Causal Survival Forests with Negative Controls

Zijun Gao* **Kyoungeui Hong†** **Leyi Ma‡**

Qianli Wu§ **Zachary Izzo¶** **Ruishan Liu‖**

Abstract

We study heterogeneous treatment-effect (HTE) estimation in observational survival studies commonly associated with both censored outcomes and unmeasured confounding. We integrate causal survival forests (CSF) with negative controls (NC) from proximal causal inference and introduce Negative Control Causal Survival Forests (NC-CSF), a flexible nonparametric HTE learner for survival analysis. Our approach uses a loss that incorporates proxy variables and Neyman orthogonalization to train the random forest, thereby mitigating bias from unobserved confounding and gaining robustness to nuisance estimation. Through extensive simulations spanning varying levels of confounding, proxy relevance, and censoring mechanisms, we demonstrate that NC-CSF substantially reduces bias and estimation error (root mean squared error) relative to existing baselines. We further demonstrate the practical utility of our method on clinical datasets, where it confirms several existing findings and also reveals new interpretable patterns of treatment-effect heterogeneity. To facilitate practical use, we provide an end-to-end Python implementation of NC-CSF that carefully handles implementation details such as nuisance estimation and clipping.

1 Introduction

Time-to-event outcomes are common in treatment-effect analysis, especially in clinical trials and biomedical studies Kalbfleisch and Prentice [2002], Imbens and Rubin [2015]. These survival outcomes are frequently right-censored, for example because follow-up ends before the event is observed or participants leave the study early. In recent years, treatment-effect heterogeneity for survival outcomes has received increasing attention [Henderson et al., 2020, Tabib and Larocque, 2020, Hu et al., 2021, Cui et al., 2023, Hu et al., 2022], where the causal effect of treatment may vary across subgroups defined by baseline covariates. To model such heterogeneity flexibly in survival settings, machine-learning-assisted estimators have been developed. Among them, causal survival forests [Cui et al., 2023, Wager and Athey, 2018, Athey et al., 2019] provide a promising framework, as they can estimate heterogeneous treatment effects in a flexible and nonparametric manner.

Many existing methods for heterogeneous treatment-effect analysis with survival outcomes rely on an ignorability assumption, that is, they assume there is no unobserved confounding. However, this assumption may be unrealistic in many observational survival datasets, where unobserved confounding can induce bias in treatment-effect estimates when it is not properly adjusted for [Imbens and Rubin, 2015, Huang et al., 2020]. To address unobserved confounding, proximal causal inference [Miao et al., 2018, Tchetgen Tchetgen et al., 2024, Cui et al., 2024, Zhang et al., 2023,

*Department of Data Sciences and Operations, USC Marshall School of Business, University of Southern California. zijungao@marshall.usc.edu

†Department of Computer Science, University of Southern California. hongkyou@usc.edu

‡Department of Computer Science, University of Southern California. rma99604@usc.edu

§Department of Computer Science, University of Southern California. ewu96023@usc.edu

¶NEC Labs America. zle.izzo@gmail.com

‖Department of Computer Science, University of Southern California. ruishanl@usc.edu

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Tchetgen Tchetgen, 2023, Liu et al., 2025] offers a promising framework by leveraging negative-control treatments and negative-control outcomes, that is, auxiliary variables that are informative about latent confounders while satisfying exclusion-type restrictions with respect to the primary treatment and outcome. Negative-control variable assumptions may be more suitable in some biomedical applications than the classical no-unmeasured-confounding assumption, for example, pre-treatment measurements may share latent risk factors while remaining unaffected by treatment. This literature has also recently been extended to survival settings with censored data, primarily for average treatment effects and under parametric modeling assumptions [Li et al., 2025].

Motivated by the practical need for flexible nonparametric modeling of heterogeneous treatment effects, as well as by the appeal of addressing unobserved confounding via negative controls, we connect proximal causal inference with causal survival forests and develop *negative-control causal survival forests* (NC-CSF). In particular, our contributions are as follows.

1. *Method*. We design a loss for random forests to learn heterogeneous treatment effects with censored survival outcomes that explicitly incorporates negative controls to adjust for unobserved confounding and employs Neyman orthogonalization to achieve robustness to nuisance-function estimation.

2. *Theory*. We provide theoretical justification for the proposed approach by showing how the negative-control-based loss mitigates bias from unobserved confounding and by establishing a local doubly robust property.

3. *Applications*. We evaluate the proposed method on both simulated and real datasets. In particular, on an AIDS clinical trial dataset, our method suggests that the fitted RMST treatment effect is negative for younger individuals and positive for older individuals, consistent with existing treatment recommendations, whereas existing methods yield uniformly positive estimates across ages.

4. *Packages*. We provide a ready-to-use Python package for the proposed method, where we have carefully addressed several practical issues required for stable finite-sample performance, such as nuisance estimation and variable clipping.

Organization. Section 2 introduces the problem setup and background. Section 3 presents our proposed NC-CSF estimator and provides its theoretical justification. Section 4 studies NC-CSF's performance in simulations. Section 5 illustrates the method on real clinical data. We conclude with a discussion in Section 6. Proofs and additional numerical results are deferred to the appendix.

2 Background

We first formulate the problem of heterogeneous treatment-effect estimation in the presence of right censoring and unmeasured confounding. We then review the two key ingredients underlying our proposal: causal survival forests and negative controls.

2.1 Problem Formulation

We follow the potential-outcome framework [Rubin, 1974, Imbens and Rubin, 2015] and consider an observational survival study with n units. Each unit is associated with a binary treatment $A_i \in \{0, 1\}$, observed baseline covariates X_i, and unobserved confounders U_i. Let $T_i(a)$ and $C_i(a)$, for $a \in \{0, 1\}$, denote the potential event time and potential censoring time, respectively. The realized event and censoring time are given by $T_i = T_i(A_i)$ and $C_i = C_i(A_i)$, and the observed survival time and censoring indicator are $Y_i = \min(T_i, C_i)$ and $\Delta_i = I\{T_i \leq C_i\}$. Let $h > 0$ be a fixed horizon, and we define the horizon-adjusted follow-up time and indicator by $Y_i^{(h)} = \min(Y_i, h)$ and $\Delta_i^{(h)} = \Delta_i \, \mathbf{1}\{Y_i \leq h\} + \mathbf{1}\{Y_i > h\}$.

We consider two sets of causal estimands derived from the potential event times. The first is the restricted mean survival time (RMST), defined by $\min\{T_i(a), h\}$, and the second is the survival probability at horizon h, defined by $I\{T_i(a) > h\}$. Let $G_i(a; h)$ denote either of these two functionals,

and use $G_i(h)$ to denote $G_i(A_i; h)$. The associated heterogeneous treatment-effect estimands are

$$\begin{aligned}
\tau &:= \mathbb{E}\left[G_i(1;h) - G_i(0;h)\right], \\
\tau(x) &:= \mathbb{E}\left[G_i(1;h) - G_i(0;h) \mid X_i = x\right], \\
\tau^w(x) &:= \mathbb{E}\left[w(X_i, U_i)(G_i(1;h) - G_i(0;h)) \mid X_i = x\right], \\
\tau(x,u) &:= \mathbb{E}\left[G_i(1;h) - G_i(0;h) \mid X_i = x, U_i = u\right],
\end{aligned}$$

where $w(X_i, U_i) \geq 0$ denotes a weight function satisfying $\mathbb{E}[w(X_i, U_i) \mid X_i] = 1$; the specific NC-CSF weight is defined in Section 3.4. Here, τ is the average treatment effect, $\tau(x)$ is the conventional treatment effect conditional on the observed covariates, and $\tau(x, u)$ is the treatment effect conditional on the full confounding state. The NC-CSF target $\tau^w(x)$ is instead a weighted average of $\tau(x, U_i)$ given $X_i = x$, which can differ from $\tau(x)$ when the unobserved confounder also modifies the treatment effect (NC-CSF gives greater weight to latent strata with more bridge residualized treatment variation). It reduces to the conventional CATE under the conditions in Remark 3.4.

We state the assumptions needed for our proposal. We assume the stable unit treatment value assumption (SUTVA), which says there is no interference between units and no hidden versions of treatment. We also assume overlap given both X_i and U_i, namely

$$0 < P(A_i = 1 \mid X_i, U_i) < 1, \quad \text{a.s.}$$

Importantly, we do *not* assume ignorability given the observed covariates X_i, and we allow treatment assignment to depend on the latent confounder U_i.

2.2 Causal Survival Forest Preliminaries

To review the causal survival forests (CSF) in [Cui et al., 2023] under no unmeasured confounding, we first introduce a few new notations. Let $e(X_i) = \mathbb{P}(A_i = 1 \mid X_i)$ denote the propensity score, and let $m(X_i) = E[G_i(h) \mid X_i]$ denote the marginal mean function. Let $S_C(t \mid X_i, A_i) = \mathbb{P}(C_i > t \mid X_i, A_i)$ and $\Lambda_C(t \mid X_i, A_i) = \int_0^t \lambda_C(u \mid X_i, A_i)\, du$ denote the conditional censoring survival function and cumulative hazard, respectively, where $\lambda_C(t \mid X_i, A_i)$ is the conditional censoring hazard function defined by $\lambda_C(t \mid X_i, A_i) = \lim_{\Delta t \to 0} \mathbb{P}(t \leq C_i < t + \Delta t \mid C_i \geq t,\ X_i, A_i)/\Delta t$.

In CSF, the following loss is used for node splitting and leaf fitting within each tree,

$$\ell^{\mathrm{CSF}} = \sum_{i=1}^{n} \Big(\Gamma_i^{\mathrm{CSF}}(h) - m(X_i) - \tau(X_i)\big(A_i - e(X_i)\big) \Big)^2,$$

$$\Gamma_i^{\mathrm{CSF}}(h) = \frac{\Delta_i^{(h)} G_i(h) + \big(1 - \Delta_i^{(h)}\big)\rho_{A_i}\big(Y_i^{(h)}, X_i; h\big)}{S_C\big(Y_i^{(h)} \mid X_i, A_i\big)} - \int_0^{Y_i^{(h)}} \frac{\rho_{A_i}(t, X_i; h)}{S_C\big(t \mid X_i, A_i\big)}\, d\Lambda_C\big(t \mid X_i, A_i\big),$$

where $\Gamma_i^{\mathrm{CSF}}(h)$ is the censoring-corrected outcome and $\rho_a(t, X_i; h) = E[G_i(a; h) \mid T_i(a) > t, X_i]$ denotes the expected outcome conditional on the unit surviving to time t. Results for uncensored data can be recovered as a special case of the censored setting by assuming $C_i(0) = C_i(1) = \infty$, a.s.

A key property of this construction is its robustness to nuisance-function estimation (here the nuisance functions include ρ_a, S_C, Λ_C, m, and e). In particular, the associated score function is Neyman orthogonal, meaning that its pathwise derivative with respect to the nuisance functions is zero at the truth, and only second-order nuisance estimation errors remain locally. In Section 3, we extend the loss of CSF to incorporate negative controls to address unobserved confounders.

2.3 Negative Controls and Bridge Functions Preliminaries

To review negative controls and bridge functions for latent confounders, we first introduce some additional notation. The standard proximal causal inference framework [Miao et al., 2018, Tchetgen Tchetgen et al., 2024, Cui et al., 2024] assumes access to a negative-control outcome proxy W_i and a negative-control treatment proxy Z_i. Intuitively, W_i is associated with the latent confounder but is not affected by the treatment, while Z_i is associated with treatment assignment but does not directly affect the outcome. The conditional independence structure is shown in Figure 1; the formalized assumption adapted to our setting is in (1).

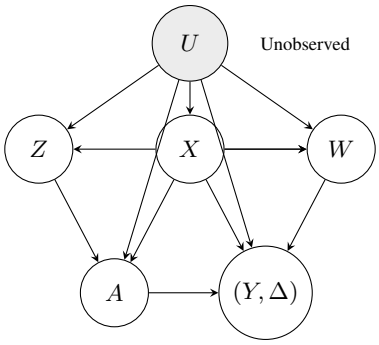

Figure 1: Directed acyclic graph illustrating the assumptions of the negative-control treatment Z and negative-control outcome W.

To remove bias induced by unobserved confounding, a common way of using negative controls is through bridge functions. In the standard setting of estimating the average treatment effect, one may consider bridge functions satisfying $\mathbb{E}\left[f_a(W_i, X_i) \mid X_i, U_i\right] = \mathbb{E}\left[Y_i(a) \mid X_i, U_i\right]$. When such functions exist, a key benefit of these bridge relations is that, although U_i is not observed, we have $\mathbb{E}[f_a(W_i, X_i)] = \mathbb{E}[\mathbb{E}[f_a(W_i, X_i) \mid X_i, U_i]] = \mathbb{E}[\mathbb{E}[Y_i(a) \mid X_i, U_i]] = \mathbb{E}[Y_i(a)]$, so that unbiased estimation can still be achieved using observable variables alone. In existing works, estimators are typically formed by averaging bridge-function-based functionals over units, whereas we instead use the bridge functions to construct a loss to train random forests that allows more flexible modeling of $\tau(x)$ and facilitates the use of well-developed machine learning tools.

3 Method

3.1 Negative Control and Bridge Functions for Survival Analysis

In this paper, we assume access to both a negative-control outcome and a negative-control treatment, satisfying the conditional independence structure

$$(A_i, Z_i) \perp (T_i(a), C_i(a), W_i) \mid U_i, X_i. \tag{1}$$

This means that, conditional on the full confounding state (X_i, U_i), the treatment A_i and negative-control treatment Z_i are independent of the potential event time, the potential censoring time, and the negative-control outcome W_i. Although the assumption (1) is not testable from the observed data alone, it may be plausible in biomedical studies by temporal ordering; for example, pre-treatment measurements can serve as negative-control outcomes as they may share latent risk factors [Tchetgen Tchetgen et al., 2024, Li et al., 2025].

Given W_i and Z_i, we consider bridge functions $f_a(w, x)$, $a \in \{0, 1\}$, and $q(z, x)$ satisfying

$$\begin{aligned} \mathbb{E}\left[f_a(W_i, X_i) \mid X_i, U_i\right] &= \mathbb{E}\left[G_i(a; h) \mid X_i, U_i\right], \\ \mathbb{E}\left[q(Z_i, X_i) \mid X_i, U_i\right] &= \mathbb{P}(A_i = 1 \mid X_i, U_i). \end{aligned} \tag{2}$$

These bridge relations imply that, although U_i is unobserved, the latent conditional outcome and treatment mechanisms can be represented through observable functions of (W_i, X_i) and (Z_i, X_i), respectively. We note that our specification of q differs mildly from that in Miao et al. [2018], where the conditional expectation of the bridge function is specified to match $1/\mathbb{P}(A_i = a \mid X_i, U_i)$.

Finally, for censoring, we assume the standard conditional independent censoring assumption [Kalbfleisch and Prentice, 2002, Cui et al., 2023] given the observed variables including the negative controls:

$$C_i(a) \perp T_i(a) \mid X_i, W_i, Z_i, A_i = a, \quad a \in \{0, 1\}. \tag{3}$$

We next introduce notation for censored survival outcomes in the presence of negative controls. Write $q_i = q(Z_i, X_i)$, $f_{a,i} = f_a(W_i, X_i)$, and define the bridge-adjusted mean

$$m_i = q_i f_{1,i} + (1 - q_i) f_{0,i}.$$

Define $S_C(t \mid X_i, W_i, Z_i, A_i)$, $\rho_a(t \mid X_i, W_i, Z_i)$, $\Lambda_C(t \mid X_i, W_i, Z_i, A_i)$, and $\lambda_C(t \mid X_i, W_i, Z_i, A_i)$ analogously to Section 2.2.

3.2 Oracle NC-CSF

We construct the loss for NC-CSF using both bridge functions q and f_a, and then fit a random forest by minimizing it. We refer to this as the oracle NC-CSF, since we assume that all true nuisance functions, including the bridge functions, are provided. Below we focus on the loss construction.

For the RMST target, analogous to the residualization in [Robinson, 1988, Nie and Wager, 2021], we introduce residualized treatment and outcome,

$$\begin{aligned} A_i^{\mathrm{res}} &= A_i - q_i, \\ Y_i^{\mathrm{res}} &= \frac{\Delta_i^{(h)}\big(Y_i^{(h)} - m_i\big) + \big(1-\Delta_i^{(h)}\big)\big(\rho_{A_i}(Y_i^{(h)}, X_i, W_i, Z_i; h) - m_i\big)}{S_C(Y_i^{(h)} \mid X_i, W_i, Z_i, A_i)} \\ &\quad - \int_0^{Y_i^{(h)}} \frac{\rho_{A_i}(t, X_i, W_i, Z_i; h) - m_i}{S_C(t \mid X_i, W_i, Z_i, A_i)}\, d\Lambda_C(t \mid X_i, W_i, Z_i, A_i). \end{aligned} \tag{4}$$

Here, A_i^{res} is the bridge-adjusted treatment residual satisfying $E[A_i^{\mathrm{res}} \mid X_i, U_i] = 0$, and Y_i^{res} denotes the censoring-corrected residualized outcome. In the absence of censoring, Y_i^{res} simplifies to $Y_i^{(h)} - m_i$, as illustrated in Example 3.1. The survival-probability analogue is given in Appendix (18). Finally, we define the NC-CSF loss based on the residuals again motivated by [Robinson, 1988, Nie and Wager, 2021],

$$\ell^{\text{NC-CSF}} = \sum_{i=1}^{n} \left(Y_i^{\mathrm{res}} - \tau(X_i)\, A_i^{\mathrm{res}}\right)^2. \tag{5}$$

Example 3.1 (Uncensored case) *If there is no censoring, then $C_i = \infty$ almost surely and $\Delta_i = 1$ for all i. Hence $\Delta_i^{(h)} = 1$, $S_C(t \mid X_i, W_i, Z_i, A_i) = 1$, $\Lambda_C(t \mid X_i, W_i, Z_i, A_i) = 0$, and $Y_i^{\mathrm{res}} = Y_i^{(h)} - m_i$.*

3.3 Empirical NC-CSF

In this section, we discuss how to estimate the nuisance quantities and plug them into the oracle loss for fitting, yielding the (empirical) NC-CSF. Because U_i is unobserved, the latent bridge equations in (2) cannot be evaluated directly from the observed data. We therefore use the observable conditional moment equations stated in Proposition 3.2 below, which do not involve U_i and are implied by Eq. (1) and Eq. (2).

Proposition 3.2 *Under* (1)*, suppose f and q satisfy* (2)*, then, for each a,*

$$\mathbb{E}[A_i - q(Z_i, X_i) \mid X_i, W_i] = 0, \quad \mathbb{E}[G_i(a; h) - f_a(W_i, X_i) \mid X_i, Z_i, A_i = a] = 0. \tag{6}$$

Conversely, if the following completeness conditions hold: $\mathbb{E}[r(X_i, U_i) \mid X_i, W_i] = 0$ a.s. implies $r(X_i, U_i) = 0$ a.s. and $\mathbb{E}[s(X_i, U_i) \mid X_i, Z_i, A_i = a] = 0$ a.s. implies $s(X_i, U_i) = 0$ a.s. for any square-integrable measurable function r, s, then (6) *implies* (2)*.*

When censoring is present, $G_i(a; h)$ may be unobserved. Under (1) and (3), it can be replaced by an inverse-probability-of-censoring-weighted pseudo-outcome satisfying the same conditional moment equation; details are given in the Appendix.

To learn survival-related nuisance functions, under (3), those nuisance functions are not required to satisfy conditions involving the unobserved confounder U_i, and thus can be estimated using standard survival-analysis procedures, such as Cox models. Finally, to avoid overfitting in the nuisance stage, we employ cross-fitting Chernozhukov et al. [2018]. Details are given in Appendix Algorithm 1.

3.4 Theory

Proposition 3.3 (Validity of the NC-CSF score) *Assume* (1) *and* (2)*. Then the score has conditional mean zero given $X_i = x$ at*

$$\tau^w(x) = \frac{\mathbb{E}\big[(A_i - q(Z_i, X_i))^2 \tau(X_i, U_i) \mid X_i = x\big]}{\mathbb{E}[(A_i - q(Z_i, X_i))^2 \mid X_i = x]},$$

where $\tau(x,u) = \mathbb{E}[G(1;h) - G(0;h) \mid X = x, U = u]$. In particular, when the latent variable U does not modify the treatment effect beyond X, that is, $\tau(x,u) = \tau(x)$, we have $\tau^w(x) = \tau(x)$.

Remark 3.4 (Interpretation of the weighted target) *Define*

$$\omega(x,u) = \mathbb{E}\big[\{A_i - q(Z_i, X_i)\}^2 \mid X_i = x, U_i = u\big].$$

Then $\tau^w(x)$ is a residual-variation-weighted average of the latent effects and remains heterogeneous over the observed covariates even when U also modifies treatment effects. In particular,

$$\tau^w(x) - \tau(x) = \frac{\mathrm{Cov}\{\omega(x,U_i), \tau(x,U_i) \mid X_i = x\}}{\mathbb{E}\{\omega(x,U_i) \mid X_i = x\}}. \tag{7}$$

Thus $\tau^w(x) = \tau(x)$ when U does not modify the effect beyond X, when $\omega(x,U)$ and $\tau(x,U)$ are conditionally uncorrelated given $X = x$, or when $\omega(x,U)$ does not vary with U. The weighting is analogous to the overlap weighting induced by an R-learner with omitted effect modifiers [Nie and Wager, 2021]. Among positive latent weighting functions, $w'(x,u) \propto \omega(x,u)$ minimizes the oracle conditional-noise variance bound, so $\tau^w(x)$ is variance-optimal within this weighted class.

Proposition 3.5 (Neyman orthogonality of the NC-CSF score) *Let $\eta = (q, m, S_C, \rho_a, \Lambda_C)$ denote the set of nuisance functions. At the true nuisance value η_0, the pathwise derivative of the NC-CSF score with respect to each nuisance component has conditional expectation zero given X.*

Proposition 3.5 implies the loss is only second-order sensitive to the nuisance estimation error.

Proposition 3.6 *Let $\mathcal{S}$ be a fixed subgroup, or a subgroup learned on data independent of the sample used for within-subgroup estimation, and let $n_{\mathcal{S}} \to \infty$ denote its estimation-sample size. Define*

$$\tau^w_{\mathcal{S}} = \frac{\mathbb{E}[\{A_i - q(Z_i, X_i)\}^2 \tau(X_i, U_i) \mid X_i \in \mathcal{S}]}{\mathbb{E}[\{A_i - q(Z_i, X_i)\}^2 \mid X_i \in \mathcal{S}]}.$$

Assume the following:

1. *Boundedness: the outcomes, covariates, and nuisance functions q, m, S_C, ρ_a, and Λ_C are uniformly bounded, and S_C is bounded away from zero on the time range of interest;*
2. *Residual overlap: $\mathbb{E}[\{A_i - q(Z_i, X_i)\}^2 \mid X_i \in \mathcal{S}]$ is bounded away from zero;*
3. *Nuisance estimation: the cross-fitted estimators $\widehat{q}$, $\widehat{m}$, $\widehat{S}_C$, $\widehat{\rho}_a$, and $\widehat{\Lambda}_C$ are consistent and each has L_2 error $o_p(n_{\mathcal{S}}^{-1/4})$ on the subgroup.*

Then the cross-fitted NC-CSF estimator is asymptotically equivalent to its oracle-nuisance counterpart and

$$\sqrt{n_{\mathcal{S}}}\,(\widehat{\tau}_{\mathcal{S}} - \tau^w_{\mathcal{S}}) \xrightarrow{d} N(0, \sigma^2_{\mathcal{S}}),$$

where

$$\sigma^2_{\mathcal{S}} = \frac{\mathrm{Var}(A_i^{\mathrm{res}}(Y_i^{\mathrm{res}} - \tau^w_{\mathcal{S}} A_i^{\mathrm{res}}) \mid X_i \in \mathcal{S})}{\mathbb{E}[(A_i^{\mathrm{res}})^2 \mid X_i \in \mathcal{S}]^2}.$$

In Proposition 3.6, if $\tau(x,u)$ is constant on $\mathcal{S}$, then $\tau^w_{\mathcal{S}}$ equals that shared subgroup effect. However, no homogeneity assumption is needed for the asymptotic normality around the weighted within-subgroup target.

Remark 3.7 *Proposition 3.6 considers the estimation and inference within a fixed or honest subgroup, following the sample-splitting principle used by honest causal trees and forests [Athey and Imbens, 2016, Wager and Athey, 2018]. It does not prove that a recursively learned forest recovers a unique oracle partition. Exact partition recovery generally requires additional separation and structural conditions, such as minimum jumps between adjacent regions [Madrid Padilla et al., 2021].*

4 Simulation

4.1 Methods for Comparison

We distinguish a matched ablation from external competitors. This distinction is important because the methods differ in software, nuisance estimators, and forest engines.

Matched ablations and oracle references. In the uncensored setting, *Uncensored Baseline* applies `CausalForestDML` [Battocchi et al., 2019] directly to $[X, W, Z]$, while *Uncensored NC-CSF* adds the bridge-adjusted residualization. In the censored setting, *Censored Baseline* is NC-CSF with the bridge layer removed: it retains the same preprocessing, censoring-robust residual construction, event and censoring nuisances, final features, train/test splits, and forest family. It is therefore the comparison that isolates the contribution of the negative-control bridge. The oracle variants have access to U and serve only as synthetic upper-bound references.

External survival and HTE learners. We additionally evaluate the R implementation of causal survival forests in `grf` [Cui et al., 2023], random survival forests used as T- and S-learners [Ishwaran et al., 2008], and a discrete-time survival BART S-learner [Sparapani et al., 2016]. Every external learner receives $[X, W, Z]$ as ordinary pre-treatment covariates; NC-CSF differs by assigning W and Z their negative-control roles. Thus no external method is disadvantaged by lack of feature access. The no-negative-control arm in our simulation is the matched Censored Baseline, not the `grf` implementation. We also report population-level AIPW and TMLE comparisons in Appendix D.5, while noting that they target a marginal ATE rather than an individual CATE.

Implementation and metrics. NC-CSF uses logistic regression for the treatment bridge, ExtraTrees [Geurts et al., 2006] for outcome-bridge regressions, Cox models [Cox, 1972] for censored nuisances, and EconML's `CausalForestDML` for the final forest. All benchmark methods use identical training/test splits, horizons, and target values. We summarize RMSE and MAE over the common $12 \times 22 = 264$ censored configurations; pseudocode for the matched estimators appears in Appendix 3–5.

Table 1: Average error over the $12 \times 22 = 264$ censored benchmark configurations. External learners receive $[X, W, Z]$ as ordinary covariates; lower is better.

Model	**RMST**		**Survival Probability**	
	RMSE	**MAE**	**RMSE**	**MAE**
NC-CSF Oracle	0.0516	0.0380	0.0361	0.0275
NC-CSF	0.0994	0.0768	0.0613	0.0491
`grf` R-CSF	0.1211	0.0971	0.0686	0.0581
Censored Baseline	0.1662	0.1247	0.0788	0.0642
RSF T-learner	0.1773	0.1416	0.0796	0.0721
RSF S-learner	0.1637	0.1274	0.0889	0.0638
Survival BART	0.1762	0.1318	0.0699	0.0657

NC-CSF has the lowest error among all feasible methods. Relative to the matched Censored Baseline, it reduces RMST RMSE by 40.2% and survival-probability RMSE by 22.2%; relative to the strongest external tree-ensemble comparator, `grf` R-CSF, the reductions are 17.9% and 10.6%. The component ablation in Appendix 10 shows that the bridge adjustment is the largest contributor: adding the bridge while holding the forest and censoring pipeline fixed reduces RMST RMSE from 0.1662 to 0.1081 and survival-probability RMSE from 0.0788 to 0.0675; nuisance-summary augmentation yields the remaining improvement to 0.0994 and 0.0613.

4.2 Data-Generating Mechanism

We generate n i.i.d. samples with covariates X_i and latent confounder U_i, where $X_i \sim \mathcal{N}(0, I_p)$ and $U_i \sim \mathcal{N}(0, 1)$. The negative-control proxies Z_i and W_i are constructed as noisy linear functions of (X_i, U_i), so that both carry information about the latent confounder. Treatment assignment follows a Bernoulli model with logistic mean $\mathbb{P}(A_i = 1 \mid X_i, U_i) = \sigma(b_0 + \gamma_u U_i + \alpha^\top X_i)$. Potential event times are generated from a Weibull proportional hazards model, with $T_i(a) \mid X_i, U_i \sim$

$\text{Weibull}(k, \lambda \exp(-\eta_i(a)/k))$ and linear predictor $\eta_i(a) = \beta_t^\top X_i + \beta_u U_i + \tau a$. Because the proxy construction is linear-Gaussian, the conditional law of U_i given (X_i, Z_i, W_i) is available in closed form, which yields an explicit expression for the observed-data conditional treatment effect $\mathbb{E}[T_i(1) - T_i(0) \mid X_i, Z_i, W_i]$ after integrating out U_i. Appendix D.3 further replaces the proxy mechanism with nonlinear, heavy-tailed, noninjective, and binary-coarsened alternatives.

We evaluate performance across regimes varying proxy quality, confounding strength, treatment effect size, and covariate dimensionality, under both linear and nonlinear settings, which results in 12×22 configurations (Table 16 in Appendix).

4.3 Results with Uncensored Data

We evaluate estimation accuracy using root mean squared error, $\text{RMSE} = \sqrt{n^{-1} \sum_{i=1}^{n} (\widehat{\tau}(X_i) - \tau_i)^2}$, and mean absolute error, $\text{MAE} = n^{-1} \sum_{i=1}^{n} |\widehat{\tau}(X_i) - \tau_i|$. NC-CSF closely matches the oracle across all n and covariate dimensions, while the baseline exhibits higher error, particularly in small-sample and high-dimensional regimes. Table 17 in Appendix further shows that NC-CSF improves over the Uncensored Baseline in both RMSE and MAE and remains closer to the oracle; additional results appear in Appendix Figure 5, Figure 6, and Table 4.

Across 30 paired Monte Carlo replications of the representative sample-size and dimension sweeps, all nine NC-CSF-minus-baseline 95% intervals lie below zero, and NC-CSF wins 28–30 of 30 replications in every setting. Representative sample-size and dimensionality results, together with Monte Carlo uncertainty, are reported in Appendix D.4 and Figures 3–4.

4.4 Results with Censored Data

Following the uncensored experiments, we evaluate performance under right censoring. We fix the censoring rate at 35% across all configurations and calibrate the Weibull scale parameter accordingly to achieve the desired censoring level. Table 1 summarizes results for both RMST (Restricted Mean Survival Time) and survival probability. These results indicate that the benefits of proxy-based adjustment become more pronounced in the presence of censoring, where standard approaches are more susceptible to bias and instability. Detailed RMSE and MAE results across all censored configurations are reported in Appendix Figures 7–9.

To assess heterogeneity, we compute subgroup bias on an independent test set of size 20,000 for each setting–case pair. Subgroups are predefined using quantile bins of selected covariates, yielding 20 groups per configuration: 3 bins for X_0, three for X_1, 5 for the true CATE, and nine for the $X_0 \times X_1$ interaction bins. After training on an independent dataset, we predict CATEs on the test set and compute, for each subgroup g, the bias $\text{Bias}(g) = |g|^{-1} \sum_{i \in g} \{\widehat{\tau}(X_i) - \tau_i\}$, excluding groups below a minimum size. These summaries assess local CATE accuracy beyond aggregate performance. Subgroup-bias results across all prespecified partitions are reported in Appendix Figures 10–14.

5 Real Data

5.1 ACTG175 Trial

We apply NC-CSF to the ACTG175 ddI versus ZDV+ddI subset [Hammer et al., 1996, Cui et al., 2023], containing 1,083 subjects and targeting RMST at $h = 1{,}000$ days. With $W = (\texttt{cd420}, \texttt{cd820})$, $Z = (\texttt{race}, \texttt{preanti})$, and 12 baseline variables in X, NC-CSF estimates a smaller treatment benefit for younger reference profiles (-5.69 days at age 20) and a positive benefit for older profiles (peaking at 6.37 days around age 62), whereas the re-run R-CSF curve is positive throughout. Because the pointwise intervals are wide, we interpret this pattern as exploratory; three clinically motivated proxy routings and the complete age-profile comparison are reported in Appendix D.7.

5.2 Right Heart Catheterization

We also analyze the SUPPORT right-heart-catheterization data [Connors et al., 1996], following the proxy specification of Kallus et al. [2021]. NC-CSF estimates a 0.0433 reduction in 30-day survival under RHC, lying between the flexible minimax-bridge estimates and the more negative

linear/no-hidden-confounding estimates reported in prior work. We treat this as a consistency check; full estimates are reported in Appendix Table 6.

5.3 MIMIC-IV pharmacologic positive and inert controls

We analyze MIMIC-IV [Johnson et al., 2023] to study whether combining tacrolimus with a CYP3A inhibitor increases acute kidney injury (AKI) within seven days. This interaction is well documented [U.S. Food and Drug Administration, 2023, Birdwell et al., 2015, Naesens et al., 2009]. Survival means remaining AKI-free, so negative estimates indicate harm. Pretreatment laboratory values serve as outcome-side proxies for latent severity, and administrative variables serve as treatment-side proxies for prescribing patterns. The analysis contains $2{,}257$ records in the positive-control arm and $1{,}398$ in an inert-control arm in which the added drug has no CYP3A activity.

Table 2: MIMIC-IV average-effect estimates for 7-day AKI-free survival and AKI-free RMST. Negative values indicate harm.

Method	**Survival**	**95% CI**	**E-value**	**RMST**	**95% CI**
Positive-control arm: tacrolimus with versus without a CYP3A inhibitor					
RSF-S	-0.0033	$[-0.0128, +0.0018]$	1.15	-0.0205	$[-0.0773, +0.0017]$
BART	-0.0462	$[-0.1529, +0.0311]$	–	-0.2365	$[-0.7806, +0.1800]$
RSF-T	-0.0758	$[-0.1245, -0.0401]$	2.17	-0.4279	$[-0.6396, -0.2223]$
CSF	-0.1176	$[-0.1939, -0.0758]$	2.69	-0.6232	$[-0.9129, -0.4440]$
NC-CSF	$\mathbf{-0.2410}$	$[\mathbf{-0.2857}, \mathbf{-0.1816}]$	**4.40**	$\mathbf{-1.6248}$	$[\mathbf{-2.1263}, \mathbf{-1.2023}]$
Inert-control arm: added drug has no CYP3A activity					
RSF-S	$+0.0053$	$[+0.0001, +0.0189]$	–	$+0.0224$	$[-0.0051, +0.0986]$
BART	$+0.0092$	$[-0.0694, +0.0965]$	–	$+0.0473$	$[-0.3791, +0.5054]$
RSF-T	$+0.0147$	$[-0.0187, +0.0492]$	–	-0.0221	$[-0.1486, +0.0990]$
CSF	$+0.0013$	$[-0.0414, +0.0466]$	–	$+0.0056$	$[-0.1574, +0.2161]$
NC-CSF	$\mathbf{-0.0323}$	$[\mathbf{-0.1403}, \mathbf{+0.0171}]$	–	$\mathbf{+0.0265}$	$[\mathbf{-0.2453}, \mathbf{+0.4034}]$

In the positive-control arm, NC-CSF estimates a 0.2410 reduction in seven-day AKI-free survival and 1.6248 fewer AKI-free days, more than twice the corresponding CSF magnitudes. In the inert-control arm, both NC-CSF intervals include zero. Thus NC-CSF detects the established harmful interaction while remaining silent under the inert substitution. Among four prespecified comorbidity axes, only prior liver disease shows clear amplification of harm; Appendix D.8 reports all available subgroup summaries.

6 Discussion

In this paper, we provide a practical approach to survival analysis with flexible heterogeneous treatment-effect modeling in the presence of unobserved confounding. Empirically, the proposed method performs favorably in extensive experiments and yields interesting findings in real data applications. We also provide an off-the-shelf Python package to improve practical usability.

We discuss limitations. First, $\tau^w(x)$ is the general estimand and need not equal the conventional CATE when latent treatment-effect heterogeneity covaries with bridge-residualized treatment variation. Second, negative-control validity is application-specific and cannot be verified from the observed law alone. Third, completeness does not resolve finite-sample ill-posedness: orthogonality limits downstream sensitivity to sufficiently accurate nuisance estimates but does not create accurate bridge estimates when the moment operator is poorly conditioned. Regularized bridge learning and diagnostics for proxy strength therefore remain practically relevant.

Several directions remain for future work. From the causal perspective, since the resulting target is weighted in general, it would be useful to further study the interpretation and behavior of different estimands. On the theoretical side, one direction is to analyze the recursive splitting behavior of both the oracle forest and its feasible counterpart with estimated nuisance functions.

References

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016. doi: 10.1073/pnas.1510489113.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019. doi: 10.1214/18-AOS1709.

Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. Econml: A python package for ml-based heterogeneous treatment effects estimation. `https://github.com/py-why/EconML`, 2019.

Kelly A. Birdwell et al. Clinical pharmacogenetics implementation consortium (CPIC) guidelines for CYP3A5 genotype and tacrolimus dosing. *Clinical Pharmacology & Therapeutics*, 98(1):19–24, 2015. doi: 10.1002/cpt.113.

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. doi: 10.1111/ectj.12097.

Alfred F. Jr. Connors, Theodore Speroff, Neal V. Dawson, Charles Thomas, Frank E. Jr. Harrell, Douglas Wagner, Norman Desbiens, Lee Goldman, Albert W. Wu, Robert M. Califf, William J. Jr. Fulkerson, Humberto Vidaillet, Steven Broste, Paul Bellamy, Joanne Lynn, and William A. Knaus. The effectiveness of right heart catheterization in the initial care of critically ill patients. *JAMA*, 276(11):889–897, 1996. doi: 10.1001/jama.1996.03540110043030.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–220, 1972.

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. doi: 10.1093/jrsssb/qkac001.

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric J. Tchetgen Tchetgen. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359, 2024. doi: 10.1080/01621459.2023.2191817.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006. doi: 10.1007/s10994-006-6226-1.

Scott M. Hammer, David A. Katzenstein, Michael D. Hughes, Holly Gundacker, Richard 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. doi: 10.1056/NEJM199610103351501.

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.

Liangyuan Hu, Jiayi Ji, and Fan Li. Estimating heterogeneous survival treatment effect in observational data using machine learning. *Statistics in Medicine*, 40(21):4691–4713, 2021.

Liangyuan Hu, Jiayi Ji, Hanzhong Liu, and Ronald Ennis. A flexible approach for assessing heterogeneity of causal treatment effects on patient survival using large datasets with clustered observations. *International Journal of Environmental Research and Public Health*, 19(22):14903, 2022.

Rong Huang, Ronghui Xu, and Parambir S. Dulai. Sensitivity analysis of treatment effect to unmeasured confounding in observational studies with survival and competing risks outcomes. *Statistics in Medicine*, 39(24):3397–3411, 2020. doi: 10.1002/sim.8672.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

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. doi: 10.1214/08-AOAS169.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023. doi: 10.1038/s41597-022-01899-x.

John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, Hoboken, NJ, 2 edition, 2002. doi: 10.1002/9781118032985.

Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*, 2021.

Kendrick Qijun Li, George C. Linderman, Xu Shi, and Eric J. Tchetgen Tchetgen. Regression-based proximal causal inference for right-censored time-to-event data. *Epidemiology*, 36(5):694–704, 2025. doi: 10.1097/EDE.0000000000001884. URL `https://journals.lww.com/epidem/fulltext/9900/regression_based_proximal_causal_inference_for.388.aspx`.

Jiewen Liu, Chan Park, Kendrick Li, and Eric J. Tchetgen Tchetgen. Regression-based proximal causal inference. *American Journal of Epidemiology*, 194(7):2030–2036, 2025. doi: 10.1093/aje/kwae370.

Oscar Hernan Madrid Padilla, Yi Yu, and Alessandro Rinaldo. Lattice partition recovery with dyadic CART. In *Advances in Neural Information Processing Systems*, volume 34, pages 26143–26155, 2021.

Wang Miao, Zhi Geng, and Eric J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018. doi: 10.1093/biomet/asy038.

Maarten Naesens, Dirk R. J. Kuypers, and Minnie Sarwal. Calcineurin inhibitor nephrotoxicity. *Clinical Journal of the American Society of Nephrology*, 4(2):481–508, 2009. doi: 10.2215/CJN.04800908.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021. doi: 10.1093/biomet/asaa076.

Peter M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988. doi: 10.2307/1912705.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.

Rodney A. Sparapani, Brent R. Logan, Robert E. McCulloch, and Purushottam W. Laud. Nonparametric survival analysis using bayesian additive regression trees (BART). *Statistics in Medicine*, 35(16):2741–2753, 2016. doi: 10.1002/sim.6893.

Sami Tabib and Denis Larocque. Non-parametric individual treatment effect estimation for survival data with random forests. *Bioinformatics*, 36(2):629–636, 2020.

Eric J. Tchetgen Tchetgen. Proximal causal inference for complex longitudinal studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 85(3):684–704, 2023. doi: 10.1093/jrsssb/qkad020.

Eric J. Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal inference. *Statistical Science*, 39(3):375–390, 2024. doi: 10.1214/23-STS911.

U.S. Food and Drug Administration. PROGRAF (tacrolimus) prescribing information. DailyMed, August 2023. URL `https://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=7f667de1-9dfa-4bd6-8ba0-15ee2d78873b`. Revised August 2023, Section 5.5.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839.

Jeffrey Zhang, Wei Li, Wang Miao, and Eric J. Tchetgen Tchetgen. Proximal causal inference without uniqueness assumptions. *Statistics & Probability Letters*, 198:109836, 2023. doi: 10.1016/j.spl.2023.109836.

A Additional Algorithms

This section provides pseudocode for the estimators used in the simulation comparisons. Algorithm 2 describes the uncensored NC-CSF pipeline, while Algorithms 3 and 4 give the corresponding observed-data baselines, which use (X, W, Z) directly and remove the negative-control bridge layer. For the censored setting, the matched baseline keeps the same preprocessing, survival nuisance estimation, censoring correction, residual construction, and final causal-forest step as Censored NC-CSF, isolating the effect of proxy-based bridge adjustment. Algorithm 1 gives the full censored NC-CSF implementation, and Algorithm 5 gives the synthetic-only oracle comparator, which has access to U and DGP-specific nuisance functions and serves only as an upper-bound reference. We additionally compare against `grf::causal_survival_forest`, RSF T-/S-learners, and survival BART in the common simulation benchmark; the `grf` implementation is also used in the ACTG175 analysis.

Algorithm 1 Censored NC-CSF Training

Require: Data $\{(X_i, W_i, Z_i, A_i, Y_i, \Delta_i)\}_{i=1}^n$; target type; horizon h; number of folds K; clipping levels $\epsilon_q, \alpha_{\mathrm{br}}, \alpha_{\mathrm{res}}$

Ensure: Fitted effect predictor $\widehat{\tau}(\cdot)$

Partition $\{1, \ldots, n\}$ into cross-fitting folds $\mathcal{I}_1, \ldots, \mathcal{I}_K$.
for $k = 1, \ldots, K$ **do**
  Let $\mathcal{I}_{-k} = \{1, \ldots, n\} \setminus \mathcal{I}_k$ and set$(Y_i^\star, \Delta_i^\star) = (Y_i^{(h)}, \Delta_i^{(h)})$ for RMST, and $(Y_i^\star, \Delta_i^\star) = (Y_i, \Delta_i)$ for survival probability.
  Fit a censoring Cox model on $\mathcal{I}_{-k}$ with covariates (X, W, Z, A), times $Y^\star$, and censoring indicators $1 - \Delta^\star$, yielding the censoring survival $\widehat{S}_C^{(-k)}$ and cumulative censoring hazard increments $d\widehat{\Lambda}_C^{(-k)}$.
  Construct target-specific bridge pseudo-outcomes $\widetilde{Y}_i^{\mathrm{bridge}}$ on $\mathcal{I}_{-k}$, and clip them to empirical $(\alpha_{\mathrm{br}}, 1 - \alpha_{\mathrm{br}})$ quantiles.
  Fit the treatment bridge $\widehat{q}^{(-k)}$ using logistic regression on $\mathcal{I}_{-k}$, with labels A and bridge covariates (X, Z).
  for $a \in \{0, 1\}$ **do**
    Let $\mathcal{I}_{-k,a} = \{i \in \mathcal{I}_{-k} : A_i = a\}$.
    Fit the outcome bridge $\widehat{f}_a^{(-k)}$ using a random forest regressor on $\mathcal{I}_{-k,a}$, with labels $\widetilde{Y}^{\mathrm{bridge}}$ and bridge covariates (X, W).
    Fit event-survival model $\widehat{S}_a^{(-k)}$ using a Cox model on $\mathcal{I}_{-k,a}$, with covariates (X, W, Z), times $Y^\star$, and event indicators $\Delta^\star$.
  end for
  Construct continuation functions $\widehat{\rho}_a^{(-k)}(t; h)$ from $\widehat{S}_a^{(-k)}$, $a \in \{0, 1\}$:
  $\widehat{\rho}_a^{(-k)}(t; h) = t + \int_t^h \frac{\widehat{S}_a^{(-k)}(u|X,W,Z)}{\widehat{S}_a^{(-k)}(t|X,W,Z)}\, du$ for RMST, and $\widehat{\rho}_a^{(-k)}(t; h) = \frac{\widehat{S}_a^{(-k)}(h|X,W,Z)}{\widehat{S}_a^{(-k)}(t|X,W,Z)}$ for survival probability.
  for $i \in \mathcal{I}_k$ **do**
    Predict $\widehat{q}_i = \textsc{Clip}\{\widehat{q}^{(-k)}(X_i, Z_i), \epsilon_q, 1 - \epsilon_q\}$, $\widehat{f}_{a,i} = \widehat{f}_a^{(-k)}(X_i, W_i)$ for $a \in \{0, 1\}$.
    Form $\widehat{m}_i = \widehat{q}_i \widehat{f}_{1,i} + (1 - \widehat{q}_i)\widehat{f}_{0,i}$, and $A_i^{\mathrm{res}} = A_i - \widehat{q}_i$.
    Compute $\widetilde{Y}_i^{\mathrm{res}}$ from Eq. (18) using $\widehat{m}_i$, $\widehat{\rho}_{A_i}^{(-k)}$, $\widehat{S}_C^{(-k)}$, and $d\widehat{\Lambda}_C^{(-k)}$.
    Form $X_i^{\mathrm{final}} = (X_i, W_i, Z_i, \widehat{q}_i, \widehat{f}_{1,i}, \widehat{f}_{0,i}, \widehat{m}_i, \widehat{r}_{1,i}, \widehat{r}_{0,i}, \widehat{r}_{1,i} - \widehat{r}_{0,i})$, where $\widehat{r}_{a,i}$ is the target-specific event-survival summary.
  end for
  Clip $\{\widetilde{Y}_i^{\mathrm{res}} : i \in \mathcal{I}_k\}$ to empirical $(\alpha_{\mathrm{res}}, 1 - \alpha_{\mathrm{res}})$ quantiles, yielding Y_i^{res}.
end for
Fit the final causal forest $\widehat{\mathcal{F}}$ on the cross-fitted training triples $\{(X_i^{\mathrm{final}}, A_i^{\mathrm{res}}, Y_i^{\mathrm{res}})\}_{i=1}^n$.
Refit only the nuisance models on the full training sample. For a new point (x, w, z), construct $X_{\mathrm{new}}^{\mathrm{final}}$ using the same formula above and return $\widehat{\tau}(x) = \widehat{\mathcal{F}}(X_{\mathrm{new}}^{\mathrm{final}})$.
return $\widehat{\tau}(\cdot)$.

Algorithm 2 Uncensored NC-CSF Training

Require: Data $\{(X_i, W_i, Z_i, A_i, Y_i^{\mathrm{unc}})\}_{i=1}^n$; number of folds K; clipping levels $\epsilon_q, \alpha_y, \alpha_{\mathrm{res}}$
Ensure: Fitted effect predictor $\widehat{\tau}(\cdot)$

Partition $\{1, \ldots, n\}$ into cross-fitting folds $\mathcal{I}_1, \ldots, \mathcal{I}_K$.
for $k = 1, \ldots, K$ **do**
  Let $\mathcal{I}_{-k} = \{1, \ldots, n\} \setminus \mathcal{I}_k$.
  Clip $\{Y_i^{\mathrm{unc}} : i \in \mathcal{I}_{-k}\}$ to empirical $(\alpha_y, 1-\alpha_y)$ quantiles, yielding $Y_i^{\mathrm{unc,clip}}$.
  Fit the treatment bridge $\widehat{q}^{(-k)}$ using logistic regression on $\mathcal{I}_{-k}$, with labels A and bridge covariates (X, Z).
  for $a \in \{0, 1\}$ **do**
    Let $\mathcal{I}_{-k,a} = \{i \in \mathcal{I}_{-k} : A_i = a\}$.
    Fit the outcome bridge $\widehat{f}_a^{(-k)}$ using a random forest regressor on $\mathcal{I}_{-k,a}$, with labels $Y^{\mathrm{unc,clip}}$ and bridge covariates (X, W).
  end for
  for $i \in \mathcal{I}_k$ **do**
    Predict $\widehat{q}_i = \textsc{Clip}\{\widehat{q}^{(-k)}(X_i, Z_i), \epsilon_q, 1-\epsilon_q\}$, $\widehat{f}_{a,i} = \widehat{f}_a^{(-k)}(X_i, W_i)$ for $a \in \{0, 1\}$.
    Form $\widehat{m}_i = \widehat{q}_i \widehat{f}_{1,i} + (1-\widehat{q}_i)\widehat{f}_{0,i}$, $A_i^{\mathrm{res}} = A_i - \widehat{q}_i$, and $\widetilde{Y}_i^{\mathrm{res}} = Y_i^{\mathrm{unc}} - \widehat{m}_i$.
    Form $X_i^{\mathrm{final}} = (X_i, W_i, Z_i, \widehat{q}_i, \widehat{f}_{1,i}, \widehat{f}_{0,i}, \widehat{m}_i)$.
  end for
  Clip $\{\widetilde{Y}_i^{\mathrm{res}} : i \in \mathcal{I}_k\}$ to empirical $(\alpha_{\mathrm{res}}, 1-\alpha_{\mathrm{res}})$ quantiles, yielding Y_i^{res}.
end for
Fit the final causal forest $\widehat{\mathcal{F}}$ on the cross-fitted training triples $\{(X_i^{\mathrm{final}}, A_i^{\mathrm{res}}, Y_i^{\mathrm{res}})\}_{i=1}^n$.
Refit only the nuisance models on the full training sample. For a new point (x, w, z), construct $X_{\mathrm{new}}^{\mathrm{final}}$ using the same formula above and return $\widehat{\tau}(x) = \widehat{\mathcal{F}}(X_{\mathrm{new}}^{\mathrm{final}})$.
return $\widehat{\tau}(\cdot)$.

Algorithm 3 Uncensored Baseline

Require: Training data $\{(X_i, W_i, Z_i, A_i, Y_i^{\mathrm{unc}})\}_{i=1}^n$; test point (x, w, z)
Ensure: Baseline CATE estimate $\widehat{\tau}_{\mathrm{base}}^{\mathrm{unc}}(x, w, z)$

Form raw covariates $X_i^{\mathrm{raw}} = (X_i, W_i, Z_i)$ for all training samples.
Fit a causal forest using outcomes Y_i^{unc}, treatments A_i, and covariates X_i^{raw}.
Form the test covariate vector $x^{\mathrm{raw}} = (x, w, z)$.
return $\widehat{\tau}_{\mathrm{base}}^{\mathrm{unc}}(x, w, z)$ by evaluating the fitted causal forest at x^{raw}.

Algorithm 4 Censored Baseline

Require: Training data $\{(X_i, W_i, Z_i, A_i, Y_i, \Delta_i)\}_{i=1}^n$; target type; horizon h if needed; test point (x, w, z)
Ensure: $\widehat{\tau}_{\mathrm{base}}^{\mathrm{cens}}(x, w, z)$

Form the raw observed feature stack

$$X_i^{\mathrm{raw}} = [X_i, W_i, Z_i].$$

Initialize the censored baseline with the same target, horizon, and final forest family as the proposed censored model.
for each nuisance fold **do**
  Fit an observed-data propensity model on $[X, W, Z]$.
  Fit treated and control Cox event-survival models using the target-preprocessed time/event outcome.
  Fit a Cox censoring model on $(Y_i, 1-\Delta_i)$ with features $[X_i, W_i, Z_i, A_i]$.
  Construct the residual pair $(Y_i^{\mathrm{res}}, A_i^{\mathrm{res}})$ from these observed-data nuisances.
end for
Fit the final causal forest on $\{(X_i^{\mathrm{raw}}, A_i^{\mathrm{res}}, Y_i^{\mathrm{res}})\}_{i=1}^n$.
Form the test feature vector $x^{\mathrm{raw}} = [x, w, z]$.
return $\widehat{\tau}_{\mathrm{base}}^{\mathrm{cens}}(x, w, z) = \texttt{effect}(x^{\mathrm{raw}})$.

Algorithm 5 NC-CSF Oracle Comparator

Require: Synthetic training data with oracle access to latent confounders U and DGP nuisances; setting type $\in \{\text{uncensored}, \text{censored}\}$; test point (x, u)
Ensure: Oracle CATE estimate $\widehat{\tau}_{\text{oracle}}(x, u)$
Use (X, U) in place of proxy-based bridge inputs and obtain oracle nuisances from the DGP.
if uncensored setting **then**
    For each training unit, obtain $\widehat{q}_i^{\text{oracle}} = q_0(X_i, U_i)$ and $\widehat{f}_{a,i}^{\text{oracle}} = f_a(X_i, U_i)$ for $a \in \{0, 1\}$.
    Form $\widehat{m}_i^{\text{oracle}} = \widehat{q}_i^{\text{oracle}} \widehat{f}_{1,i}^{\text{oracle}} + (1 - \widehat{q}_i^{\text{oracle}}) \widehat{f}_{0,i}^{\text{oracle}}$, $A_i^{\text{res}} = A_i - \widehat{q}_i^{\text{oracle}}$, and $Y_i^{\text{res}} = Y_i^{\text{unc}} - \widehat{m}_i^{\text{oracle}}$, with the same clipping rules as Uncensored NC-CSF.
    Form X_i^{final} by augmenting (X_i, U_i) with oracle nuisance summaries.
else
    Set $(Y_i^\star, \Delta_i^\star) = (Y_i^{(h)}, \Delta_i^{(h)})$ for RMST, and $(Y_i^\star, \Delta_i^\star) = (Y_i, \Delta_i)$ for survival probability.
    For each training unit, obtain oracle bridge nuisances $\widehat{q}_i^{\text{oracle}} = q_0(X_i, U_i)$ and $\widehat{f}_{a,i}^{\text{oracle}} = f_a(X_i, U_i; h)$, and oracle survival nuisances $\widehat{S}_a^{\text{oracle}}(t \mid X_i, U_i)$, $S_C^{\text{oracle}}(t \mid X_i, U_i, A_i)$, and $d\Lambda_C^{\text{oracle}}(t \mid X_i, U_i, A_i)$.
    Form $\widehat{m}_i^{\text{oracle}} = \widehat{q}_i^{\text{oracle}} \widehat{f}_{1,i}^{\text{oracle}} + (1 - \widehat{q}_i^{\text{oracle}}) \widehat{f}_{0,i}^{\text{oracle}}$ and $A_i^{\text{res}} = A_i - \widehat{q}_i^{\text{oracle}}$.
    Compute Y_i^{res} using the oracle version of Eq. (18).
    Form X_i^{final} by augmenting (X_i, U_i) with oracle bridge and survival nuisance summaries.
end if
Fit a causal forest to $\{(X_i^{\text{final}}, A_i^{\text{res}}, Y_i^{\text{res}})\}_{i=1}^n$.
Construct the corresponding oracle final feature vector $X_{\text{new}}^{\text{final}}$ for test point (x, u).
return the fitted causal forest prediction at $X_{\text{new}}^{\text{final}}$.

B Bridge Estimation and Operator Stability

Completeness is an identification condition, not a numerical-stability guarantee. The latent completeness conditions in Proposition 3.2 ensure that the observable moments imply the latent bridge relations. Separately, uniqueness of the bridge as a function of the observed proxies requires injectivity, on the chosen bridge class, of the observable conditional-expectation operators

$$(\mathcal{T}_q q)(x, w) = \mathbb{E}\{q(Z, X) \mid X = x, W = w\},$$
$$(\mathcal{T}_{f,a} f_a)(x, z) = \mathbb{E}\{f_a(W, X) \mid X = x, Z = z, A = a\}.$$

Neither form of identification implies that the inverse is continuous or well conditioned.

For a correctly specified finite-dimensional bridge b_θ, let $M(\theta)$ be the vector of observable moment restrictions and $J_0 = \nabla_\theta M(\theta_0)$. Local identification requires J_0 to have full column rank, whereas numerical stability depends on its smallest singular value. After standardizing the moments and parameters, the condition number is

$$\text{cond}(J_0) = \frac{\sigma_{\max}(J_0)}{\sigma_{\min}(J_0)}.$$

A moment perturbation of size δ can therefore produce a bridge-parameter error of order $\delta / \sigma_{\min}(J_0)$. Thus completeness may hold even when the condition number is large and bridge estimation is highly sensitive to sampling or approximation error.

For a nonparametric bridge class $\mathcal{B}$, the analogous restricted stability constant is

$$\kappa_{\mathcal{B}} = \inf_{\substack{b \in \mathcal{B} \\ b \neq 0}} \frac{\|\mathcal{T} b\|_{\mathcal{H}}}{\|b\|_{L_2}}.$$

Stable recovery on the effective class requires $\kappa_{\mathcal{B}}$ to be bounded away from zero; completeness alone only rules out an exact null direction. Ridge or related regularization stabilizes a poorly conditioned inverse by trading a smaller estimation term against regularization bias. Neyman orthogonality addresses a separate issue: it limits how sufficiently small bridge-estimation errors propagate into the final effect estimator, but it does not itself ensure accurate bridge estimation. In practice, we examine cross-fitted moment residuals, sensitivity to regularization and clipping, and stability across scientifically defensible proxy routings.

C Proofs

Proof C.1 (Proof of Proposition 3.2) *We first prove* (6)*. For the treatment bridge,*

$$\begin{aligned}\mathbb{E}[A_i - q(Z_i, X_i) \mid X_i, U_i, W_i] &= \mathbb{E}[A_i \mid X_i, U_i, W_i] - \mathbb{E}[q(Z_i, X_i) \mid X_i, U_i, W_i] \\ &= \mathbb{P}(A_i = 1 \mid X_i, U_i) - \mathbb{E}[q(Z_i, X_i) \mid X_i, U_i] \\ &= 0,\end{aligned}$$

where the second equality uses $(A_i, Z_i) \perp W_i \mid X_i, U_i$*, and the third follows from* (2)*. Taking conditional expectation with respect to* (X_i, W_i) *yields*

$$\mathbb{E}[A_i - q(Z_i, X_i) \mid X_i, W_i] = 0.$$

Similarly, for the outcome bridge,

$$\begin{aligned}&\mathbb{E}[G_i(a; h) - f_a(W_i, X_i) \mid X_i, U_i, Z_i, A_i = a] \\ &\quad= \mathbb{E}[G_i(a; h) \mid X_i, U_i, Z_i, A_i = a] - \mathbb{E}[f_a(W_i, X_i) \mid X_i, U_i, Z_i, A_i = a] \\ &\quad= \mathbb{E}[G_i(a; h) \mid X_i, U_i] - \mathbb{E}[f_a(W_i, X_i) \mid X_i, U_i] \\ &\quad= 0,\end{aligned}$$

where the third equality uses $(A_i, Z_i) \perp (G_i(a; h), W_i) \mid X_i, U_i$*, and the last follows from* (2)*. Taking conditional expectation with respect to* $(X_i, Z_i, A_i = a)$ *yields*

$$\mathbb{E}[G_i(a; h) - f_a(W_i, X_i) \mid X_i, Z_i, A_i = a] = 0.$$

We now prove the converse. Suppose (6) *holds. Define*

$$\begin{aligned}r(X_i, U_i) &= \mathbb{P}(A_i = 1 \mid X_i, U_i) - \mathbb{E}[q(Z_i, X_i) \mid X_i, U_i]\,, \\ s(X_i, U_i) &= \mathbb{E}[G_i(a; h) \mid X_i, U_i] - \mathbb{E}[f_a(W_i, X_i) \mid X_i, U_i]\,.\end{aligned}$$

By iterated expectation and $(A_i, Z_i) \perp W_i \mid X_i, U_i$*,*

$$\begin{aligned}\mathbb{E}[r(X_i, U_i) \mid X_i, W_i] &= \mathbb{E}[\mathbb{E}[A_i - q(Z_i, X_i) \mid X_i, U_i, W_i] | X_i, W_i] \\ &= \mathbb{E}[A_i - q(Z_i, X_i) \mid X_i, W_i] \\ &= 0.\end{aligned}$$

By the first completeness condition, this implies $r(X_i, U_i) = 0$ *a.s., that is,*

$$\mathbb{E}[q(Z_i, X_i) \mid X_i, U_i] = \mathbb{P}(A_i = 1 \mid X_i, U_i).$$

Similarly, using $(A_i, Z_i) \perp (G_i(a; h), W_i) \mid X_i, U_i$*,*

$$\begin{aligned}\mathbb{E}[s(X_i, U_i) \mid X_i, Z_i, A_i = a] &= \mathbb{E}[\mathbb{E}[G_i(a; h) - f_a(W_i, X_i) \mid X_i, U_i, Z_i, A_i = a] | X_i, Z_i, A_i = a] \\ &= \mathbb{E}[G_i(a; h) - f_a(W_i, X_i) \mid X_i, Z_i, A_i = a] \\ &= 0.\end{aligned}$$

Similarly for the second completeness condition.

Proof C.2 (Proof of Proposition 3.3) *We first prove the uncensored case, and then note that the censored case follows by the same argument using the censoring-corrected pseudo-outcome as in the proof of Proposition 3.5.*

For the uncensored case, we aim to show

$$\mathbb{E}[(A_i - q_i)(G_i - m_i - \tau^w(X_i)(A_i - q_i))|X_i = x] = 0.$$

In fact, we can decompose $(A_i - q_i)(G_i - m_i)$ *as*

$$A_i(A_i - q_i)(Y_i(1) - f_{1,i}) + (1 - A_i)(A_i - q_i)(Y_i(0) - f_{0,i}) + (A_i - q_i)^2(f_{1,i} - f_{0,i}).$$

Taking conditional expectation given (X_i, U_i)*, the first two terms are zero by* (1) *and* (2) *as*

$$\mathbb{E}[A_i(A_i - q_i)(Y_i(1) - f_{1,i}) \mid X_i, U_i] = \mathbb{E}[A_i(A_i - q_i) \mid X_i, U_i]\,\mathbb{E}[Y_i(1) - f_{1,i} \mid X_i, U_i] = 0,$$

and similarly for the second term. Therefore,

$$\begin{aligned}\mathbb{E}[(A_i - q_i)(G_i - m_i) \mid X_i, U_i] &= \mathbb{E}\big[(A_i - q_i)^2 \mid X_i, U_i\big]\, \mathbb{E}[f_{1,i} - f_{0,i} \mid X_i, U_i] \\ &= \mathbb{E}\big[(A_i - q_i)^2 \tau(X_i, U_i) \mid X_i, U_i\big]\,.\end{aligned}$$

Taking conditional expectation given $X_i = x$ yields

$$\mathbb{E}[(A_i - q_i)(G_i - m_i) \mid X_i = x] = \mathbb{E}\big[(A_i - q_i)^2 \tau(X_i, U_i) \mid X_i = x\big]\,.$$

Therefore, the NC-CSF loss satisfies

$$\begin{aligned}&\mathbb{E}[(A_i - q_i)(G_i - m_i - \tau(X_i)(A_i - q_i))|X_i = x] \\ &\quad = \mathbb{E}\big[(A_i - q_i)^2 \tau(X_i, U_i) \mid X_i = x\big] - \tau(x)\mathbb{E}\big[(A_i - q_i)^2 \mid X_i = x\big]\,,\end{aligned}$$

which is zero at

$$\tau^w(x) = \frac{\mathbb{E}\big[(A_i - q_i)^2 \tau(X_i, U_i) \mid X_i = x\big]}{\mathbb{E}[(A_i - q_i)^2 \mid X_i = x]}.$$

Finally, if $\tau(x, u) = \tau(x)$, then the numerator of $\tau^w(x)$ simplifies to $\tau(x)\mathbb{E}\big[(A_i - q_i)^2 \mid X_i = x\big]$, and as a result $\tau^w(x) = \tau(x)$.

Proof C.3 (Proof of Proposition 3.5) *Let $V_i = (X_i, W_i, Z_i, A_i)$. Under* (3)*, $\mathbb{E}[\Gamma_i \mid V_i] = \mathbb{E}[G_i \mid V_i]$ according to Cui et al. [2023]. Therefore,*

$$\mathbb{E}[Y^{\text{res}} \mid X] = \mathbb{E}[\mathbb{E}[Y^{\text{res}} \mid V]|X] = \mathbb{E}[\mathbb{E}[G - m \mid V]|X] = \mathbb{E}[G - m \mid X]\,.$$

Next,

$$\begin{aligned}&\mathbb{E}[Y^{\text{res}} \mid X] = \mathbb{E}[G - m \mid X] \\ =&\mathbb{E}\big[\mathbb{E}\big[AG(1;h) + (1-A)G(0;h) - q(Z,X)f_1(W,X) - \big(1 - q(Z,X)\big)f_0(W,X) \mid X, U\big]\big|X\big]\,.\end{aligned}$$

By (1) *and* (2)*,*

$$\begin{aligned}\mathbb{E}[AG(1;h) \mid X, U] &= \mathbb{E}[A \mid X, U] \cdot \mathbb{E}[G(1;h) \mid X, U] \\ &= \mathbb{E}[q(Z,X) \mid X, U] \cdot \mathbb{E}[f_1(W,X) \mid X, U] \\ &= \mathbb{E}[q(Z,X) f_1(W,X) \mid X, U],\end{aligned}$$

and similarly

$$\mathbb{E}[(1-A)G(0;h) \mid X, U] = \mathbb{E}[(1 - q(Z,X)) f_0(W,X) \mid X, U].$$

Therefore, $\mathbb{E}[Y^{\text{res}}_{\eta_0} \mid X] = 0$. Again by (2) *and the tower property,*

$$\mathbb{E}\big[A^{\text{res}}_{\eta_0} \mid X\big] = \mathbb{E}[\mathbb{E}[A - q(Z,X) \mid X, U]|X] = 0.$$

For the bridge q, since $A^{\text{res}} = A - q$, the derivative of the score with respect to q is proportional to

$$-\big(Y^{\text{res}}_{\eta_0} - 2\tau_0(X) A^{\text{res}}_{\eta_0}\big).$$

Taking conditional expectation given X and using the two centering identities above yields zero. For m, the derivative of the score is proportional to $-A^{\text{res}}_{\eta_0}$, whose conditional expectation given X is also zero by the above argument.

For the survival-related nuisance functions S_C, ρ, define the censoring counting process and at-risk process by $N_{C,i}(t) = I(Y_i \le t, \Delta_i = 0)$ and $R_i(t) = I(Y_i \ge t)$. Under (3)*,*

$$M_{C,i}(t \mid V_i) = N_{C,i}(t) - \int_0^t R_i(u)\, d\Lambda_C(u \mid V_i)$$

is a martingale. For any perturbation $\delta(t;h)$ of $\rho(t;h)$, the first-order derivative of the score with respect to ρ is proportional to

$$A^{\text{res}}_i \int_0^h \frac{\delta_{A_i}(t;h)}{S_C(t \mid V_i)}\, dM_{C,i}(t \mid V_i),$$

which has mean zero conditional on X_i. The same argument applies to perturbations of S_C and Λ_C.

Proof C.4 (Proof of Proposition 3.6) *Let $\widetilde{\tau}_{\mathcal{S}}$ be the oracle estimator using the true nuisance functions. The argument in the proof of Proposition 3.3, now averaging over $X \in \mathcal{S}$, gives*

$$\mathbb{E}[A_i^{\mathrm{res}}(Y_i^{\mathrm{res}} - tA_i^{\mathrm{res}}) \mid X_i \in \mathcal{S}] = \mathbb{E}[(A_i^{\mathrm{res}})^2 \tau(X_i, U_i) \mid X_i \in \mathcal{S}] - t\,\mathbb{E}[(A_i^{\mathrm{res}})^2 \mid X_i \in \mathcal{S}],$$

whose unique zero is $t = \tau_{\mathcal{S}}^w$. Thus the oracle estimator is a ratio of empirical moments centered at $\tau_{\mathcal{S}}^w$. Boundedness, residual overlap, and $n_{\mathcal{S}} \to \infty$ yield the displayed central limit theorem and variance by the multivariate CLT and delta method.

For the feasible estimator, cross-fitting makes each observation independent of the nuisance estimates used to construct its score, conditional on the training folds. Proposition 3.5 removes the first-order Gateaux derivative at the truth. The remaining score difference consists of products of nuisance errors and is $o_p(n_{\mathcal{S}}^{-1/2})$ under the stated $o_p(n_{\mathcal{S}}^{-1/4})$ rates and boundedness conditions. Hence

$$\widehat{\tau}_{\mathcal{S}} - \widetilde{\tau}_{\mathcal{S}} = o_p(n_{\mathcal{S}}^{-1/2}),$$

and Slutsky's theorem gives the result. If $\tau(x, u)$ is constant on $\mathcal{S}$, the weighted target reduces to that common value.

D Additional Numerical Details and Results

D.1 Numerical Experiment Details

We describe more implementation details of NC-CSF using our proposed loss with estimated nuisance functions.

D.1.1 Uncensored NC-CSF

In the uncensored setting, the nuisance layer consists of

$$q_i = q(Z_i, 1, X_i), \qquad f_{1,i} = f(W_i, 1, X_i), \qquad f_{0,i} = f(W_i, 0, X_i),$$

together with the bridge mixture

$$m_i = q(Z_i, 1, X_i) f(W_i, 1, X_i) + \big(1 - q(Z_i, 1, X_i)\big) f(W_i, 0, X_i).$$

We then form the orthogonal residuals

$$A_i^{\mathrm{res}} = A_i - q_i, \qquad Y_i^{\mathrm{res}} = Y_i^{\mathrm{unc}} - m_i. \tag{8}$$

The final-stage feature vector augments the raw observed covariates with nuisance summaries:

$$X_i^{\mathrm{final}} = [X_i,\ W_i,\ Z_i,\ q_i,\ f_{1,i},\ f_{0,i},\ m_i]. \tag{9}$$

The final learner is a causal forest fit on the bridge-adjusted residualized data. In our implementation, we use the `CausalForestDML` interface in EconML with externally constructed residuals and final-stage features.

D.1.2 Censored NC-CSF

We now extend the estimator to right-censored outcomes. As in the uncensored case, we estimate the treatment bridge q, the treatment-specific outcome bridges $f(W, 1, X)$ and $f(W, 0, X)$, and the bridge mixture m, and define

$$A_i^{\mathrm{res}} = A_i - \widehat{q}_i. \tag{10}$$

For censored outcomes, the outcome bridges are trained using a target-specific IPCW bridge pseudo-outcome $\widetilde{Y}_i^{\mathrm{bridge}}$, used only for nuisance estimation. For the RMST target, using the truncated time/event pair $(\widetilde{Y}_i, \widetilde{\Delta}_i)$, we use

$$\widetilde{Y}_i^{\mathrm{bridge}} = \frac{\widetilde{Y}_i \widetilde{\Delta}_i}{\widehat{S}_C(\widetilde{Y}_i \mid X_i, W_i, Z_i, A_i)}. \tag{11}$$

For the survival-probability target, let

$$\widetilde{Y}_i^{(h)} = \min(Y_i, h), \qquad \widetilde{\Delta}_i^{(h)} = \Delta_i \mathbf{1}\{Y_i \le h\} + \mathbf{1}\{Y_i > h\}, \tag{12}$$

and define

$$\widetilde{Y}_i^{\text{bridge}} = \frac{\mathbf{1}\{Y_i > h\}}{\widehat{S}_C(\widetilde{Y}_i^{(h)} \mid X_i, W_i, Z_i, A_i)}. \tag{13}$$

We also fit arm-specific event-survival curves

$$\widehat{S}_a(t \mid X_i, W_i, Z_i) \approx \widehat{\mathbb{P}}(T_i > t \mid X_i, W_i, Z_i, A_i = a), \qquad a \in \{0, 1\},$$

using separate Cox models within treatment arms. For RMST, we use

$$T_{i,\text{nuis}} = \min(Y_i, h), \qquad \Delta_{i,\text{nuis}} = \Delta_i \mathbf{1}\{Y_i < h\} + \mathbf{1}\{Y_i \geq h\}. \tag{14}$$

For survival probability, the Cox duration/event pair is (Y_i, Δ_i).

These event-survival curves define target-specific continuation terms. For RMST,

$$\rho_a^{\text{RMST}}(t, X_i; h) = t + \int_t^h \frac{\widehat{S}_a(u \mid X_i, W_i, Z_i)}{\widehat{S}_a(t \mid X_i, W_i, Z_i)}\, du, \qquad 0 \leq t \leq h, \quad a \in \{0, 1\}. \tag{15}$$

For survival probability,

$$\rho_a(t, X_i; h) = \frac{\widehat{S}_a(h \mid X_i, W_i, Z_i)}{\widehat{S}_a(t \mid X_i, W_i, Z_i)}, \qquad 0 \leq t \leq h, \quad a \in \{0, 1\}. \tag{16}$$

Because U_i is unobserved, our main specification uses a proxy-adjusted working model for censoring,

$$\widehat{S}_C(t \mid X_i, W_i, Z_i, A_i), \tag{17}$$

estimated by a censoring Cox model using $(Y_i, 1 - \Delta_i)$ and covariates (X_i, W_i, Z_i, A_i).

The outcome residual plugs the bridge-adjusted regression term into the three-term censoring-robust score of Cui et al. [2023]. For survival probability,

$$\begin{aligned} Y_i^{\text{res}} = {} & \frac{\widetilde{\Delta}_i^{(h)}\big(\mathbf{1}\{Y_i > h\} - \widehat{m}_i\big) + \big(1 - \widetilde{\Delta}_i^{(h)}\big)\big(\rho_{A_i}(\widetilde{Y}_i^{(h)}, X_i; h) - \widehat{m}_i\big)}{\widehat{S}_C(\widetilde{Y}_i^{(h)} \mid X_i, W_i, Z_i, A_i)} \\ & - \int_0^{\widetilde{Y}_i^{(h)}} \frac{\rho_{A_i}(t, X_i; h) - \widehat{m}_i}{\widehat{S}_C(t \mid X_i, W_i, Z_i, A_i)}\, d\widehat{\Lambda}_C(t \mid X_i, W_i, Z_i, A_i). \end{aligned} \tag{18}$$

For RMST, we replace $\mathbf{1}\{Y_i > h\}$ with $\widetilde{Y}_i \widetilde{\Delta}_i$, and replace $\rho_a(t, x; h)$ with $\rho_a^{\text{RMST}}(t, x; h)$.

The final-stage feature vector is

$$X_i^{\text{final}} = [X_i,\ W_i,\ Z_i,\ \widehat{q}_i,\ \widehat{f}_{1,i},\ \widehat{f}_{0,i},\ \widehat{m}_i,\ \widehat{S}_{1,i}(h),\ \widehat{S}_{0,i}(h),\ \widehat{S}_{1,i}(h) - \widehat{S}_{0,i}(h)]. \tag{19}$$

Here $\widehat{S}_{a,i}(h)$ denotes the target-specific scalar summary derived from the fitted event-survival curve under arm a. Clipping thresholds are specified in Appendix Algorithm 1.

D.2 Data-Generating Mechanism

We generate n i.i.d. samples with covariates $X_i \sim \mathcal{N}(0, I_p)$ and latent confounder $U_i \sim \mathcal{N}(0, 1)$. To model negative controls, we construct proxies Z_i and W_i as noisy linear functions of (X_i, U_i):

$$Z_i = a_z U_i + b_z^\top X_i + \varepsilon_{z,i}, \quad W_i = a_w U_i + b_w^\top X_i + \varepsilon_{w,i},$$

with independent Gaussian noise.

Treatment assignment follows

$$A_i \sim \text{Bernoulli}\big(\sigma(b_0 + \gamma_u U_i + \alpha^\top X_i)\big),$$

introducing unobserved confounding via U_i. Potential event times follow a Weibull proportional hazards model:

$$T_i(a) \mid X_i, U_i \sim \text{Weibull}\big(k, \lambda e^{-\eta_i(a)/k}\big), \quad \eta_i(a) = \beta_t^\top X_i + \beta_u U_i + \tau a.$$

The conditional treatment effect under full observability is

$$\mathbb{E}[T_i(1) - T_i(0) \mid X_i, U_i] = \lambda\Gamma(1 + 1/k) \exp\big(-\tfrac{1}{k}(\beta_t^\top X_i + \beta_u U_i)\big) \big(e^{-\tau/k} - 1\big). \tag{20}$$

Although U_i is unobserved, its conditional distribution given (Z_i, W_i) is available due to the linear-Gaussian structure. Defining residualized proxies $\tilde{Z}_i = Z_i - b_z^\top X_i$ and $\tilde{W}_i = W_i - b_w^\top X_i$, we have

$$U_i \mid Z_i, W_i \sim \mathcal{N}\big(\mu_{U|Z,W},\, \sigma^2_{U|Z,W}\big),$$

with closed-form expressions for the mean and variance. Integrating over this distribution yields the observed-data treatment effect:

$$\begin{aligned} &\mathbb{E}[T_i(1) - T_i(0) \mid X_i, Z_i, W_i] \\ &= \lambda\,\Gamma\left(1 + \frac{1}{k}\right) \exp\left(-\frac{1}{k}\beta_t^\top X_i - \frac{\beta_u}{k}\mu_{U|Z,W} + \frac{1}{2}\frac{\beta_u^2}{k^2}\sigma^2_{U|Z,W}\right) \left(e^{-\tau/k} - 1\right). \end{aligned} \tag{21}$$

We evaluate performance across regimes varying proxy quality (σ_z, σ_w), confounding strength (γ_u, β_u), treatment effect size $\tau_{\text{log-hr}}$, and dimensionality (p_x, p_z, p_w), under both linear and nonlinear models. We fix $a_z = a_w = 1.5$; crossing the 12 base cases with the 22 dimensional settings yields 12×22 configurations.

D.3 Proxy-map stress tests beyond linear Gaussian proxies

To isolate the proxy mechanism, we replace the baseline construction by

$$Z = a_Z \widetilde{g}(U) + b_Z^\top X + \varepsilon_Z, \qquad\qquad W = a_W \widetilde{g}(U) + b_W^\top X + \varepsilon_W,$$

where $\widetilde{g}$ is standardized to have mean zero and variance one under $U \sim N(0,1)$. Treatment, event times, censoring, X, U, and the target effects are held fixed across arms. P0 is the baseline linear-Gaussian proxy model. P1 uses $g(u) = \tanh(u)$; P2 uses $g(u) = u$ with t_3 noise scaled to preserve variance; P3 uses the noninjective map $g(u) = u + 0.5u^2$; and P4 coarsens each noisy proxy to one bit. P1 and P2 remove joint Gaussianity while preserving an essentially invertible signal. P3 and P4 remove exact invertibility or fine-grained proxy information.

Table 3: Proxy-map stress tests over all 12×22 configurations. Entries are average RMSE; gain is the relative reduction from the matched Censored Baseline.

Proxy arm	RMST: Baseline → NC-CSF	Survival: Baseline → NC-CSF
P0 (linear-Gaussian)	0.1662 → 0.0994 (40.2%)	0.0788 → 0.0613 (22.2%)
P1 (tanh)	0.1641 → 0.1012 (38.3%)	0.0791 → 0.0620 (21.6%)
P2 (t_3 noise)	0.1621 → 0.0993 (38.8%)	0.0798 → 0.0603 (24.5%)
P3 (quadratic, noninjective)	0.1795 → 0.1159 (35.4%)	0.0856 → 0.0693 (19.1%)
P4 (binary coarsening)	0.1769 → 0.1079 (39.0%)	0.0830 → 0.0661 (20.4%)

NC-CSF improves on the matched baseline in every arm, with RMST reductions of 35.4–40.2% and survival-probability reductions of 19.1–24.5%. P3 and P4 nevertheless degrade the bridge-based method in the predicted direction. In an n-sweep from 500 to 16,000, excess RMST RMSE relative to P0 averages only $+0.9\%$ for P1 and $+0.8\%$ for P2 and changes sign across sample sizes, whereas it is positive at every sample size and averages $+12.8\%$ for P3 and $+9.2\%$ for P4. Thus more data repair an identified but difficult moment system more readily than proxy information lost through a noninjective or coarsened map.

P3 degrades NC-CSF RMST accuracy by 16.6% relative to P0, compared with 8.1% for the no-bridge ablation. This approximately twofold differential degradation localizes the damage to the bridge component rather than to the forest or censoring correction. The bridge is degraded, not disabled: $\text{corr}\{g(U), U\} = 0.816$, and the linear R^2 of U on (Z, W, X) falls from 0.733 under P0 to 0.489 under P3 and 0.503 under P4.

D.4 Paired Monte Carlo uncertainty

We reran the representative sample-size and dimension sweeps 30 times. NC-CSF and the matched baseline use the same training draw within each replication, and performance is evaluated on an independent test set of 10,000 observations against the observed-data target $\mathbb{E}[T(1)-T(0) \mid X, Z, W]$. Table 4 reports mean RMSE plus or minus Monte Carlo standard error, the paired NC-CSF-minus-baseline difference with a pointwise 95% interval, and the number of NC-CSF wins.

Table 4: Monte Carlo uncertainty over 30 paired replications.

Setting	n/d	Baseline RMSE	NC-CSF RMSE	NC-CSF − baseline [95% CI]	Reduction	Wins
Case 10 (n)	1,000	0.1404 ± 0.0057	0.1079 ± 0.0058	-0.0325 $[-0.0411, -0.0240]$	23.2%	29/30
Case 10 (n)	4,000	0.1282 ± 0.0060	0.0892 ± 0.0027	-0.0390 $[-0.0490, -0.0291]$	30.5%	30/30
Case 10 (n)	8,000	0.1158 ± 0.0044	0.0810 ± 0.0028	-0.0348 $[-0.0421, -0.0275]$	30.1%	30/30
Case 10 (n)	16,000	0.1031 ± 0.0025	0.0793 ± 0.0024	-0.0238 $[-0.0298, -0.0179]$	23.1%	29/30
Case 12 (d)	2	0.0588 ± 0.0027	0.0308 ± 0.0023	-0.0280 $[-0.0330, -0.0229]$	47.6%	29/30
Case 12 (d)	5	0.0321 ± 0.0008	0.0191 ± 0.0009	-0.0130 $[-0.0149, -0.0112]$	40.5%	30/30
Case 12 (d)	10	0.0508 ± 0.0015	0.0347 ± 0.0027	-0.0161 $[-0.0209, -0.0113]$	31.7%	28/30
Case 12 (d)	20	0.2207 ± 0.0245	0.1256 ± 0.0069	-0.0952 $[-0.1443, -0.0460]$	43.1%	30/30
Case 12 (d)	40	0.4359 ± 0.0313	0.2816 ± 0.0103	-0.1543 $[-0.2205, -0.0881]$	35.4%	30/30

Every paired interval lies below zero, and NC-CSF wins 28–30 of 30 replications in each setting. These intervals quantify Monte Carlo uncertainty in mean test RMSE, not uncertainty for an individual CATE.

D.5 Marginal ATE learners

AIPW and TMLE target a marginal average effect and therefore are not directly comparable with individual-level CATE learners. For a transparent auxiliary comparison, we estimate the population-average survival-probability effect at the same horizon, assign that scalar estimate to every test individual, and evaluate it against the true individual effects from the simulation.

Table 5: Population-level learners evaluated as constant individual-effect predictors.

Model	Survival-probability RMSE	Survival-probability MAE
AIPW	0.1127	0.0879
TMLE	0.1063	0.0875

The table is included to situate the benchmark, not as a head-to-head HTE comparison: a constant marginal prediction cannot represent genuine individual heterogeneity.

D.6 Right Heart Catheterization Details

Table 6 reports the SUPPORT-study comparison used in the main text. Negative values indicate lower 30-day survival under RHC.

Table 6: RHC average treatment-effect estimates for binary 30-day survival. Literature estimates are from Kallus et al. [2021].

Estimator	ATE estimate	SE	95% CI
NC-CSF	-0.0433	0.0110	[-0.0649, -0.0217]
Kallus et al. DR	-0.0219	0.0051	[-0.0319, -0.0119]
Kallus et al. DR(sta)	-0.0271	0.0043	[-0.0355, -0.0188]
CPSMT20 linear bridge	-0.0607	0.0055	[-0.0714, -0.0500]
VV15 no-unmeasured-confounding	-0.0612	0.0141	[-0.0889, -0.0335]

D.7 ACTG175 Proxy-Split Sensitivity

The main ACTG175 analysis uses the clinical $X/W/Z$ split described in Table 7. Because ACTG175 does not include variables that were designed as negative-control proxies, we also fit two additional clinically motivated proxy splits. These specifications should not be interpreted as definitive proximal identification designs; rather, they test how much the fitted age profile changes when immune-progression, disease-severity, and treatment-selection variables are routed through the proxy channels.

All sensitivity analyses used the same ACTG175 subset, treatment coding, RMST target, and horizon as the main analysis: $n = 1{,}083$, with 561 ddI controls, 522 ZDV+ddI treated subjects, and $h = 1{,}000$ days. For a fair comparison, the three NC-CSF proxy-split specifications and the re-run R-CSF baseline were fit with 10,000 trees in the final forest stage. Bootstrap uncertainty summaries used 200 nonparametric bootstrap replications, with model seed 42, bootstrap seed 2026, and the same fixed age-reference grid across all specifications.

Table 7: ACTG175 proxy-split sensitivity specifications. Variables not listed under W or Z are kept in X.

Specification	W **variables**	Z **variables**	**Rationale**
Clinical split	CD4 count at 20 ± 5 weeks, CD8 count at 20 ± 5 weeks	Race, prior antiretroviral history	Follow-up immune markers are outcome-side progression proxies; race and antiretroviral history may proxy treatment-selection patterns in the analyzed subset.
Rec A	Karnofsky score, symptomatic status, baseline CD4, baseline CD8, CD4 at 20 ± 5 weeks, CD8 at 20 ± 5 weeks	Prior antiretroviral history, intravenous drug use, haemophilia	W emphasizes disease severity and immune progression; Z emphasizes variables that may affect regimen selection or anticipated adherence risk.
Rec B	Baseline CD4, baseline CD8, CD4 at 20 ± 5 weeks, CD8 at 20 ± 5 weeks	Karnofsky score, symptomatic status, prior antiretroviral history, haemophilia	W isolates lab-based immune measurements; Z collects decision-relevant clinical status and history variables.

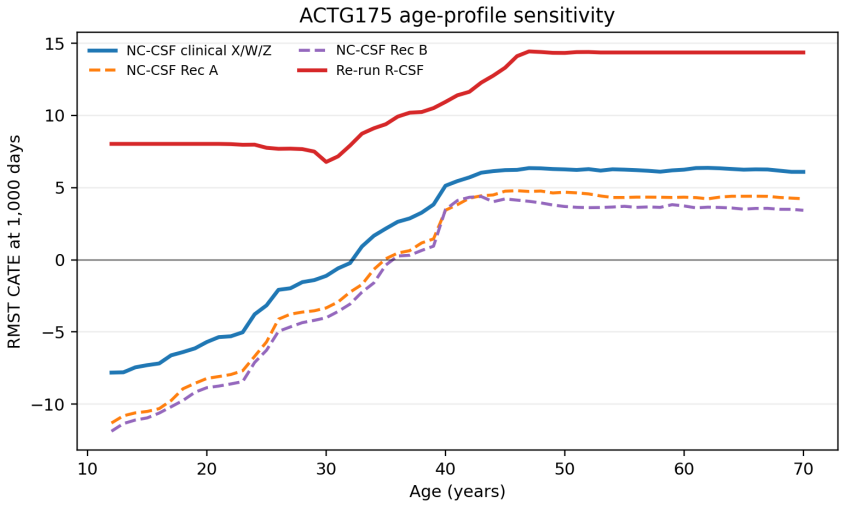

Figure 2: Unified ACTG175 age-profile comparison for the main NC-CSF clinical split, two additional proxy-split sensitivity specifications, and our re-run R-CSF baseline. All curves target RMST at $h = 1{,}000$ days and use the same 10,000-tree final-stage setting.

The sensitivity curves in Figure 2 preserve the main qualitative pattern: estimated benefit is smaller for younger patients and larger for middle-aged or older patients. The proxy-split NC-CSF models have average fitted effects between 3.76 and 5.93 days, compared with 9.68 days for the re-run R-CSF baseline. Their peak ages range from 43 to 62 years, while their peak fitted effects range

from 4.38 to 6.37 days. This indicates that the split choice changes the magnitude and smoothness of the age profile, but does not overturn the broad conclusion that age is the most visible heterogeneity dimension in this dataset.

D.8 MIMIC-IV Prespecified Comorbidity Axes

The four comorbidity axes were prespecified from clinical mechanism before the estimates were inspected. Prior liver disease is the only axis directly linked to tacrolimus disposition through hepatic CYP3A metabolism. The available results are summarized in Table 8.

Table 8: Summary of all four prespecified MIMIC-IV comorbidity axes in the positive-control arm.

Axis	Relative magnitude	Difference CI	Summary
Prior liver disease	1.66	Excludes zero	CATE $= -0.3532$; E-value $= 11.16$
Three nonmetabolic axes	0.85, 1.02, 0.82	All include zero	No clear heterogeneity

For prior liver disease, the estimated effect is -0.3532 (95% CI $[-0.3990, -0.2737]$) among 449 patients, compared with -0.2131 (95% CI $[-0.2692, -0.1569]$) among 1,808 patients without prior liver disease. The difference from the cohort effect is -0.1122 with 95% CI $[-0.1395, -0.0463]$. In contrast, all three prespecified axes without a metabolic link have difference intervals covering zero.

D.9 Marginal versus Conditional Censoring

Our main specification uses conditional censoring adjustment, modeling the censoring survival as $S_C(t \mid X_i, W_i, Z_i, A_i)$. We compare this with a marginal variant that replaces the subject-specific nuisance by a pooled curve $S_C(t)$. The conditional pseudo-outcome is

$$Y_i^{\mathrm{res,cond}} = \frac{\widetilde{\Delta}_i(G_i - \widehat{m}_i) + (1 - \widetilde{\Delta}_i)(\rho_{A_i}(\widetilde{Y}_i, X_i) - \widehat{m}_i)}{\widehat{S}_C(\widetilde{Y}_i \mid X_i, W_i, Z_i, A_i)} - \int_0^{\widetilde{Y}_i} \frac{\rho_{A_i}(t, X_i) - \widehat{m}_i}{\widehat{S}_C(t \mid X_i, W_i, Z_i, A_i)}\, d\widehat{\Lambda}_C(t \mid X_i, W_i, Z_i, A_i). \tag{22}$$

The marginal variant replaces $\widehat{S}_C(\cdot \mid X_i, W_i, Z_i, A_i)$ and $d\widehat{\Lambda}_C(\cdot \mid X_i, W_i, Z_i, A_i)$ with $\widehat{S}_C(\cdot)$ and $d\widehat{\Lambda}_C(\cdot)$. Thus, the bridge layer and final forest are unchanged and only the censoring nuisance differs.

Table 9 compares both variants on the common 12×22 benchmark grid, using RMST at $h = 3.0$ and survival probability at the observed-follow-up $q90$ horizon. Conditional and marginal NC-CSF perform nearly identically, while both variants outperform their corresponding baselines. Thus, the main gains come from the negative-control adjustment rather than the censoring model.

Target	Model	RMSE	MAE
RMST ($h = 3.0$)	Conditional NC-CSF	0.0994	0.0768
	Marginal NC-CSF	0.0996	0.0763
	Marginal Baseline	0.1214	0.0937
	Conditional Baseline	0.1662	0.1247
Survival ($q90$)	Conditional NC-CSF	0.0613	0.0491
	Marginal NC-CSF	0.0620	0.0495
	Marginal Baseline	0.0676	0.0562
	Conditional Baseline	0.0788	0.0642

Table 9: Marginal versus conditional censoring on the fixed-horizon benchmark. Lower RMSE and MAE indicate better performance.

We use conditional censoring as the main specification because it preserves subject-level censoring heterogeneity, aligns naturally with the proxy-adjusted nuisance construction, and gives a slight

RMST advantage in our benchmark. Since the gap between marginal and conditional censoring is small relative to the gain from NC-CSF over baseline estimators, we view the censoring specification as a secondary modeling choice rather than the primary driver of performance.

D.10 Bridge and Component Ablation

Table 10 evaluates two final-stage design choices: passing proxy-adjusted raw features to the final learner and augmenting them with nuisance summaries. The comparison uses the common 12×22 benchmark grid, with RMST at $h = 3.0$ and survival probability at the $q90$ horizon. The *NC-CSF (X)* and *NC-CSF (X,W,Z)* variants omit nuisance summaries from the final learner.

Table 10: Final-stage feature ablation on the fixed-horizon benchmark for RMST and survival probability. Lower RMSE and MAE indicate better performance.

Target	Model	RMSE	MAE
RMST ($h = 3.0$)	NC-CSF	0.0994	0.0768
	NC-CSF (X,W,Z)	0.1081	0.0861
	NC-CSF (X)	0.1147	0.0914
	Censored Baseline	0.1662	0.1247
Survival ($q90$)	NC-CSF	0.0613	0.0491
	NC-CSF (X,W,Z)	0.0675	0.0561
	NC-CSF (X)	0.0717	0.0595
	Censored Baseline	0.0788	0.0642

The bridge-adjusted feature set already gives a large gain over the baseline. Compared with the censored baseline, *NC-CSF (X,W,Z)* reduces RMST RMSE from 0.1662 to 0.1081 (35.0%) and RMST MAE from 0.1247 to 0.0861 (30.9%). For survival probability, RMSE decreases from 0.0788 to 0.0675 (14.3%), and MAE decreases from 0.0642 to 0.0561 (12.6%).

Adding nuisance summaries yields a further improvement. Compared with *NC-CSF (X,W,Z)*, the full *NC-CSF* reduces RMST RMSE from 0.1081 to 0.0994 (8.0%) and survival-probability RMSE from 0.0675 to 0.0613 (9.2%). Overall, compared with the censored baseline, the full *NC-CSF* reduces RMST RMSE by 40.2% and survival-probability RMSE by 22.2%.

This isolates the bridge as the primary source of improvement. The forest family is identical across the matched comparisons; RSF-T and RSF-S omit residualization and give RMST RMSEs of 0.1773 and 0.1637, bracketing the 0.1662 of the residualized no-bridge baseline; and conditional versus marginal censoring changes NC-CSF only from 0.0994 to 0.0996 for RMST and from 0.0613 to 0.0620 for survival probability. Thus neither the forest, residualization alone, nor the precise censoring specification explains the bridge-adjusted gain.

Intuitively, raw (X, W, Z) require the final learner to reconstruct the bridge-adjusted signal indirectly, whereas nuisance summaries such as $\widehat{q}$, $\widehat{f}_1$, and $\widehat{f}_0$ provide denoised, low-dimensional features aligned with the orthogonalized residualization step. This makes the final learner more effective at recovering treatment-effect heterogeneity.

D.11 RMST Under the Default $\max(Y)$ Horizon

The main text and fixed-horizon appendix tables evaluate RMST at a common benchmark horizon, such as $h = 3.0$. The original synthetic benchmark also supports a default convention in which, when no explicit RMST horizon is supplied, the target is evaluated at the maximum observed follow-up within each benchmark instance:

$$h_{\text{cell}} = \max_{1 \le i \le n_{\text{cell}}} Y_i.$$

This comparison serves as a robustness check, testing whether NC-CSF maintains its performance advantage when the RMST horizon is not fixed.

The comparison includes the full $12 \times 22 = 264$ case-setting cells for each model. Under the default $\max(Y)$ convention, the realized RMST horizon varies across cells because it is determined by

the maximum observed follow-up in each benchmark instance. We also report the corresponding survival-probability benchmark results from the same output grid for completeness.

Model	RMSE	MAE
Panel A: RMST ($h = \max_i Y_i$)		
NC-CSF Oracle	0.0994	0.0559
NC-CSF	0.1892	0.1155
Baseline	0.3766	0.2609
Panel B: Survival (benchmark quantile horizon)		
NC-CSF Oracle	0.0413	0.0337
NC-CSF	0.0656	0.0556
Baseline	0.0691	0.0587

Table 11: Benchmark results under the default RMST horizon $h = \max_i Y_i$. Lower RMSE and MAE indicate better performance.

Table 11 shows that the main qualitative ranking is preserved under the default $\max(Y)$ RMST convention. For RMST, NC-CSF reduces RMSE from 0.3766 to 0.1892 (49.8%) and MAE from 0.2609 to 0.1155 (55.7%) relative to the baseline. Thus, the advantage of NC-CSF is not specific to the fixed-horizon choice $h = 3.0$.

For survival probability, the gain is smaller but still favorable: NC-CSF reduces RMSE from 0.0691 to 0.0656 (5.0%) and MAE from 0.0587 to 0.0556 (5.2%). We therefore interpret this appendix result as a robustness check: changing the RMST horizon convention from a fixed benchmark horizon to the default $\max(Y)$ rule does not overturn the main conclusion that NC-CSF improves upon the baseline.

The gains are larger for RMST than for horizon-specific survival probability, likely because RMST integrates survival over time and is therefore smoother. In contrast, survival probability depends on a single time point and is more sensitive to local errors in censoring correction and continuation-term estimation.

D.12 Sensitivity to RMST and Survival Horizons

Tables 12 and 13 summarize average performance over all 12×22 case-setting combinations, separately for each horizon. For RMST, we use the four fixed horizons from the CSF synthetic benchmark, $h \in \{1.5, 2.0, 3.0, 15.0\}$. For survival probability, we sweep quantile horizons from $q60$ to $q90$. This comparison serves as a robustness check across time scales rather than a result tied to a single horizon choice.

Horizon	NC-CSF	Baseline	Gain
RMST (fixed horizons)			
$h = 1.5$	0.063	0.077	18.0%
$h = 2.0$	0.077	0.105	26.2%
$h = 3.0$	0.099	0.166	40.2%
$h = 15.0$	0.341	1.016	66.4%
Survival (quantile horizons)			
$q60$	0.066	0.069	5.0%
$q70$	0.068	0.074	7.9%
$q80$	0.068	0.080	15.1%
$q90$	0.061	0.079	22.2%

Table 12: RMSE comparison across horizon choices. Gain denotes the relative reduction from the baseline; lower RMSE is better.

Horizon	NC-CSF	Baseline	Gain
RMST (fixed horizons)			
$h = 1.5$	0.038	0.043	11.0%
$h = 2.0$	0.046	0.058	20.3%
$h = 3.0$	0.055	0.084	34.5%
$h = 15.0$	0.160	0.488	67.2%
Survival (quantile horizons)			
$q60$	0.036	0.036	0.9%
$q70$	0.039	0.041	5.2%
$q80$	0.038	0.043	12.1%
$q90$	0.031	0.037	17.3%

Table 13: Absolute bias comparison across horizon choices. Gain denotes the relative reduction from the baseline; lower absolute bias is better.

The results show that NC-CSF achieves lower RMSE and smaller absolute bias than the baseline at every RMST horizon and every survival horizon. For RMST, the gain increases with the horizon and is largest at $h = 15.0$, suggesting that bridge-adjusted nuisance learning is especially useful when long-horizon targets accumulate more censoring and survival-modeling error.

A similar pattern appears for survival probability. Gains are modest at $q60$ and $q70$ but become larger at $q80$ and $q90$, where censoring and continuation-term estimation are more influential. Overall, the advantage of NC-CSF is not confined to a single horizon choice; it persists across both the RMST horizon set and the survival quantile sweep, with the strongest improvements in more challenging long-horizon regimes.

D.13 Compatibility with the Original CSF Synthetic Benchmark

We additionally evaluate our estimator on the original synthetic simulation benchmark of Cui et al. Cui et al. [2023]. This experiment is distinct from our main proximal benchmark: the original CSF simulation study does not include a latent unobserved confounder U or negative-control proxy variables, and all confounding is induced through the observed covariates X. Thus, this appendix does not test the benefit of proximal adjustment. Instead, it checks whether the final censored learner used in NC-CSF remains stable in a standard fully observed causal survival setting.

We follow the Section 4 simulation protocol of Cui et al. Cui et al. [2023]. The four data-generating settings are the original CSF synthetic settings, with independent covariates drawn from $[0, 1]^p$ and $p = 15$. Each method is trained on $n = 2{,}000$ observations and evaluated on an independent test set of size $n_{\text{test}} = 2{,}000$, over 250 simulation replications. We consider the same two estimands as the original paper: restricted mean survival time (RMST) and survival probability. The RMST horizons are $(1.5, 2.0, 15.0, 3.0)$ for settings 1–4, respectively, and the survival probability horizons are $(0.8, 1.2, 10.0, 2.0)$.

For a fair comparison, we re-run R-CSF locally using the direct `grf` implementation under the same simulation protocol. R-CSF is fit with $2{,}000$ trees, minimum node size 5, subsampling fraction 0.5, and the default split-variable rule $\min(p, \lceil \sqrt{p} \rceil + 20)$. The nuisance survival forests use 500 trees and minimum node size 15, matching the original benchmark. Because this benchmark has no designated negative-control proxies, we evaluate the all-X variant of NC-CSF, where all observed covariates are passed through the observed feature block X. Therefore, this experiment evaluates the compatibility of the NC-CSF final censored learner with the original CSF setting, rather than the benefit of negative-control adjustment.

Setting	Metric	Re-run R-CSF	NC-CSF
Panel A: RMST			
1	MSE	0.27	0.25
	Excess MSE	1.14	1.08
2	MSE	1.15	1.24
	Excess MSE	1.03	1.16
3	MSE	13.37	14.67
	Excess MSE	1.03	1.15
4	MSE	2.97	2.33
	Excess MSE	1.38	1.02
Panel B: Survival probability			
1	MSE	0.16	0.15
	Excess MSE	1.15	1.16
2	MSE	0.37	0.38
	Excess MSE	1.05	1.11
3	MSE	0.15	0.16
	Excess MSE	1.04	1.09
4	MSE	0.28	0.28
	Excess MSE	1.14	1.09

Table 14: Original CSF synthetic benchmark results. MSE is multiplied by 100. Excess MSE is computed relative to the better method in each replication.

Setting	Re-run R-CSF	NC-CSF
Panel A: RMST		
1	0.22	0.22
2	0.17	0.18
3	0.09	0.09
4	0.01	0.03
Panel B: Survival probability		
1	0.22	0.25
2	0.23	0.24
3	0.10	0.11
4	0.04	0.14

Table 15: Classification error on the original CSF synthetic benchmark. Classification error is defined as $1 - (1/n_{\text{test}}) \sum_{i=1}^{n_{\text{test}}} \mathbf{1}\{\text{sign}(\widehat{\tau}(X_i)) = \text{sign}(\tau(X_i))\}$. For Setting 4, classification error is evaluated on the subgroup $X^{(1)} \geq 0.3$, following the original benchmark.

Tables 14 and 15 show that NC-CSF remains broadly competitive with the re-run R-CSF in the fully observed, no-latent-confounding benchmark. This is a demanding comparison for NC-CSF because the benchmark was designed for the original R-CSF framework and contains no negative-control structure. NC-CSF achieves similar MSE in most settings and improves RMST MSE in Settings 1 and 4, while R-CSF remains stronger in some classification-error comparisons, especially survival probability in Setting 4. We therefore interpret this experiment as a compatibility check: NC-CSF does not replace R-CSF in fully observed settings, but its final censored learner remains stable when applied outside the proximal benchmark.

D.14 Detailed Synthetic Setup

Value	Range
$\mathrm{Corr}(Z, U)$	$[0.1,\ 0.9]$
σ_z, σ_w	$[0.73,\ 14.93]$
(γ_u, β_u)	$[0,\ 2]$
$\tau_{\text{log-hr}}$	$[-0.7,\ 0.7]$
model	Linear / Nonlinear
n	$[500,\ 16000]$
d_x	$[2,\ 40]$
$d_z,\ d_w$	$[1,\ 20]$

(a) Parameter ranges

Case	β_u	γ_u	σ_z	σ_w	$\tau_{\text{log-hr}}$	Linear T	Linear Y
C01	0.8	0.8	1.125	1.53	-0.6	Yes	Yes
C02	0.3	0.2	1.125	1.53	-0.25	Yes	Yes
C03	0.9	0.9	7.35	4.77	-0.5	Yes	Yes
C04	0.2	0.3	7.35	4.77	-0.12	Yes	Yes
C05	0.8	0.9	1.125	1.53	-0.7	Yes	No
C06	0.2	0.2	1.125	1.53	-0.3	Yes	No
C07	1.1	0.9	7.35	4.77	-0.5	Yes	No
C08	0.15	0.25	7.35	4.77	-0.12	Yes	No
C09	1.0	1.0	1.125	1.53	-0.7	No	No
C10	0.2	0.2	1.125	1.53	-0.3	No	No
C11	1.1	0.9	7.35	4.77	-0.5	No	No
C12	0.2	0.3	7.35	4.77	-0.12	No	No

(b) Base configurations

Setting	n	d_x	d_w	d_z
S01	500	5	1	1
S02	1000	5	1	1
S03	2000	5	1	1
S04	4000	5	1	1
S05	8000	5	1	1
S06	16000	5	1	1
S07	2000	2	1	1
S08	2000	5	1	1
S09	2000	10	1	1
S10	2000	20	1	1
S11	2000	40	1	1
S12	2000	5	2	2
S13	2000	5	3	3
S14	2000	5	5	5
S15	2000	5	10	10
S16	2000	5	20	20
S17	2000	5	1	10
S18	2000	5	10	1
S19	2000	10	5	5
S20	2000	20	5	5
S21	2000	40	10	10
S22	4000	20	10	10

(c) Dimensional settings

Table 16: Synthetic experiment configuration grid. The left column reports the parameter ranges and the 12 base configurations; the right column reports the 22 dimensional settings. The full benchmark crosses the 12 cases with the 22 settings.

The 12 cases isolate key problem dimensions under fixed proxy loading $a_z = a_w = 1.5$. Cases C01–C04 consider linear treatment and outcome models, varying proxy quality via (σ_z, σ_w) (informative vs. noisy proxies) and confounding strength via (β_u, γ_u) (weak to strong). Cases C05–C08 introduce nonlinear outcome models, which capture misspecification in the survival process. Cases C09–C12 further incorporate nonlinear treatment assignment, representing the most challenging regimes. We focus on beneficial treatment effects (negative log-hazard ratios) to reflect common clinical settings where interventions are expected to improve survival. Reversing the effect direction leaves RMSE and MAE unchanged under the corresponding sign-reversed data-generating mechanism; signed subgroup bias changes sign but not magnitude.

The experimental settings are designed to isolate the effect of different problem dimensions. Settings S01–S06 focus on sample size, varying n from 500 to 16,000 while keeping covariate and proxy dimensions fixed. Settings S07–S11 vary the covariate dimensionality d_x from 2 to 40 with fixed sample size. Settings S12–S16 jointly increase proxy dimensions d_w and d_z, which capture regimes with richer proxy information. Settings S17–S22 are stress-test cases, which explore asymmetric and high-dimensional configurations.

D.15 Additional Results and Graphs for Uncensored Cases

Model	RMSE	MAE
NC-CSF Oracle	0.3347	0.1203
NC-CSF	0.3518	0.1309
Uncensored Baseline	0.3988	0.1861

Table 17: Summary performance in the uncensored setting.

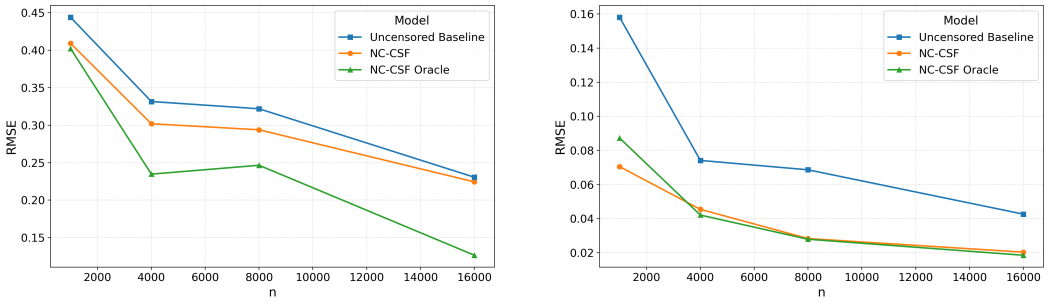

Figure 3: **Left:** RMSE as a function of sample size n under Case 11. **Right:** RMSE as a function of sample size n under Case 12. In both cases, NC-CSF consistently achieves lower RMSE.

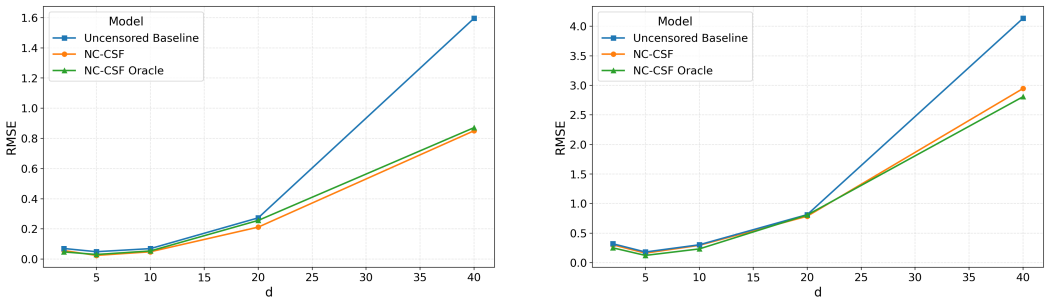

Figure 4: **Left:** RMSE as a function of covariate dimensionality d under Case 10. **Right:** RMSE as a function of d under Case 11. In both cases, NC-CSF scales more robustly with increasing dimensionality d, while the baseline degrades substantially in high-dimensional regimes.

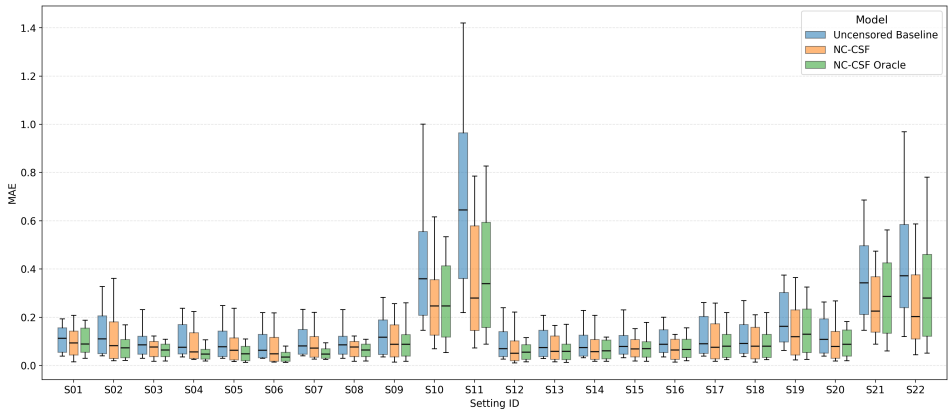

Figure 5: MAE of CATE estimation across settings S01–S22 under uncensoring. While NC-CSF generally improves over the baseline, the gains are modest.

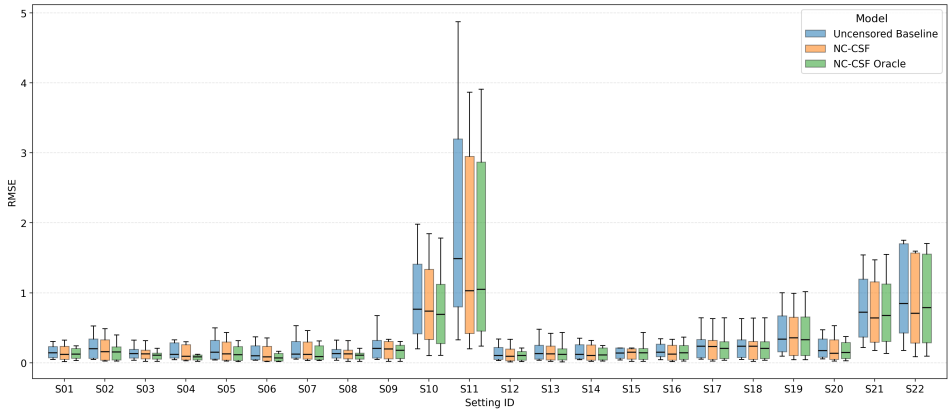

Figure 6: RMSE of CATE estimation across settings S01–S22 under uncensoring. While NC-CSF generally improves over the baseline, the gains are modest in the uncensored setting; as shown later, the advantage becomes substantially more pronounced under censoring.

D.16 Additional Results and Graphs for Censored Cases

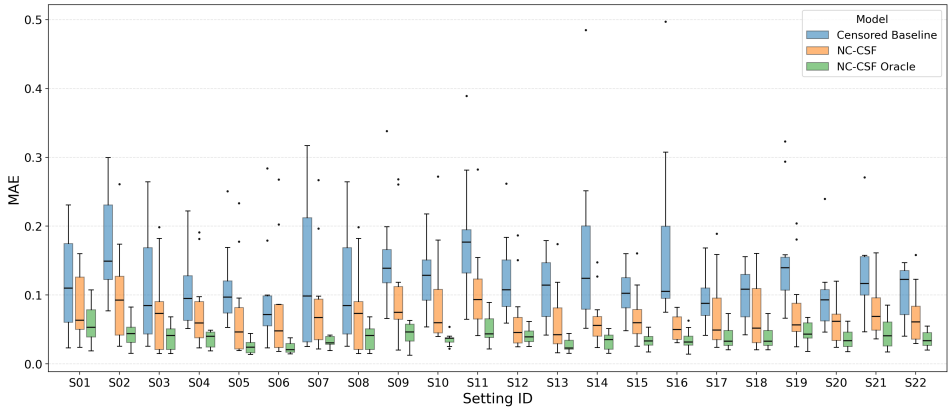

Figure 7: MAE for RMST estimation across settings S01–S22 under censoring. NC-CSF achieves substantially lower MAE than the baseline and remains close to the oracle.

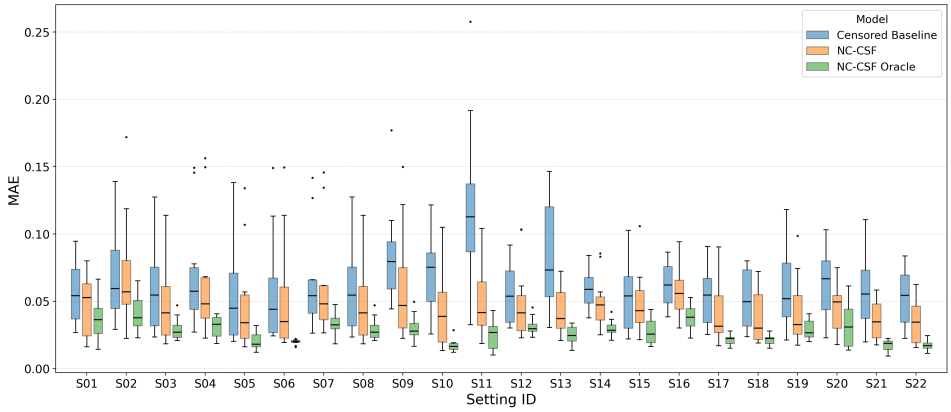

Figure 8: MAE for Survival Probability estimation across settings S01–S22 under censoring. NC-CSF achieves substantially lower MAE than the baseline and remains close to the oracle.

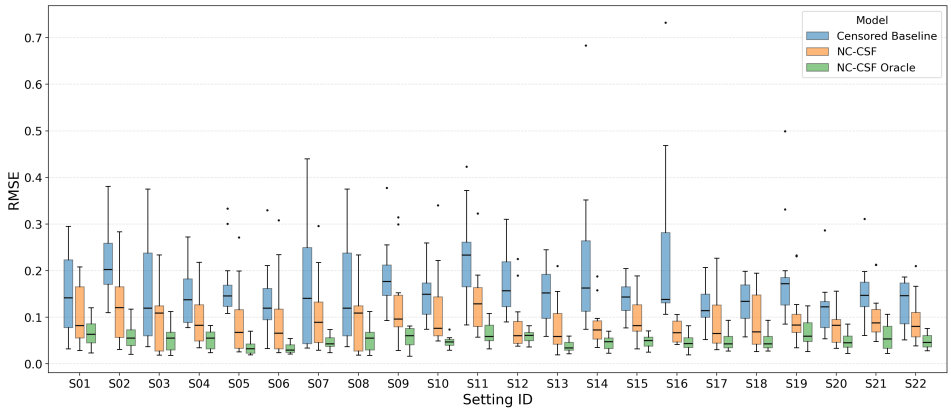

Figure 9: RMSE for RMST estimation across settings S01–S22 under censoring. NC-CSF achieves substantially lower RMSE than the baseline and remains close to the oracle.

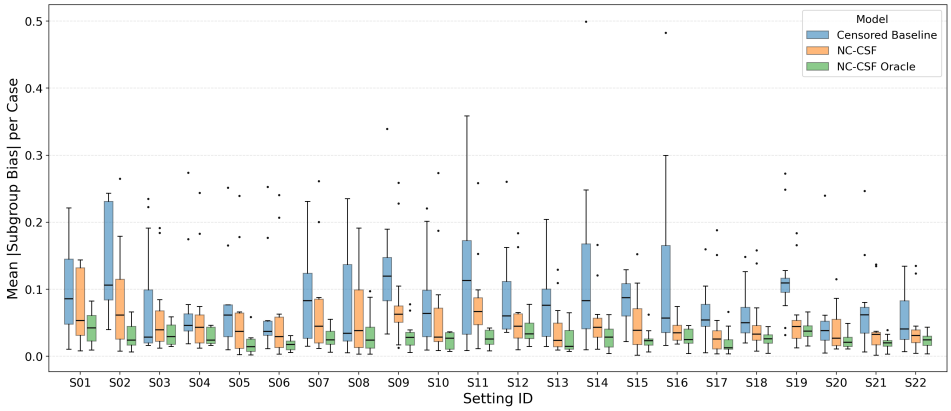

Figure 10: Subgroup bias for RMST estimation across settings S01–S22 under censoring, using X_0-based subgroups. NC-CSF consistently reduces subgroup bias relative to the baseline and remains closer to the oracle, indicating improved heterogeneous treatment effect estimation.

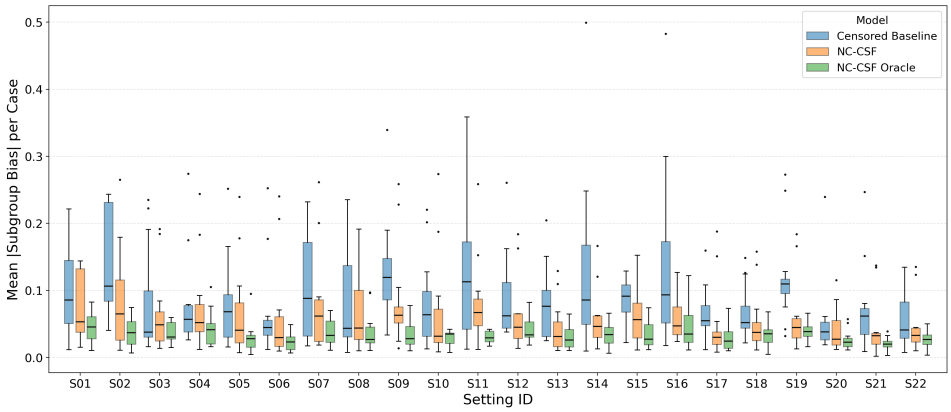

Figure 11: Subgroup bias for RMST estimation across settings S01–S22 under censoring, using X_0X_1-based subgroups. NC-CSF consistently reduces subgroup bias relative to the baseline and remains closer to the oracle, indicating improved heterogeneous treatment effect estimation.

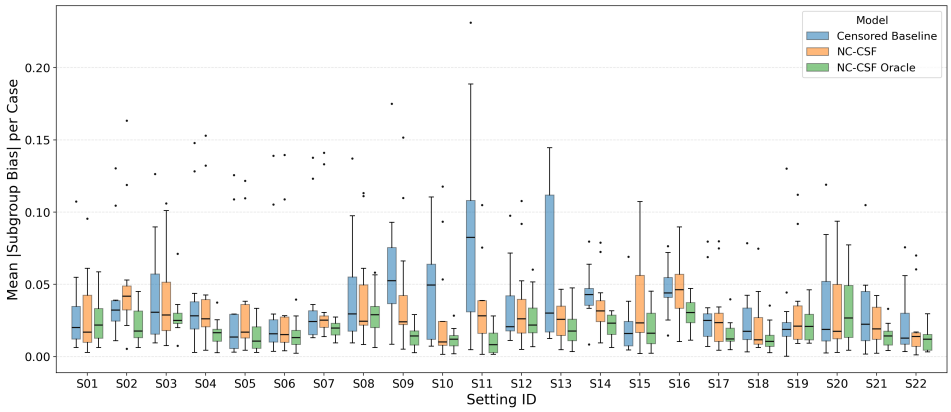

Figure 12: Subgroup bias for Survival Probability estimation across settings S01–S22 under censoring, using X_0-based subgroups. NC-CSF consistently reduces subgroup bias relative to the baseline and remains closer to the oracle, indicating improved heterogeneous treatment effect estimation.

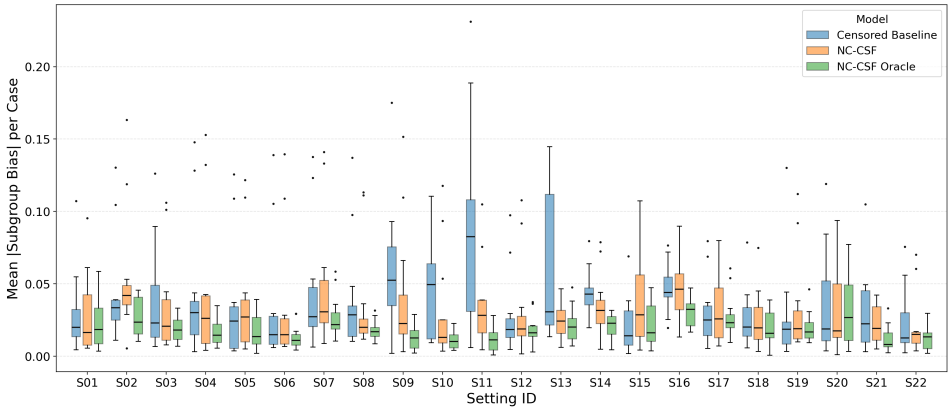

Figure 13: Subgroup bias for Survival Probability estimation across settings S01–S22 under censoring, using X_1-based subgroups. NC-CSF consistently reduces subgroup bias relative to the baseline and remains closer to the oracle, indicating improved heterogeneous treatment effect estimation.

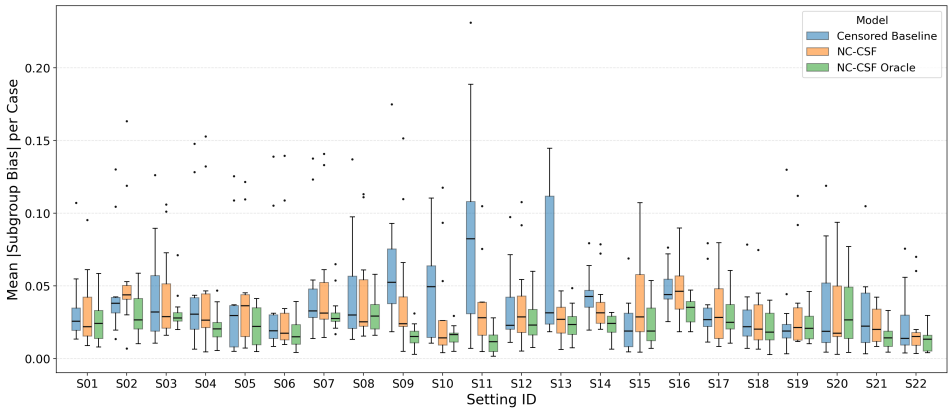

Figure 14: Subgroup bias for Survival Probability estimation across settings S01–S22 under censoring, using X_0X_1-based subgroups. NC-CSF consistently reduces subgroup bias relative to the baseline and remains closer to the oracle, indicating improved heterogeneous treatment effect estimation.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims regarding the proposed NC-CSF method, theoretical guarantees, and empirical evaluation are stated in the abstract and Introduction (Section 1), and are supported by the methodological and theoretical development in Sections 2–3 and the experiments in Sections 4–5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes].

Justification: Section 6 discusses limitations and future work, including interpretation of the weighted target, the need for further theory for recursive forest splitting with estimated nuisances, and the limited scope of real-data validation. The real-data section also states that the ACTG175 heterogeneity finding is exploratory because bootstrap intervals are wide, and that the RHC analysis is observational and lacks ground truth.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions and theoretical results are presented in Section 3, with supporting proofs and derivations provided in the appendix and supplemental material. The paper also states oracle comparisons and local doubly robust properties for the proposed estimator.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The data-generating process, evaluation metrics, and comparison methods are described in Section 4, and implementation details including cross-fitting and nuisance estimation are given in Section 3.3 and Algorithm 1, with full pseudocode provided in Appendix C (Algorithms 3–5).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The anonymized supplementary material includes the code used for the synthetic experiments and real-data analyses, configuration files and scripts for regenerating the reported tables and figures, and instructions for obtaining and preprocessing the public ACTG175 and RHC datasets. The released outputs also include intermediate summaries used to audit the reported results.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training details, data splits, model choices, and nuisance estimation procedures are described in Sections 3.3 and 4.1, with additional implementation details provided in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports uncertainty using bootstrap confidence intervals in the real-data analysis (Figure 5) and evaluates variability across simulation settings (Section 4 and Figures 2–4).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper currently does not provide detailed information about compute resources, runtime, or hardware specifications used for the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work uses synthetic experiments and secondary analyses of public, de-identified clinical datasets; it does not collect new human-subject data, use crowdsourcing, or release personally identifiable information. We cite the original data sources and prior analyses, restrict claims to methodological evaluation and observational clinical association, and discuss limitations relevant to deployment in clinical decision making.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Potential positive impacts related to clinical decision-making and observational healthcare data are discussed in Section 1.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk generative models, scraped datasets, or systems with substantial misuse potential; it studies methodological advances for causal survival analysis using synthetic and authorized clinical data.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: We cite the original sources for the ACTG175 and SUPPORT/RHC clinical datasets and the prior analyses used for comparison, and we do not redistribute restricted raw clinical data outside their applicable access terms. Open-source software dependencies and reused code assets are credited in the supplementary material; released code preserves applicable copyright and license headers.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The implementation code and documentation for reproducing the proposed NC-CSF method and experiments are provided in the anonymized repository.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or prospective research involving human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or prospective human-subject research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as part of the core methodology, experiments, or theoretical development of the proposed NC-CSF framework.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.