.022±0.002	**0.021±0.003**
	OBS-CPS	0.052±0.005	0.042±0.004	**0.040±0.003**
	OBS-UConf	0.058±0.004	0.152±0.007	**0.046±0.004**
	OBS-NoPos	0.068±0.008	0.057±0.005	**0.056±0.005**
	OBS-CPS-InfC	0.039±0.005	0.037±0.005	**0.036±0.005**
	OBS-UConf-InfC	0.040±0.004	0.140±0.005	**0.038±0.004**
	OBS-NoPos-InfC	0.048±0.007	0.046±0.008	**0.045±0.008**
	Total Win	0	3	77
C	RCT-50	**0.837±0.008**	0.838±0.008	0.841±0.007
	RCT-5	0.803±0.013	0.804±0.013	**0.793±0.009**
	OBS-CPS	0.829±0.014	0.830±0.014	**0.828±0.014**
	OBS-UConf	**0.835±0.026**	2.701±0.033	0.837±0.027
	OBS-NoPos	**0.845±0.014**	**0.845±0.015**	0.855±0.012
	OBS-CPS-InfC	2.786±0.079	**2.090±0.046**	2.858±0.055
	OBS-UConf-InfC	**2.753±0.074**	2.443±0.023	2.852±0.058
	OBS-NoPos-InfC	2.904±0.061	**2.197±0.045**	3.006±0.034
	Total Win	23	35	22
D	RCT-50	3.303±0.333	**2.241±0.065**	2.624±0.054
	RCT-5	2.897±0.257	**1.845±0.059**	2.192±0.059
	OBS-CPS	3.191±0.449	**2.109±0.062**	2.421±0.068
	OBS-UConf	3.463±0.706	**2.361±0.198**	2.610±0.073
	OBS-NoPos	3.536±0.435	**2.404±0.074**	2.853±0.072
	OBS-CPS-InfC	1.395±0.067	**1.289±0.064**	1.366±0.068
	OBS-UConf-InfC	1.524±0.069	1.737±0.054	**1.511±0.063**
	OBS-NoPos-InfC	1.689±0.074	**1.595±0.069**	1.698±0.073
	Total Win	3	68	9
E	RCT-50	2.672±0.348	**1.595±0.019**	2.033±0.022
	RCT-5	2.238±0.218	**1.468±0.023**	1.823±0.023
	OBS-CPS	2.446±0.262	**1.577±0.022**	1.992±0.032
	OBS-UConf	2.531±0.191	2.651±0.054	**2.051±0.031**
	OBS-NoPos	2.669±0.288	**1.639±0.021**	2.102±0.035
	OBS-CPS-InfC	3.324±0.136	**2.483±0.05**	3.491±0.054
	OBS-UConf-InfC	3.346±0.147	**2.686±0.071**	3.526±0.036
	OBS-NoPos-InfC	3.373±0.101	**2.541±0.038**	3.648±0.072
	Total Win	0	70	10

Table 15: S-Learner MAE

Survival Scenario	Causal Configuration	Base Regression Model		
		Lasso Reg.	Random Forest	XGBoost
A	RCT-50	0.661±0.012	**0.655±0.014**	0.671±0.013
	RCT-5	**0.645±0.011**	0.649±0.012	0.667±0.014
	OBS-CPS	0.653±0.011	**0.646±0.010**	0.662±0.012
	OBS-UConf	**0.604±0.009**	0.608±0.009	0.620±0.007
	OBS-NoPos	0.657±0.010	**0.654±0.011**	0.671±0.013
	OBS-CPS-InfC	**0.727±0.018**	0.730±0.021	0.752±0.023
	OBS-UConf-InfC	**0.675±0.019**	0.693±0.023	0.713±0.023
	OBS-NoPos-InfC	**0.724±0.014**	0.732±0.018	0.755±0.02
	Total Win	44	36	0
B	RCT-50	0.33±0.008	**0.315±0.011**	0.334±0.015
	RCT-5	0.278±0.006	**0.277±0.006**	0.292±0.008
	OBS-CPS	0.315±0.011	**0.307±0.012**	0.324±0.019
	OBS-UConf	0.354±0.007	**0.341±0.008**	0.359±0.011
	OBS-NoPos	0.345±0.009	**0.323±0.011**	0.341±0.011
	OBS-CPS-InfC	0.301±0.007	**0.294±0.006**	0.309±0.008
	OBS-UConf-InfC	0.34±0.004	**0.328±0.005**	0.347±0.006
	OBS-NoPos-InfC	0.337±0.005	**0.319±0.006**	0.337±0.007
	Total Win	3	77	0
C	RCT-50	**1.430±0.022**	1.593±0.025	1.738±0.024
	RCT-5	**1.403±0.019**	1.532±0.018	1.687±0.027
	OBS-CPS	**1.409±0.021**	1.555±0.025	1.706±0.028
	OBS-UConf	**1.419±0.023**	1.563±0.026	1.721±0.020
	OBS-NoPos	**1.453±0.019**	1.612±0.024	1.755±0.022
	OBS-CPS-InfC	**0.931±0.027**	1.011±0.03	1.148±0.045
	OBS-UConf-InfC	**0.895±0.033**	0.966±0.04	1.096±0.059
	OBS-NoPos-InfC	**0.916±0.057**	1.003±0.067	1.123±0.086
	Total Win	80	0	0
D	RCT-50	**0.941±0.18**	1.002±0.202	1.082±0.206
	RCT-5	**1.016±0.121**	1.095±0.123	1.210±0.248
	OBS-CPS	**1.031±0.258**	1.082±0.264	1.188±0.345
	OBS-UConf	**0.985±0.34**	1.015±0.300	1.073±0.383
	OBS-NoPos	**0.967±0.238**	1.03±0.249	1.107±0.295
	OBS-CPS-InfC	**1.146±0.030**	1.148±0.034	1.209±0.044
	OBS-UConf-InfC	1.179±0.024	**1.172±0.031**	1.234±0.032
	OBS-NoPos-InfC	**1.169±0.026**	**1.169±0.026**	1.230±0.028
	Total Win	48	29	3
E	RCT-50	**1.75±0.186**	1.906±0.207	2.124±0.247
	RCT-5	**1.604±0.125**	1.731±0.133	1.901±0.17
	OBS-CPS	**1.651±0.161**	1.799±0.201	1.990±0.243
	OBS-UConf	**1.630±0.127**	1.779±0.159	1.974±0.212
	OBS-NoPos	**1.698±0.139**	1.856±0.162	2.059±0.202
	OBS-CPS-InfC	**0.917±0.089**	1.012±0.113	1.140±0.141
	OBS-UConf-InfC	**0.968±0.130**	1.074±0.164	1.222±0.237
	OBS-NoPos-InfC	**0.928±0.060**	1.012±0.063	1.130±0.071
	Total Win	80	0	0

Table 16: T-Learner MAE

Survival Scenario	Causal Configuration	Base Regression Model (Treated)			Base Regression Model (Control)		
		Lasso Reg.	Random Forest	XGBoost	Lasso Reg.	Random Forest	XGBoost
A	RCT-50	0.669±0.012	**0.652±0.013**	0.680±0.015	**0.652±0.019**	0.657±0.019	0.685±0.019
	RCT-5	0.656±0.044	**0.641±0.049**	0.677±0.050	**0.644±0.011**	0.650±0.012	0.668±0.014
	OBS-CPS	0.711±0.012	**0.697±0.012**	0.727±0.012	**0.588±0.014**	0.595±0.014	0.621±0.018
	OBS-UConf	**0.640±0.009**	0.641±0.007	0.668±0.005	**0.558±0.012**	0.566±0.013	0.593±0.015
	OBS-NoPos	0.568±0.014	**0.561±0.013**	0.587±0.016	**0.733±0.015**	0.747±0.016	0.776±0.016
	OBS-CPS-InfC	0.799±0.020	**0.797±0.022**	0.838±0.023	**0.646±0.023**	0.664±0.025	0.699±0.029
	OBS-UConf-InfC	**0.723±0.023**	0.740±0.031	0.778±0.030	**0.614±0.019**	0.633±0.023	0.668±0.025
	OBS-NoPos-InfC	**0.608±0.018**	0.609±0.016	0.642±0.020	**0.824±0.021**	0.855±0.021	0.895±0.029
	Total Win	28	52	0	77	3	0
B	RCT-50	0.375±0.012	**0.350±0.014**	0.374±0.016	**0.279±0.007**	0.281±0.008	0.303±0.006
	RCT-5	0.393±0.051	**0.383±0.047**	0.404±0.043	**0.271±0.005**	0.273±0.005	0.287±0.006
	OBS-CPS	0.326±0.014	**0.302±0.014**	0.322±0.016	**0.305±0.011**	0.313±0.012	0.338±0.015
	OBS-UConf	0.375±0.008	**0.348±0.009**	0.371±0.011	**0.327±0.008**	0.334±0.007	0.363±0.012
	OBS-NoPos	0.425±0.014	**0.411±0.015**	0.442±0.017	**0.231±0.007**	0.235±0.007	0.253±0.008
	OBS-CPS-InfC	0.311±0.007	**0.292±0.006**	0.311±0.008	**0.291±0.009**	0.297±0.010	0.322±0.011
	OBS-UConf-InfC	0.365±0.004	**0.344±0.007**	0.370±0.009	**0.310±0.007**	0.313±0.007	0.337±0.008
	OBS-NoPos-InfC	0.415±0.009	**0.406±0.009**	0.438±0.011	**0.228±0.005**	0.233±0.006	0.249±0.006
	Total Win	4	76	0	68	12	0
C	RCT-50	**1.534±0.037**	1.653±0.041	1.839±0.046	**1.387±0.029**	1.542±0.020	1.747±0.027
	RCT-5	**1.643±0.091**	1.751±0.085	1.935±0.074	**1.393±0.020**	1.527±0.02	1.688±0.021
	OBS-CPS	**1.484±0.026**	1.599±0.031	1.797±0.036	**1.379±0.030**	1.529±0.029	1.726±0.033
	OBS-UConf	**1.493±0.037**	1.617±0.034	1.809±0.028	**1.363±0.019**	1.528±0.021	1.740±0.027
	OBS-NoPos	**1.568±0.040**	1.665±0.038	1.860±0.037	**1.420±0.037**	1.565±0.027	1.763±0.035
	OBS-CPS-InfC	**0.909±0.047**	1.002±0.052	1.144±0.050	**0.953±0.042**	1.040±0.048	1.198±0.060
	OBS-UConf-InfC	**0.876±0.039**	0.951±0.052	1.088±0.070	**0.916±0.044**	1.004±0.056	1.166±0.079
	OBS-NoPos-InfC	**0.902±0.057**	1.009±0.074	1.149±0.088	**0.928±0.076**	1.016±0.087	1.152±0.088
	Total Win	79	1	0	80	0	0
D	RCT-50	**0.302±0.083**	0.316±0.093	0.339±0.11	**1.546±0.304**	1.691±0.484	1.767±0.429
	RCT-5	**0.284±0.044**	0.296±0.049	0.312±0.052	**1.051±0.127**	1.118±0.129	1.285±0.304
	OBS-CPS	**0.347±0.084**	0.366±0.087	0.384±0.094	**1.651±0.416**	1.758±0.443	1.924±0.594
	OBS-UConf	**0.328±0.134**	0.343±0.130	0.362±0.138	**1.691±0.564**	1.797±0.529	1.809±0.705
	OBS-NoPos	**0.334±0.105**	0.342±0.117	0.366±0.129	**1.571±0.382**	1.659±0.392	1.939±0.612
	OBS-CPS-InfC	1.138±0.029	**1.097±0.027**	1.171±0.028	**1.147±0.041**	1.201±0.051	1.300±0.056
	OBS-UConf-InfC	1.206±0.029	**1.156±0.043**	1.240±0.044	**1.146±0.023**	1.197±0.028	1.298±0.029
	OBS-NoPos-InfC	1.216±0.030	**1.198±0.035**	1.297±0.046	**1.100±0.030**	1.143±0.037	1.222±0.035
	Total Win	45	35	0	66	10	4
E	RCT-50	**1.784±0.230**	1.955±0.255	2.22±0.310	**1.720±0.154**	1.899±0.184	2.108±0.198
	RCT-5	**1.607±0.244**	1.859±0.436	1.969±0.335	**1.605±0.121**	1.733±0.131	1.894±0.147
	OBS-CPS	**1.670±0.194**	1.852±0.261	2.092±0.312	**1.637±0.141**	1.785±0.158	1.993±0.200
	OBS-UConf	**1.650±0.145**	1.809±0.167	2.053±0.239	**1.616±0.130**	1.763±0.144	1.982±0.201
	OBS-NoPos	**1.740±0.161**	1.907±0.200	2.198±0.338	**1.663±0.140**	1.804±0.148	2.027±0.167
	OBS-CPS-InfC	**0.911±0.111**	1.007±0.123	1.146±0.140	**0.925±0.074**	1.015±0.096	1.158±0.104
	OBS-UConf-InfC	**0.953±0.107**	1.049±0.118	1.221±0.146	**0.987±0.159**	1.103±0.222	1.271±0.228
	OBS-NoPos-InfC	**0.949±0.085**	1.043±0.095	1.221±0.112	**0.908±0.046**	1.000±0.044	1.136±0.064
	Total Win	80	0	0	80	0	0

Table 17: X-Learner MAE

Survival Scenario	Causal Configuration	Base Regression Model (Treated)			Base Regression Model (Control)		
		Lasso Reg.	Random Forest	XGBoost	Lasso Reg.	Random Forest	XGBoost
A	RCT-50	**0.620±0.011**	0.622±0.013	0.631±0.013	**0.624±0.019**	**0.624±0.019**	0.631±0.018
	RCT-5	**0.613±0.048**	0.624±0.056	0.634±0.050	0.618±0.012	**0.617±0.012**	0.623±0.010
	OBS-CPS	**0.664±0.011**	0.667±0.010	0.675±0.012	**0.561±0.013**	0.562±0.014	0.568±0.013
	OBS-UConf	**0.608±0.008**	0.610±0.008	0.616±0.008	**0.530±0.012**	0.531±0.012	0.538±0.012
	OBS-NoPos	**0.532±0.013**	0.533±0.013	0.539±0.015	**0.711±0.014**	0.712±0.014	0.718±0.015
	OBS-CPS-InfC	**0.747±0.018**	0.751±0.018	0.763±0.019	**0.616±0.021**	**0.616±0.022**	0.623±0.022
	OBS-UConf-InfC	**0.689±0.024**	0.691±0.023	0.698±0.024	**0.583±0.019**	0.585±0.020	0.592±0.019
	OBS-NoPos-InfC	**0.570±0.016**	**0.570±0.017**	0.578±0.018	**0.800±0.021**	0.802±0.021	0.809±0.020
	Total Win	64	16	0	47	33	0
B	RCT-50	0.339±0.012	**0.328±0.012**	0.333±0.012	0.265±0.007	**0.261±0.007**	0.264±0.007
	RCT-5	0.365±0.045	**0.364±0.051**	0.368±0.046	0.256±0.005	**0.253±0.004**	0.255±0.004
	OBS-CPS	0.296±0.014	**0.283±0.013**	0.287±0.013	**0.290±0.010**	**0.290±0.012**	0.291±0.011
	OBS-UConf	0.339±0.009	**0.327±0.008**	0.331±0.008	0.312±0.008	**0.309±0.006**	0.311±0.008
	OBS-NoPos	0.396±0.013	**0.387±0.014**	0.393±0.013	0.221±0.006	**0.217±0.007**	0.219±0.007
	OBS-CPS-InfC	0.283±0.006	**0.272±0.007**	0.275±0.006	0.277±0.008	**0.275±0.008**	0.278±0.009
	OBS-UConf-InfC	0.332±0.005	**0.321±0.006**	0.326±0.006	0.295±0.008	**0.291±0.007**	0.293±0.007
	OBS-NoPos-InfC	0.388±0.008	**0.379±0.008**	0.386±0.008	0.219±0.005	**0.215±0.005**	0.216±0.005
	Total Win	3	73	4	5	70	5
C	RCT-50	1.542±0.041	1.547±0.032	**1.533±0.033**	1.417±0.025	1.422±0.023	**1.403±0.026**
	RCT-5	1.651±0.092	1.662±0.091	**1.640±0.087**	1.411±0.018	1.414±0.019	**1.400±0.019**
	OBS-CPS	1.492±0.031	1.492±0.031	**1.481±0.026**	1.409±0.032	1.413±0.030	**1.394±0.031**
	OBS-UConf	1.509±0.038	1.510±0.039	**1.497±0.042**	1.396±0.016	1.402±0.014	**1.385±0.016**
	OBS-NoPos	1.562±0.042	1.563±0.038	**1.561±0.038**	1.445±0.033	1.449±0.035	**1.433±0.037**
	OBS-CPS-InfC	**0.909±0.047**	0.915±0.048	**0.909±0.046**	**0.952±0.043**	0.959±0.041	0.954±0.043
	OBS-UConf-InfC	0.876±0.039	0.879±0.042	**0.875±0.039**	**0.916±0.044**	0.921±0.045	0.919±0.047
	OBS-NoPos-InfC	**0.902±0.058**	0.909±0.061	0.904±0.060	**0.928±0.076**	0.935±0.081	**0.928±0.074**
	Total Win	27	11	42	17	7	56
D	RCT-50	0.306±0.088	0.300±0.089	**0.292±0.083**	1.633±0.342	1.613±0.503	**1.506±0.302**
	RCT-5	0.283±0.046	0.280±0.042	**0.277±0.047**	1.124±0.125	1.055±0.128	**1.033±0.150**
	OBS-CPS	0.355±0.085	0.343±0.080	**0.334±0.081**	1.819±0.545	1.681±0.456	**1.613±0.436**
	OBS-UConf	0.336±0.137	0.324±0.126	**0.322±0.123**	1.772±0.571	1.737±0.498	**1.613±0.557**
	OBS-NoPos	0.336±0.110	0.322±0.106	**0.319±0.104**	1.688±0.418	1.607±0.383	**1.540±0.388**
	OBS-CPS-InfC	1.067±0.026	**1.043±0.025**	1.049±0.026	**1.133±0.041**	1.137±0.043	1.136±0.043
	OBS-UConf-InfC	1.116±0.032	**1.098±0.033**	1.103±0.027	**1.135±0.021**	1.138±0.021	1.136±0.021
	OBS-NoPos-InfC	1.160±0.029	**1.137±0.029**	1.140±0.029	**1.085±0.032**	1.088±0.032	1.087±0.031
	Total Win	4	36	40	17	21	42
E	RCT-50	**1.771±0.226**	1.798±0.237	1.772±0.230	1.734±0.154	1.748±0.156	**1.721±0.150**
	RCT-5	**1.607±0.240**	1.731±0.366	1.608±0.245	1.602±0.120	1.606±0.121	**1.597±0.118**
	OBS-CPS	**1.668±0.199**	1.722±0.244	1.668±0.197	1.638±0.138	1.652±0.139	**1.633±0.141**
	OBS-UConf	1.661±0.157	1.659±0.151	**1.646±0.152**	**1.613±0.130**	1.632±0.132	1.635±0.187
	OBS-NoPos	**1.726±0.162**	1.753±0.176	1.773±0.226	1.670±0.137	1.671±0.135	**1.656±0.136**
	OBS-CPS-InfC	**0.911±0.111**	0.919±0.108	0.913±0.107	**0.925±0.074**	0.930±0.081	0.924±0.075
	OBS-UConf-InfC	**0.953±0.107**	0.960±0.114	**0.953±0.106**	0.987±0.159	1.008±0.209	**0.986±0.151**
	OBS-NoPos-InfC	0.949±0.085	**0.947±0.077**	0.951±0.083	**0.908±0.046**	0.923±0.043	0.910±0.047
	Total Win	37	14	29	26	11	43

Table 18: DR-Learner MAE

Survival Scenario	Causal Configuration	Base Regression Model		
		Lasso Reg.	Random Forest	XGBoost
A	RCT-50	0.661±0.012	**0.658±0.013**	0.685±0.013
	RCT-5	**0.645±0.011**	0.652±0.013	0.680±0.014
	OBS-CPS	0.653±0.011	**0.649±0.012**	0.676±0.012
	OBS-UConf	**0.604±0.009**	0.611±0.009	0.635±0.010
	OBS-NoPos	0.657±0.010	**0.656±0.011**	0.684±0.014
	OBS-CPS-InfC	**0.727±0.018**	0.735±0.023	0.770±0.025
	OBS-UConf-InfC	**0.675±0.019**	0.696±0.022	0.730±0.023
	OBS-NoPos-InfC	**0.724±0.014**	0.735±0.015	0.770±0.020
	Total Win	54	26	0
B	RCT-50	0.330±0.008	**0.318±0.011**	0.341±0.015
	RCT-5	**0.278±0.006**	**0.278±0.006**	0.300±0.007
	OBS-CPS	0.316±0.011	**0.308±0.013**	0.332±0.017
	OBS-UConf	0.354±0.007	**0.344±0.008**	0.370±0.010
	OBS-NoPos	0.345±0.009	**0.326±0.011**	0.349±0.012
	OBS-CPS-InfC	0.301±0.007	**0.295±0.006**	0.316±0.008
	OBS-UConf-InfC	0.340±0.004	**0.331±0.004**	0.355±0.006
	OBS-NoPos-InfC	0.337±0.005	**0.321±0.005**	0.345±0.007
	Total Win	6	74	0
C	RCT-50	**1.430±0.022**	1.591±0.021	1.791±0.024
	RCT-5	**1.404±0.019**	1.540±0.018	1.742±0.021
	OBS-CPS	**1.410±0.021**	1.564±0.026	1.763±0.021
	OBS-UConf	**1.420±0.023**	1.572±0.024	1.777±0.019
	OBS-NoPos	**1.454±0.019**	1.614±0.020	1.805±0.015
	OBS-CPS-InfC	**0.931±0.027**	1.021±0.030	1.173±0.038
	OBS-UConf-InfC	**0.895±0.033**	0.977±0.044	1.133±0.054
	OBS-NoPos-InfC	**0.916±0.057**	1.009±0.065	1.172±0.083
	Total Win	80	0	0
D	RCT-50	**0.943±0.180**	0.986±0.187	1.096±0.253
	RCT-5	**1.020±0.120**	1.095±0.118	1.177±0.176
	OBS-CPS	**1.031±0.256**	1.103±0.259	1.158±0.339
	OBS-UConf	**0.986±0.340**	1.024±0.345	1.119±0.357
	OBS-NoPos	**0.969±0.238**	1.038±0.292	1.124±0.243
	OBS-CPS-InfC	**1.146±0.030**	1.154±0.037	1.238±0.048
	OBS-UConf-InfC	1.179±0.024	**1.176±0.031**	1.262±0.034
	OBS-NoPos-InfC	**1.169±0.026**	1.173±0.026	1.260±0.028
	Total Win	65	15	0
E	RCT-50	**1.753±0.185**	1.917±0.206	2.161±0.243
	RCT-5	**1.605±0.127**	1.742±0.144	1.949±0.161
	OBS-CPS	**1.652±0.163**	1.811±0.199	2.033±0.241
	OBS-UConf	**1.632±0.127**	1.785±0.166	2.041±0.255
	OBS-NoPos	**1.701±0.139**	1.859±0.162	2.112±0.202
	OBS-CPS-InfC	**0.917±0.089**	1.016±0.104	1.164±0.114
	OBS-UConf-InfC	**0.969±0.130**	1.080±0.158	1.276±0.240
	OBS-NoPos-InfC	**0.928±0.060**	1.022±0.068	1.182±0.096
	Total Win	80	0	0

Table 19: DR-Learner propensity score AUC. Note that we use the `econml` package in Python, which by default uses logistic regression for predicting the treatment assignment. Thus, we report the AUC of the treatment prediction by the logistic regression.

Causal Configuration	Logistic Regression
RCT-50	0.501±0.005
RCT-5	0.497±0.011
OBS-CPS	0.661±0.007
OBS-UConf	0.548±0.007
OBS-NoPos	0.820±0.005
OBS-CPS-InfC	0.661±0.007
OBS-UConf-InfC	0.548±0.007
OBS-NoPos-InfC	0.820±0.005

Table 20: Survival S-Learner concordance index

Survival Scenario	Causal Configuration	Base Regression Model		
		RSF	DeepSurv	DeepHit
A	RCT-50	0.568±0.008	**0.595±0.003**	0.557±0.007
	RCT-5	0.551±0.008	**0.580±0.004**	0.558±0.006
	OBS-CPS	0.565±0.004	**0.596±0.004**	0.567±0.008
	OBS-UConf	0.556±0.005	**0.587±0.006**	0.558±0.010
	OBS-NoPos	0.565±0.009	**0.594±0.004**	0.553±0.006
	OBS-CPS-InfC	0.563±0.005	**0.597±0.004**	0.546±0.010
	OBS-UConf-InfC	0.557±0.006	**0.585±0.006**	0.538±0.008
	OBS-NoPos-InfC	0.562±0.006	**0.591±0.003**	0.539±0.008
	Total Win	0	80	0
B	RCT-50	0.640±0.003	**0.645±0.004**	**0.645±0.004**
	RCT-5	0.616±0.003	**0.622±0.005**	0.621±0.004
	OBS-CPS	0.631±0.005	**0.632±0.003**	0.631±0.003
	OBS-UConf	0.632±0.005	**0.634±0.005**	**0.634±0.004**
	OBS-NoPos	0.650±0.003	**0.656±0.002**	**0.656±0.002**
	OBS-CPS-InfC	0.630±0.004	**0.632±0.004**	0.629±0.003
	OBS-UConf-InfC	0.630±0.004	**0.633±0.005**	0.631±0.005
	OBS-NoPos-InfC	0.649±0.003	**0.655±0.003**	0.654±0.003
	Total Win	10	50	20
C	RCT-50	0.545±0.009	**0.576±0.004**	0.570±0.005
	RCT-5	0.522±0.007	**0.554±0.007**	0.540±0.014
	OBS-CPS	0.538±0.006	**0.573±0.005**	0.562±0.004
	OBS-UConf	0.536±0.007	**0.566±0.007**	0.561±0.008
	OBS-NoPos	0.550±0.007	**0.583±0.005**	0.575±0.007
	OBS-CPS-InfC	0.498±0.015	**0.558±0.026**	0.546±0.017
	OBS-UConf-InfC	0.502±0.023	**0.560±0.029**	0.541±0.020
	OBS-NoPos-InfC	0.511±0.029	**0.586±0.019**	0.561±0.023
	Total Win	0	70	10
D	RCT-50	0.633±0.027	0.676±0.021	**0.696±0.013**
	RCT-5	0.569±0.019	0.626±0.017	**0.628±0.011**
	OBS-CPS	0.610±0.029	0.668±0.019	**0.683±0.011**
	OBS-UConf	0.634±0.027	**0.702±0.015**	0.696±0.018
	OBS-NoPos	0.615±0.032	0.678±0.016	**0.683±0.015**
	OBS-CPS-InfC	0.626±0.011	**0.634±0.005**	0.629±0.007
	OBS-UConf-InfC	0.639±0.005	**0.646±0.005**	0.643±0.007
	OBS-NoPos-InfC	0.635±0.006	**0.644±0.006**	0.640±0.005
	Total Win	4	40	36
E	RCT-50	0.544±0.010	**0.591±0.011**	0.578±0.011
	RCT-5	0.513±0.009	**0.554±0.015**	0.547±0.012
	OBS-CPS	0.538±0.013	**0.583±0.010**	0.566±0.018
	OBS-UConf	0.533±0.016	**0.574±0.018**	0.567±0.017
	OBS-NoPos	0.544±0.015	**0.599±0.010**	0.589±0.012
	OBS-CPS-InfC	0.482±0.041	**0.546±0.030**	0.538±0.028
	OBS-UConf-InfC	0.445±0.029	**0.542±0.045**	0.534±0.017
	OBS-NoPos-InfC	0.474±0.017	**0.565±0.035**	0.563±0.022
	Total Win	0	60	20

Table 21: Survival T-Learner concordance index

Survival Scenario	Causal Configuration	Base Regression Model (Treated)			Base Regression Model (Control)		
		RSF	DeepSurv	DeepHit	RSF	DeepSurv	DeepHit
A	RCT-50	0.579±0.009	**0.612±0.006**	0.581±0.015	0.546±0.010	**0.578±0.007**	0.549±0.014
	RCT-5	0.567±0.031	**0.604±0.018**	0.592±0.025	0.549±0.007	**0.581±0.005**	0.557±0.012
	OBS-CPS	0.569±0.008	**0.603±0.009**	0.582±0.011	0.546±0.006	**0.577±0.007**	0.548±0.009
	OBS-UConf	0.546±0.008	**0.579±0.010**	0.553±0.012	0.557±0.009	**0.585±0.007**	0.554±0.009
	OBS-NoPos	0.567±0.009	**0.598±0.008**	0.564±0.016	0.534±0.006	**0.564±0.009**	0.544±0.005
	OBS-CPS-InfC	0.569±0.006	**0.602±0.009**	0.559±0.018	0.546±0.007	**0.578±0.006**	0.541±0.017
	OBS-UConf-InfC	0.546±0.009	**0.578±0.008**	0.540±0.012	0.555±0.012	**0.584±0.006**	0.538±0.019
	OBS-NoPos-InfC	0.564±0.009	**0.598±0.006**	0.545±0.013	0.531±0.009	**0.564±0.007**	0.537±0.010
	Total Win	0	75	5	0	80	0
B	RCT-50	0.651±0.004	**0.656±0.004**	0.654±0.005	0.610±0.005	**0.618±0.005**	0.616±0.005
	RCT-5	**0.628±0.017**	0.627±0.027	0.609±0.022	0.610±0.007	0.619±0.004	**0.620±0.004**
	OBS-CPS	0.630±0.006	**0.637±0.005**	0.634±0.005	0.605±0.006	**0.612±0.006**	0.607±0.007
	OBS-UConf	0.644±0.008	**0.648±0.006**	0.646±0.006	0.598±0.005	**0.605±0.007**	0.602±0.008
	OBS-NoPos	0.628±0.005	**0.636±0.004**	0.631±0.005	0.593±0.008	**0.601±0.004**	0.600±0.007
	OBS-CPS-InfC	0.628±0.005	**0.633±0.005**	0.632±0.003	0.603±0.006	**0.611±0.008**	0.610±0.007
	OBS-UConf-InfC	0.642±0.004	0.645±0.005	**0.646±0.005**	0.598±0.009	**0.605±0.004**	0.603±0.007
	OBS-NoPos-InfC	0.624±0.007	**0.634±0.005**	0.628±0.005	0.592±0.006	**0.602±0.004**	0.600±0.007
	Total Win	14	44	22	7	51	22
C	RCT-50	0.532±0.015	**0.565±0.013**	0.557±0.015	0.512±0.008	**0.541±0.011**	0.524±0.010
	RCT-5	0.536±0.031	0.541±0.050	**0.544±0.027**	0.518±0.014	**0.547±0.010**	0.541±0.007
	OBS-CPS	0.536±0.008	**0.568±0.009**	0.555±0.011	0.516±0.007	**0.540±0.011**	0.523±0.009
	OBS-UConf	0.537±0.010	**0.568±0.012**	0.557±0.007	0.509±0.010	**0.533±0.012**	0.521±0.012
	OBS-NoPos	0.524±0.016	**0.555±0.012**	0.543±0.014	0.521±0.012	**0.547±0.011**	0.535±0.008
	OBS-CPS-InfC	0.487±0.041	**0.550±0.032**	0.542±0.038	0.497±0.018	**0.537±0.021**	0.522±0.033
	OBS-UConf-InfC	0.491±0.044	**0.552±0.031**	0.550±0.032	0.484±0.021	0.513±0.035	**0.519±0.018**
	OBS-NoPos-InfC	0.470±0.045	**0.549±0.037**	0.544±0.032	0.489±0.029	**0.538±0.023**	0.513±0.033
	Total Win	3	57	20	0	64	16
D	RCT-50	0.646±0.038	0.683±0.084	**0.727±0.024**	0.565±0.025	0.614±0.018	**0.623±0.025**
	RCT-5	0.447±0.174	0.412±0.158	**0.672±0.135**	0.573±0.019	**0.626±0.016**	0.625±0.015
	OBS-CPS	0.584±0.052	0.646±0.038	**0.672±0.022**	0.543±0.023	0.609±0.032	**0.620±0.024**
	OBS-UConf	0.655±0.038	0.731±0.028	**0.745±0.012**	0.536±0.034	0.588±0.036	**0.597±0.029**
	OBS-NoPos	0.668±0.051	0.658±0.085	**0.773±0.024**	0.547±0.027	**0.593±0.019**	0.586±0.021
	OBS-CPS-InfC	0.632±0.010	**0.639±0.005**	0.637±0.006	0.556±0.010	**0.575±0.010**	0.566±0.010
	OBS-UConf-InfC	0.673±0.005	0.675±0.007	**0.676±0.007**	0.549±0.011	**0.569±0.007**	0.556±0.008
	OBS-NoPos-InfC	0.664±0.009	**0.671±0.007**	0.668±0.007	0.549±0.007	**0.563±0.006**	0.553±0.009
	Total Win	5	25	50	2	49	29
E	RCT-50	0.539±0.020	**0.589±0.024**	0.575±0.017	0.514±0.020	**0.547±0.019**	0.537±0.014
	RCT-5	0.481±0.065	**0.518±0.047**	0.516±0.065	0.518±0.011	**0.554±0.010**	0.544±0.013
	OBS-CPS	0.533±0.021	**0.574±0.022**	0.562±0.020	0.508±0.018	**0.544±0.015**	0.535±0.014
	OBS-UConf	0.534±0.023	**0.587±0.024**	0.552±0.014	0.510±0.014	0.520±0.023	**0.531±0.022**
	OBS-NoPos	0.520±0.024	0.539±0.032	**0.547±0.024**	0.516±0.020	**0.546±0.016**	0.534±0.015
	OBS-CPS-InfC	0.485±0.047	**0.551±0.042**	0.520±0.034	0.454±0.038	**0.515±0.045**	0.508±0.042
	OBS-UConf-InfC	0.437±0.064	0.525±0.048	**0.541±0.065**	0.455±0.025	0.495±0.037	**0.499±0.041**
	OBS-NoPos-InfC	0.464±0.046	**0.520±0.038**	0.505±0.040	0.453±0.043	0.514±0.027	**0.537±0.023**
	Total Win	2	53	25	5	43	32

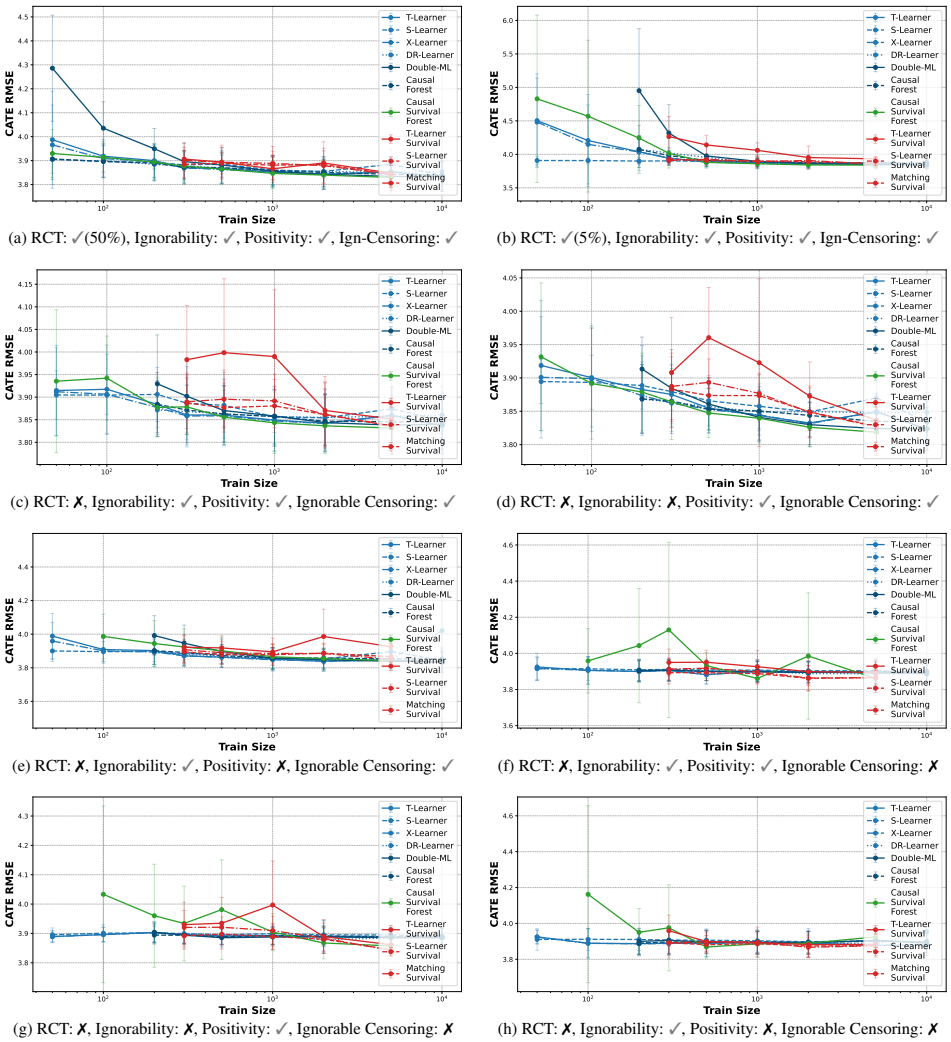
Table 22: Survival Matching-Learner concordance index

Survival Scenario	Causal Configuration	Base Survival Model		
		RSF	DeepSurv	DeepHit
A	RCT-50	0.568±0.008	**0.595±0.003**	0.557±0.007
	RCT-5	0.551±0.008	**0.580±0.004**	0.558±0.006
	OBS-CPS	0.556±0.005	**0.596±0.004**	0.567±0.008
	OBS-UConf	0.556±0.005	**0.587±0.006**	0.558±0.010
	OBS-NoPos	0.565±0.009	**0.594±0.004**	0.553±0.006
	OBS-CPS-InfC	0.563±0.005	**0.597±0.004**	0.546±0.010
	OBS-UConf-InfC	0.557±0.006	**0.585±0.006**	0.538±0.008
	OBS-NoPos-InfC	0.562±0.006	**0.591±0.003**	0.539±0.008
	Total Win	0	80	0
B	RCT-50	0.640±0.003	**0.645±0.004**	**0.645±0.004**
	RCT-5	0.616±0.003	**0.622±0.005**	0.621±0.004
	OBS-CPS	0.631±0.005	**0.632±0.003**	0.631±0.003
	OBS-UConf	0.632±0.005	**0.634±0.005**	**0.634±0.004**
	OBS-NoPos	0.650±0.003	**0.656±0.002**	**0.656±0.002**
	OBS-CPS-InfC	0.630±0.004	**0.632±0.004**	0.629±0.003
	OBS-UConf-InfC	0.630±0.004	**0.633±0.005**	0.631±0.005
	OBS-NoPos-InfC	0.649±0.003	**0.655±0.003**	0.654±0.003
	Total Win	9	50	21
C	RCT-50	0.545±0.009	**0.576±0.004**	0.570±0.005
	RCT-5	0.522±0.007	**0.554±0.007**	0.540±0.014
	OBS-CPS	0.538±0.006	**0.573±0.005**	0.562±0.004
	OBS-UConf	0.536±0.007	**0.566±0.007**	0.561±0.008
	OBS-NoPos	0.550±0.007	**0.583±0.005**	0.575±0.007
	OBS-CPS-InfC	0.498±0.015	**0.558±0.026**	0.546±0.017
	OBS-UConf-InfC	0.502±0.023	**0.560±0.029**	0.541±0.020
	OBS-NoPos-InfC	0.511±0.029	**0.586±0.019**	0.561±0.023
	Total Win	0	70	10
D	RCT-50	0.633±0.027	0.676±0.021	**0.696±0.013**
	RCT-5	0.569±0.019	0.626±0.017	**0.628±0.011**
	OBS-CPS	0.610±0.029	0.668±0.019	**0.683±0.011**
	OBS-UConf	0.634±0.027	**0.702±0.015**	0.696±0.018
	OBS-NoPos	0.615±0.032	0.678±0.016	**0.683±0.015**
	OBS-CPS-InfC	0.626±0.011	**0.634±0.005**	0.629±0.007
	OBS-UConf-InfC	0.639±0.005	**0.646±0.005**	0.643±0.007
	OBS-NoPos-InfC	0.635±0.006	**0.644±0.006**	0.640±0.005
	Total Win	4	40	36
E	RCT-50	0.544±0.010	**0.591±0.011**	0.578±0.011
	RCT-5	0.513±0.009	**0.554±0.015**	0.547±0.012
	OBS-CPS	0.538±0.013	**0.583±0.010**	0.566±0.018
	OBS-UConf	0.533±0.016	**0.574±0.018**	0.567±0.017
	OBS-NoPos	0.544±0.015	**0.599±0.010**	0.589±0.012
	OBS-CPS-InfC	0.482±0.041	**0.546±0.030**	0.538±0.028
	OBS-UConf-InfC	0.445±0.029	**0.542±0.045**	0.534±0.017
	OBS-NoPos-InfC	0.474±0.017	**0.565±0.035**	0.563±0.022
	Total Win	0	60	20

F.7 CONVERGENCE RESULTS

Figure 17 presents the convergence behavior of different causal inference methods under eight configurations of assumptions, all within Scenario C that was the main focus in Section 4.1. The x-axis shows increasing training set sizes (ranging from 50 to 10,000), while the y-axis plots the root mean squared error (RMSE) of the estimated CATE on the test set. (Note that, all models are selected based on performance on the validation set).

Across all configurations, we observe general convergence trends where CATE RMSE decreases as training size increases. Among the survival methods, the T-learner Survival consistently converges the slowest, especially under small training sizes. This may be due to the model requiring sufficient uncensored samples per treatment arm to function effectively. Double-ML also tends to require more data to stabilize, particularly in the presence of low treatment rate or lack of positivity. The Causal Survival Forest shows slower convergence under settings with non-ignorable censoring or positivity violations, reflecting its convergence sensitivity to these assumptions despite its nonparametric structure. Overall, while standard meta-learners and tree-based methods show relatively stable convergence behavior, survival-specific adaptations appear more data-hungry and assumption-sensitive for convergence. These trends highlight the importance of choosing appropriately robust methods with respect to the dataset size in practice, especially in real-world settings where assumptions like positivity or ignorability may be compromised.



(a) RCT: ✓(50%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓

(b) RCT: ✓(5%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓

(c) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✓

(d) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✓

(e) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✓

(f) RCT: ✗, Ignorability: ✓, Positivity: ✓, Ignorable Censoring: ✗

(g) RCT: ✗, Ignorability: ✗, Positivity: ✓, Ignorable Censoring: ✗

(h) RCT: ✗, Ignorability: ✓, Positivity: ✗, Ignorable Censoring: ✗

Figure 17: Convergence properties: CATE RMSE in Scenario C as number of training data increases.

G SEMI-SYNTHETIC DATASETS: SETUP AND ADDITIONAL RESULTS

G.1 SEMI-SYNTHETIC DATASETS SETUP

To complement synthetic benchmarks and real-world case studies, we construct semi-synthetic datasets that pair real covariates with simulated treatments and survival outcomes. This strategy preserves realistic covariate distributions and correlations while enabling controlled evaluation against known ground-truth CATEs. Below, we describe the ACTG- and MIMIC-based semi-synthetic datasets and provide detailed summary statistics. Table 23 reports overall dataset sizes, covariate counts, and censoring/treatment rates.

Table 23: Semi-synthetic dataset overview.

	Data size	**No. covariates**	**Censoring rate**	**Treatment Rate**
ACTG semi-synthetic	2,139	23	51.19%	56.15%
MIMIC-i semi-synthetic	25,170	36	88.49%	49.92%
MIMIC-ii semi-synthetic	25,170	36	81.65%	49.92%
MIMIC-iii semi-synthetic	25,170	36	74.10%	49.92%
MIMIC-iv semi-synthetic	25,170	36	66.34%	49.92%
MIMIC-v semi-synthetic	25,170	36	53.35%	49.92%

G.1.1 ACTG SEMI-SYNTHETIC DATASET

The ACTG semi-synthetic dataset is derived from the ACTG 175 HIV clinical trial (Hammer et al., 1996), which contains 23 baseline covariates. Following the construction procedure of Chapfuwa et al. (2021), we simulate treatment assignments and event times. This dataset captures realistic treatment imbalance and moderate censoring ($\sim$51%). It serves as a smaller-scale but clinically grounded benchmark, preserving the covariate structures observed in trial participants.

G.1.2 MIMIC SEMI-SYNTHETIC DATASETS

The second family of datasets is derived from MIMIC-IV ICU records (Johnson et al., 2023). We extract 36 covariates that span laboratory test abnormalities (e.g., creatinine, glucose, hemoglobin), demographic features (e.g., age, sex, race, marital status), and admission descriptors (e.g., admission type, recurrent admissions, night admission). Treatments (W) are simulated as Bernoulli(0.5), and event times are generated following the formulation of Meir et al. (2025), where baseline hazards depend on subsets of laboratory and demographic covariates. Five variants are created by altering the censoring distribution, resulting in censoring rates ranging from 53% to 88%. This design mimics the range of censoring observed in longitudinal EHR studies, from moderate censoring to highly censored survival outcomes.

Tables 24 and 25 provide detailed covariate statistics and demographic distributions for the MIMIC datasets and Figure 18 provides an overview of correlation among the MIMIC covariates.

Treatment assignment. Treatment is assigned independently as a Bernoulli random variable with probability 0.5:

$$W \sim \text{Bernoulli}(0.5).$$

This ensures balanced treatment groups while maintaining independence from baseline covariates.

Potential outcomes. Let $X_{1:5}$ denote the first five binary covariates corresponding to abnormal laboratory values (*Anion gap*, *Bicarbonate*, *Calcium total*, *Chloride*, *Creatinine*), and let X_{36} denote the standardized *Admission age*. The sum of abnormal indicators is written as

$$S = \sum_{j=1}^{5} X_j.$$

Table 24: Summary statistics of MIMIC semi-synthetic covariates. Reported values are mean ± standard deviation. Physiological covariates are coded as *indicators for abnormal values* where mean reflects prevalence of abnormality.

Covariate	Mean ± Std	**Covariate**	Mean ± Std
Sodium	0.12 ± 0.32	Admission age	61.39 ± 17.97
Potassium	0.08 ± 0.28	Sex:Male	0.51 ± 0.50
Chloride	0.19 ± 0.39	Race:White	0.70 ± 0.46
Bicarbonate	0.24 ± 0.43	Race:Black	0.14 ± 0.35
Anion gap	0.09 ± 0.29	Race:Hispanic	0.05 ± 0.22
Creatinine	0.28 ± 0.45	Race:Other	0.07 ± 0.25
Urea nitrogen	0.40 ± 0.49	Insurance:Medicare	0.42 ± 0.49
Glucose	0.65 ± 0.48	Insurance:Other	0.52 ± 0.50
Calcium total	0.29 ± 0.45	Marital status:Married	0.45 ± 0.50
Magnesium	0.09 ± 0.28	Marital status:Single	0.33 ± 0.47
Phosphate	0.28 ± 0.45	Marital status:Widowed	0.14 ± 0.34
Hemoglobin	0.73 ± 0.44	Direct emergency:Yes	0.11 ± 0.31
Hematocrit	0.69 ± 0.46	Night admission:Yes	0.54 ± 0.50
MCV	0.20 ± 0.40	Previous admission this month: Yes	0.08 ± 0.27
MCH	0.26 ± 0.44	Admissions number:2	0.16 ± 0.37
MCHC	0.31 ± 0.46	Admissions number:3+	0.22 ± 0.42
Platelet count	0.29 ± 0.45		
RDW	0.29 ± 0.45		
White blood cells	0.40 ± 0.49		
Red blood cells	0.76 ± 0.43		

Potential survival times under control ($T(0)$) and treatment ($T(1)$) are drawn from Poisson distributions with means linearly dependent on S and X_{36}:

$$T(0) \sim \text{Poisson}(30 + 0.75S + 0.75X_{36}),$$
$$T(1) \sim \text{Poisson}(30 + 0.75X_{36} - 0.45).$$

The individual treatment effect is defined as

$$\tau(x) = \mathbb{E}[T(1) - T(0) \mid X = x].$$

We record the true CATE for each unit as $T(1) - T(0)$.

Observed outcome. The realized survival time depends on treatment assignment:

$$T = W \cdot T(1) + (1 - W) \cdot T(0).$$

Censoring. Censoring times are drawn independently from a Poisson distribution with mean parameter λ_c:

$$C \sim \text{Poisson}(\lambda_c),$$

where $\lambda_c \in \{21, 23, 24.7, 26.5, 29\}$ controls the censoring severity across the five dataset variants (MIMIC-[i–v]).

Final observed data. For each individual, we define the observed time and event indicator as

$$Y = \min(T, C),$$
$$\delta = \mathbb{1}\{T \leq C\},$$

where Y is the observed follow-up time and δ is the event indicator (1 if the event was observed, 0 if censored).

Table 25: Demographic and categorical distributions in MIMIC semi-synthetic datasets. Reported values are proportions.

Demographics

Variable	Proportion
Sex	
Male	0.512
Female	0.488
Race	
White	0.699
Black	0.141
Other	0.066
Hispanic	0.053
Asian	0.041
Insurance	
Other	0.522
Medicare	0.421
Medicaid	0.057
Marital status	
Married	0.449
Single	0.334
Widowed	0.136
Divorced	0.081

Admission-related

Variable	Proportion
Direct emergency	
Yes	0.110
No	0.890
Night admission	
Yes	0.539
No	0.461
Previous admission this month	
Yes	0.081
No	0.919
Admissions number	
1	0.615
2	0.164
3+	0.222

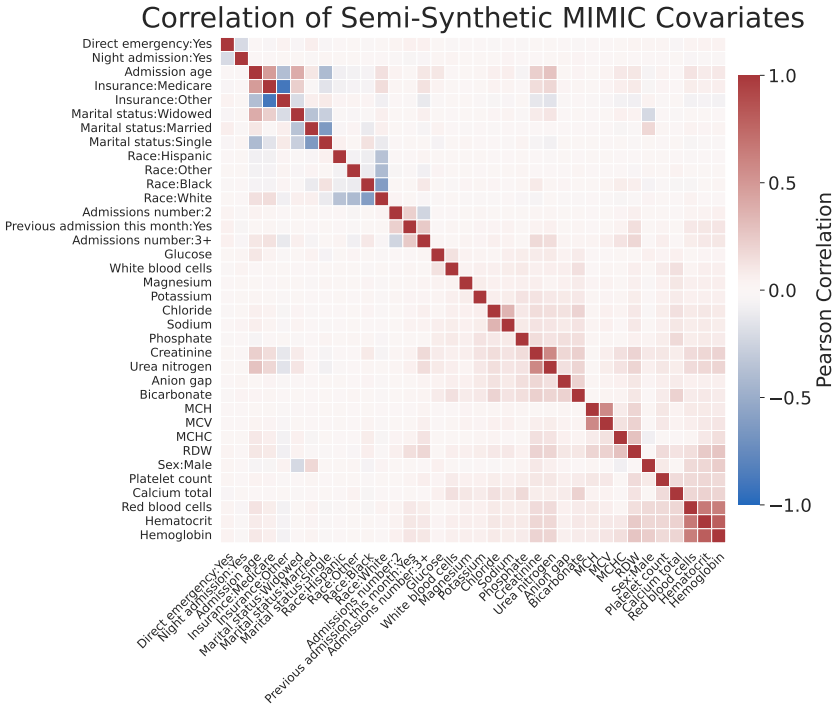


Figure 18: Correlation heatmap of the 36 semi-synthetic MIMIC covariates. Variables include demographic features, admission descriptors, insurance and marital status indicators, and laboratory measurements. Most correlations are weak to moderate, with stronger dependencies visible among related laboratory values (e.g., hematocrit, hemoglobin, and red blood cell count).

G.2 Detailed Analysis of Semi-Synthetic Results

This section provides additional analysis of semi-synthetic results in Section 4.2 and Table 3, examining performance patterns, stability characteristics, and practical implications for method selection.

Method-Specific Performance Analysis. The ACTG dataset reveals clear performance hierarchies. Double-ML achieves the lowest RMSE (10.651 ± 0.239), representing a 3.9% improvement over the next-best method, X-Learner (11.072 ± 0.196). This advantage aligns with our synthetic findings where sophisticated causal machinery excels in moderate-dimensional settings. However, survival meta-learners struggle on ACTG, with S-Learner Survival (11.713 ± 0.237) and Matching Survival (12.523 ± 0.289) showing the worst performance.

The pattern reverses on MIMIC data. S-Learner Survival achieves the best performance on four of five MIMIC variants: MIMIC-i (7.921 ± 0.044), MIMIC-iii (7.900 ± 0.045), MIMIC-iv (7.901 ± 0.046), and MIMIC-v (7.897 ± 0.042). This consistency demonstrates robustness across varying censoring intensities in high-dimensional settings.

Censoring Gradient Analysis. The MIMIC censoring rate range (53% - 88%) enables analysis of degradation patterns. S-Learner Survival maintains consistent performance across this range, with RMSE ranging from 7.897 to 7.921—less than 0.3% variation despite 35 percentage points of censoring difference. In contrast, T-Learner Survival shows clear instability, particularly at MIMIC-ii (82% censoring), where standard deviation jumps to ±0.233, indicating unreliable estimates.

Double-ML exhibits an interesting non-monotonic pattern: performing well at extreme censoring (MIMIC-i: 7.954 ± 0.047) and low censoring (MIMIC-v: 7.891 ± 0.050) but showing degradation at intermediate levels. This suggests that Double-ML's robustness may depend on specific censoring-covariate interactions rather than censoring rate alone.

Variance and Stability Patterns. Standard deviations reveal important stability trade-offs. On ACTG, imputation methods show relatively low variance (0.175-0.239 range) while survival methods exhibit higher variability (0.160-0.289). This pattern suggests that survival-specific methods may be more sensitive to the particular covariate-outcome relationships present in clinical trial data.

However, this relationship inverts on MIMIC data. Survival meta-learners achieve consistently low variance (0.042-0.075 range), while imputation methods show slightly higher variability (0.043-0.050 range). The exception is T-Learner Survival's instability at high censoring, which appears dataset-specific rather than method-inherent.

Performance Convergence in High-Dimensional Settings. The MIMIC results demonstrate performance convergence absent in synthetic experiments. Across all MIMIC variants, the RMSE range spans only 7.891 to 8.007—approximately 1.4% variation. This convergence contrasts sharply with ACTG's 15% performance spread (10.651 to 12.523), suggesting that high-dimensional, realistic covariate structures may level the playing field between method families.

This convergence has practical implications: in EHR-like settings with many correlated covariates, method selection may prioritize stability and interpretability over raw performance, since performance differences become negligible.

Cross-Dataset Generalization Challenges. No single method achieves consistent top-tier performance across both datasets. Double-ML excels on ACTG but performs moderately on MIMIC. S-Learner Survival dominates MIMIC but struggles on ACTG. This inconsistency highlights a critical limitation of synthetic-only evaluation: performance rankings established on one data structure may not transfer to others, even when both represent realistic medical scenarios.

Practical Method Selection Guidelines. Based on these findings, a suggested actionable recommendation can be formed as follows: (1) For clinical trial-like data with moderate dimensionality and balanced censoring, prioritize Double-ML or X-Learner. (2) For EHR-like data with high dimensionality and variable censoring, S-Learner Survival provides the best combination of performance and stability. (3) Avoid T-Learner Survival in high-censoring scenarios due to variance instability. (4) When performance differences are small ($<$2%), prioritize methods with lower computational cost and better interpretability.

These semi-synthetic results demonstrate that while our synthetic benchmark captures important performance trends, real-world method selection requires considering dataset-specific characteristics that pure synthetic evaluation cannot fully capture.

H REAL-WORLD DATASETS: SETUP AND ADDITIONAL RESULTS

We evaluate our benchmark on two real-world datasets: the Twins dataset (with known ground truth) and the ACTG 175 HIV clinical trial dataset (without known ground truth). This section provides detailed descriptions of data preprocessing and additional experimental results.

H.1 TWINS DATASET

The Twins dataset is derived from all births in the USA between 1989-1991 (Almond et al., 2005) focusing on twin births. Following Curth et al. (2021), we artificially create a binary treatment where $W = 1$ ($W = 0$) denotes being born the heavier (lighter) twin. The outcome of interest is the time-to-mortality (in days) of each twin in their first year, administratively censored at $t = 365$ days. Since we have records for both twins, we treat their time-to-event outcomes as two potential outcomes $\tau(1)$ and $\tau(0)$ with respect to the treatment assignment of being born heavier.

We obtained 30 features (43 feature dimensions after one-hot encoding categorical features) for each twin relating to the parents, pregnancy, and birth characteristics including marital status, race, residence, number of previous births, pregnancy risk factors, quality of care during pregnancy, and number of gestation weeks prior to birth. We select only twins weighing less than 2kg and without missing features, resulting in more than 11,000 twin pairs.

To create an observational time-to-event dataset with known ground truth, we follow the semi-synthetic experimental design from Curth et al. (2021). The treatment assignment is given by $W|x \sim \text{Bernoulli}(\sigma(\beta_1^\top x + e))$ where $\beta_1 \sim \text{Uniform}(-0.1, 0.1)^{43\times 1}$ and $e \sim \mathcal{N}(0, 1^2)$. The time-to-censoring is given by $C \sim \text{Exp}(100 \cdot \sigma(\beta_2^\top x))$ where $\beta_2 \sim \mathcal{N}(0, 1^2)$. This results in a treatment rate of 68.1% and censoring rate of 38.2%.

We split the data 50/25/25 for training/validation/testing samples and repeat all the experiments 10 times with different random splits. CATE RMSE are reported on the testing sets. In Section 4.3, we display the CATE RMSE with horizon $h = 30$ days. Here, we show CATE RMSE results for the Twins dataset with horizon $h = 180$ days in Figure 19, and we can see it indicates similar results as $h = 30$.

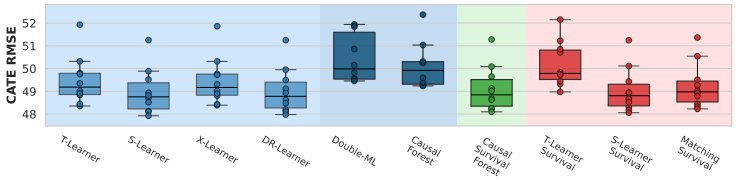


Figure 19: CATE RMSE for twin birth data using different estimator families with $h = 180$. Box plots show the distribution of error across 10 experimental runs.

H.2 ACTG 175 HIV CLINICAL TRIAL DATASET

We use data from the AIDS Clinical Trials Group Protocol 175 (ACTG 175) (Hammer et al., 1996), a double-blind, randomized controlled trial that compared four treatment regimens in adults infected with HIV type I: monotherapy with zidovudine (ZDV), monotherapy with didanosine (ddI), combination therapy with ZDV and ddI, or combination therapy with ZDV and zalcitabine (Zal). The publicly available dataset[3] includes 2,139 HIV-infected patients randomized into four groups with assigned treatments: ZDV, ZDV+ddI, ZDV+Zal, and ddI. An event occurrence was defined as the first of either a decline in CD4 cell count, an event indicating AIDS progression, or death.

Following Meir et al. (2025), after fetching raw data from the UCI Machine Learning Repository, we change the resolution from days to months and add synthetic censoring based on a Bernoulli distribution with parameter $p = 0.6 + 0.25 \cdot Z30$, where $Z30$ is a feature that is available in the data and indicates whether a patient started taking ZDV prior to the assigned treatment, and it is not included in the covariates for CATE estimation. We conduct three pairwise comparisons with ZDV as the baseline treatment ($W = 0$): ZDV vs. ZDV+ddI (HIV1), ZDV vs. ZDV+Zal (HIV2), and

[3] https://archive.ics.uci.edu/dataset/890/aids+clinical+trials+group+study+175

ZDV vs. ddI (HIV3). The baseline censoring rate is less than 15% for different treatment groups. After applying the censoring injection procedure from Meir et al. (2025), increasing censoring rates to over 90%. For each treatment group, we establish baseline CATE estimates by running Causal Survival Forest 10 times and averaging the estimated conditional average treatment effects. Since there are many variants of outcome imputation and survival meta-learner families due to different imputation and base learner options, for display purposes in the HIV dataset results, we use a model selection criterion based on closeness (CATE RMSE) to estimation by Causal Survival Forest. We have looked the results using other variants of same CATE estimator as well, and similar trends are observed.

In Section 4.3, we display the comparisons of CATE estimates between baseline and high-censoring conditions for group HIV1. Here we display the same sets of results for HIV2 and HIV3 groups in Figure 20, 21. Consistent patterns emerge across all three treatment comparisons: Causal Survival Forest produces estimates that cluster tightly around their baseline CATE estimations on data before additional censoring injection; outcome imputation methods show higher variation in baseline estimates but more concentrated predictions under high censoring, and survival meta-learners display substantial deviations from the 45-degree line, indicating sensitivity to censoring conditions. The consistency of these patterns across different treatment pairs reinforces the robustness of our findings regarding how different estimator families respond to increased censoring.

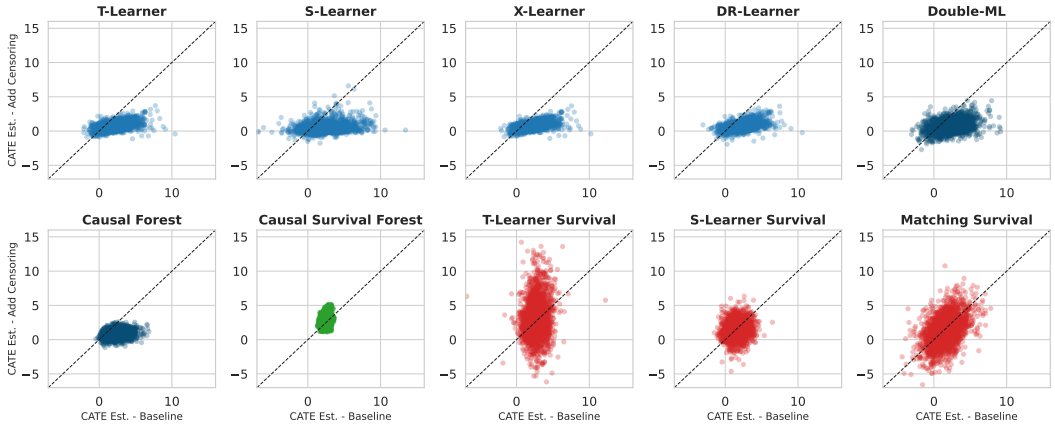


Figure 20: CATE Estimation comparison between baseline and high-censoring conditions under ZDV vs. ZDV+Zal treatments (HIV2). Each point represents an individual patient in test sets, with the dashed diagonal line indicating perfect consistency between baseline CATE estimation and that with the additional censoring injected.

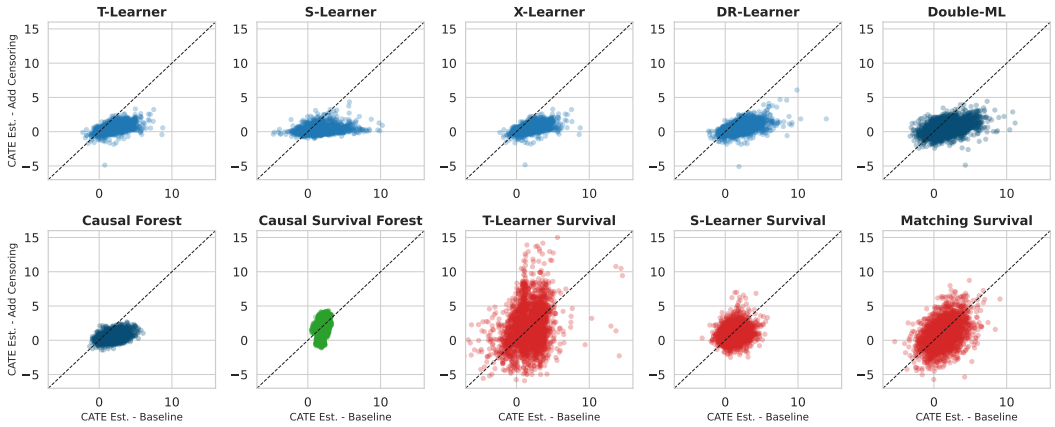


Figure 21: CATE Estimation comparison between baseline and high-censoring conditions under ZDV vs. ddI treatments (HIV3). Each point represents an individual patient in test sets, with the dashed diagonal line indicating perfect consistency between baseline CATE estimation and that with the additional censoring injected.

I Additional Informative Censoring via Unobserved Confounding

In the main paper, we model informative censoring by making censoring times stochastically dependent on event times, reflecting realistic scenarios where patients with shorter expected survival may drop out earlier. Here we complement this setting with an alternative mechanism where the ignorable censoring assumption is violated due to unobserved confounding. This extension illustrates the extensibility of our modular data generation framework.

Data generation process. We follow the same covariate generation procedure as in our synthetic datasets: observed covariates $X \sim \text{Uniform}(0,1)^5$ and an unobserved covariate $U \sim \text{Uniform}(0,1)$. Treatment assignment follows the `OBS-UConf` configuration, where U enters into both treatment assignment and outcome generation but remains unobserved during estimation.

We focus on survival Scenario `C` (Poisson hazards with medium censoring). Event times and censoring times are generated as follows, where $w \in \{0,1\}$ is the treatment indicator:

$$\lambda(w) = X_2^2 + X_3 + 6 + 2\left(\sqrt{0.3 \cdot X_1 + 0.7 \cdot U} - 0.3\right) \cdot w + \epsilon, \tag{6}$$

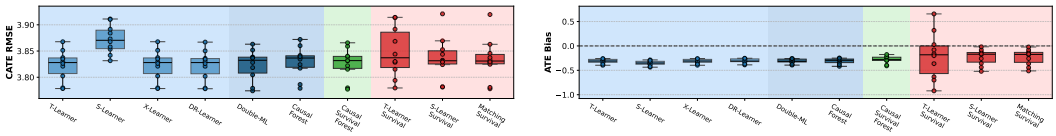
$$T(w) \sim \text{Poisson}(\lambda(w)), \tag{7}$$

$$C = \begin{cases} \infty & \text{if } U \leq 0.6, \\ 1 + \mathbb{1}(X_4 < 0.5) & \text{otherwise,} \end{cases} \tag{8}$$

where $\epsilon \sim \mathcal{N}(0, 0.1)$ adds stochastic variation. The censoring distribution thus depends directly on the unobserved variable U, creating dependence between censoring and survival that cannot be explained away by the observed X alone.

Summary statistics Similar to the other synthetic datasets, we include up to 50,000 samples with treatment assigned according to an observational study mechanism. The treatment rate is 53.9%, the censoring rate is 39.7% (driven by U), and the population-level ATE is 0.7737 (computed from the 50,000 samples by averaging the CATEs). This setup mirrors real-world contexts such as clinical trials with dropout patterns influenced by latent health status.

Experimental results We evaluated representative estimators from all three method families. Figure 22 reports CATE RMSE (mean $\pm$ standard error) across 10 random splits.



(a) RCT: ✓(50%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓

(b) RCT: ✓(5%), Ignorability: ✓, Positivity: ✓, Ign-Censoring: ✓

Figure 22: CATE RMSE and ATE bias under informative censoring induced by unobserved confounding.

The results indicate that causal survival forest and survival meta-learners with matching tend to perform best under this setting, consistent with findings from the main synthetic datasets.

Extensibility to other settings. Here we illustrate one case: `OBS-UConf` combined with Scenario `C`. However, the same mechanism can be straightforwardly extended to other causal configurations (e.g., randomized trials with imbalance) and survival scenarios (e.g., AFT or Cox models). We leave systematic exploration of these additional combinations for future work, but their ease of inclusion highlights the flexibility of SurvHTE-Bench to accommodate alternative censoring mechanisms.