Assessing the robustness of heterogeneous treatment effects in survival analysis under informative censoring

Yuxin Wang[*], **Dennis Frauen, Jonas Schweisthal, Maresa Schröder, Stefan Feuerriegel**
LMU Munich
Munich Center of Machine Learning (MCML)

Abstract

Dropout is common in clinical studies, with up to half of patients leaving early due to side effects or other reasons. When dropout is informative (i.e., dependent on survival time), it introduces censoring bias, because of which treatment effect estimates are also biased. In this paper, we propose an assumption-lean framework to assess the robustness of conditional average treatment effect (CATE) estimates in survival analysis when facing censoring bias. Unlike existing works that rely on strong assumptions, such as non-informative censoring, to obtain point estimation, we use partial identification to derive informative bounds on the CATE. Thereby, our framework helps to identify patient subgroups where treatment is effective despite informative censoring. We further propose a novel model-agnostic meta-learner, called **SurvB-learner**, to estimate the bounds that can be used in combination with arbitrary machine-learning models, and that has favorable theoretical properties such as double-robustness and quasi-oracle efficiency. We finally demonstrate the effectiveness of our meta-learner across various experiments using both simulated and real-world data.

1 Introduction

Dropout is widespread in survival analysis, especially in oncology. A recent systematic review found that nearly all cancer studies, both prospective and retrospective, reported dropout, with rates as high as 53% in colorectal cancer and 43% in non–small cell lung cancer (NSCLC) [76]. Patients drop out for many reasons, including severe side effects, personal circumstances, or loss to follow-up [21]. However, due to dropout, survival analyses can be biased [28, 82], making therapies appear more or less effective than they truly are. Here, we present a framework to assess the robustness of heterogeneous treatment effects in survival analysis under dropout.

When dropout is related to survival time (known as *informative censoring*), it introduces systematic bias into survival analysis because event times such as disease-free or overall survival are censored in a non-random way [50, 83]. We refer to this problem as *censoring bias*. Such bias directly undermines the reliable estimation of treatment effects and is especially problematic when studying heterogeneous treatment effects, which aim to identify patient subgroups that respond differently to treatment. Because dropout mechanisms often vary across subgroups, for example, patients with advanced disease may leave due to rapid progression, while others may withdraw due to toxicity [21, 35, 62], censoring bias can affect subgroup-specific estimates even more severely than population-wide effects.

In this paper, we propose an assumption-lean framework to assess the robustness of CATE estimates in survival analysis under informative censoring (see overview figure in Fig. 3 in Appendix A). When censoring depends on survival time, unbiased point estimation of the CATE is generally

[*]Correspondence to: Yuxin.Wang1@lmu.de

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

impossible. Rather than assuming censoring is independent of the survival time, we explicitly account for informative censoring and therefore reframe the task using *partial identification*, which shifts the goal from estimating a single point value to constructing informative lower and upper bounds on the CATE. This allows clinicians to assess whether treatment benefits remain positive for certain subgroups even in the presence of censoring. To achieve this, we build upon the idea of worst-case survival bounds [3, 77, 98], incorporating censoring strength and domain knowledge (e.g., such as expected survival time after dropout from the trial) via sensitivity functions to construct informative bounds.

We further introduce a novel two-stage meta-learner called **SurvB-learner** to efficiently estimate these bounds. Our SurvB-learner is carefully designed to correct for the plug-in bias of naïve learners. Our SurvB-learner is flexible (i.e., model-agnostic) and can be used together with arbitrary machine learning models. Further, it has favorable theoretical properties such as double robustness and quasi-oracle efficiency. Double robustness means that our estimator remains consistent even if one of the nuisance models (e.g., outcome or censoring) is misspecified, while quasi-oracle efficiency guarantees asymptotic performance close to that of an estimator with knowledge of the true nuisance models. Finally, beyond discrete treatments, we extend our framework to continuous treatments (Appendix E.1) and further show that it is applicable to both randomized data from clinical trials as well as observational data from real-world studies with hidden confounding (Appendix E.2).

In this paper, we make three **contributions:**[2] (1) We propose an assumption-lean framework to audit censoring bias in the CATE estimates from a censored dataset. Our method replaces the non-informative censoring assumption with sensitivity functions that use censoring strength and domain knowledge (e.g., expected survival after dropout) to form informative bounds. (2) We further introduce a model-agnostic meta-learner called **SurvB-learner** to efficiently estimate bounds. (3) We provide theoretical results for our meta-learners by showing consistency, double robustness, and quasi-oracle efficiency properties. Finally, we confirm the effectiveness of our meta-learners by performing various experiments using both synthetic and real-world data.

2 Related work

We review the main related literature streams on causal inference with survival outcomes. We refer to Appendix B for details.

CATE estimation under non-informative censoring assumption: Standard methods for survival analysis (e.g., the Kaplan-Meier estimator [45], the Cox proportional hazards model [7, 12], and the accelerated failure time model [12, 85]) rely on the assumption of *non-informative censoring*, meaning that dropout time is independent of survival time. Existing methods for estimating heterogeneous treatment effects from survival data (e.g., including Cox-based models [25, 4], tree-based models [e.g., 13, 30, 80, 96], neural network-based methods [e.g., 46, 72]. and orthogonal meta-learners tailored to censored outcomes [14, 24, 91, 92]) all rely on the assumption of non-informative censoring. $\Rightarrow$ *When the non-informative censoring assumption is violated, estimates obtained from the previous methods will be biased.*

Survival methods for informative censoring: Several methods in survival analysis address the problem of *dependent* censoring. Early works [3, 77, 98] introduced worst-case bounds, which estimate survival functions by assuming that each censored individual experiences the event immediately after their censoring time. While these studies focused on bounding the survival function itself, more recent work has shifted to partially-identify covariate effects on survival time under dependent censoring [70, 88]. $\Rightarrow$ *These studies remain strictly within the survival analysis framework, without any causal inference task.*

Causal survival methods for informative censoring: While the above works address dependent censoring purely within a survival analysis framework, a parallel line of research tackles informative censoring in causal inference settings. Several methods have been proposed to address censoring bias when estimating the average treatment effect (ATE) from censored data [1, 55, 69, 71, 84]. In contrast, clinicians and researchers currently lack practical tools for assessing the robustness of heterogeneous treatment effect estimates when informative censoring is present. $\Rightarrow$ *None of these are directly applicable to CATE estimation.*

[2] Code available at: `https://github.com/Yuxin217/survival_bias`.

Research gap: So far, existing heterogeneous treatment effect learners fail to account for informative censoring. To the best of our knowledge, we are thus the first to provide a partial identification tool that includes the censoring bias to ensure a robust CATE estimation.

3 Problem setup

Data: We consider the standard setting for estimating CATEs based on censored, time-to-event data[3] [13, 15, 24, 83, 96]. That is, we consider the population $(X, A, T, C) \sim \mathbb{P}$, where $X \in \mathcal{X} \subseteq \mathbb{R}^p$ are observed covariates, $A \in \mathcal{A}$ is the treatment[4], $T \in \mathcal{T} \subseteq [0, t_{\max}] \subseteq \mathbb{R}^+$ is the event time of interest (e.g., the time of overall survival (OS) or the time with disease-free survival (DFS)), $C \in \mathcal{T}$ is the censoring time (e.g., the time at which a patient drops out of the study). Importantly, $t_{\max}$ is typically known in medical applications: it is naturally provided by the theoretical maximum human lifespan in general survival analysis, while, in oncology, the maximum survival time is often limited by disease biology. For instance, patients with certain forms of advanced lung cancer rarely survive beyond a few months [37].

The challenge with censored datasets is typically that T is not fully observed, but only $\tilde{T} = \min\{T, C\}$ [7, 12, 45]. Thus, one can not observe data from the full population (X, A, T, C). Instead, we only observe a dataset $\mathcal{D} = \left\{\left(x_i, a_i, \tilde{t}_i, \delta_i\right)\right\}_{i=1}^n$ of size $n \in \mathbb{N}$ sampled i.i.d. from the population $\mathbb{Z} = (X, A, \tilde{T}, \Delta)$, where $\Delta = \mathbb{1}(C \leq T)$ is a censoring indicator for the event and the observed time is given by $\tilde{T} = \min\{T, C\}$. The corresponding causal graph is shown in Fig. 1.

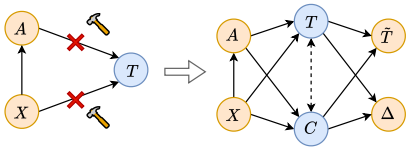

Figure 1: **Causal graph for survival data under informative censoring**. Variables in yellow are observed, while variables in blue are unobserved. Solid edges denote causal relationships, and dashed edges reflect dependence. Note that the absence of arrows encodes causal assumptions, not their presence.

Causal estimand: We make use of the potential outcome framework [68] to formalize our causal inference task. Let $T(a) \in \mathcal{T}$ denote the potential event time under treatment intervention $A = a$. We are interested in the CATE on survival time, i.e.,

$$\tau_{a_1,a_2}(x) = \tau_{a_1}(x) - \tau_{a_2}(x) = \mathbb{E}_{\mathbb{P}}\left[T(a_1) - T(a_2) \mid X = x\right]. \tag{1}$$

Hence, our estimand $\tau_{a_1,a_2}(x)$ measures the difference of conditional means between two treatments for a patient with covariates $X = x$. For completeness, we define the conditional average potential outcome (CAPO) for survival time via $\tau_a(x) = \mathbb{E}\left[T(a) \mid X = x\right]$.

Key definitions: We define the *propensity score* for treatment a as $\pi_a(x) = \mathbb{P}(A = a \mid X = x)$, for $a \in \mathcal{A}$, which captures the treatment assignment mechanism. Further, we denote the *censoring strength* as $\xi(x, a) = \mathbb{P}(\Delta = 1 \mid X = x, A = a)$, which gives the probability that the event is censored, meaning that a patient does not complete the study based on covariates $X = x$ and treatment $A = a$. We further denote the *expected survival time function* by $\mu(x, a) = \mathbb{E}[T \mid X = x, A = a]$ and the *expected conditional survival time function* by $\nu(\delta, x, a) = \mathbb{E}[\tilde{T} \mid \Delta = \delta, X = x, A = a]$. Here, $\mu(x, a)$ is the expected survival time given covariates $X = x$ and treatment $A = a$, while $\nu(\delta, x, a)$ is the expected survival time further conditioned on the censoring indicator $\Delta = \delta$. Finally, we define the *post-dropout survival function* $\gamma(x, a)$, which denotes the expected remaining survival time after dropout, given covariates $X = x$ and treatment $A = a$. Note that the post-dropout survival function serves as a sensitivity function in our method; we treat it as *known* rather than estimating it from the observed data.

Identifiability: We make use of the following standard Assumption 3.1.

Assumption 3.1. For all $x \in \mathcal{X}$, $a \in \mathcal{A}$, it holds: (i) Consistency: $A = a \Rightarrow \tilde{T} = \min\{T, C\} = \min\{T(a), C(a)\}$; (ii) Treatment overlap: $0 < \pi_a(x) < 1, \forall x \in \mathcal{X}$, and $\forall a \in \mathcal{A}$; (iii) Unconfoundedness: $T(a) \perp\!\!\!\perp A \mid X$; (iv) Censoring overlap: $0 \leq \xi(x, a) < 1, \forall x \in \mathcal{X}$, and $\forall a \in \mathcal{A}$.

Assumptions (i)–(iii) are standard in the causal inference literature and are widely used for estimating CATEs [9, 15, 38, 68, 75]. (i) Consistency ensures that an individual's observed outcome under

[3] We deal with the problem of right censoring, which is common in survival analysis settings. We thus assume that events have not happened before time $t = 0$.

[4] Our method is applicable to both discrete and continuous treatment settings. For simplicity, we focus on discrete treatments (i.e., $A \in \mathcal{A} \subseteq \mathbb{N}$) in the main paper. We provide an extension to continuous treatments in Appendix E.1.

treatment a equals their potential outcome $T(a)$; it also implies no interference between individuals. (ii) Treatment overlap means that every patient has a positive probability of receiving each treatment, which thereby ensures sufficient support in the data. (iii) Unconfoundedness assumes that there are no unobserved confounders, which implies that, in the uncensored subgroup, the expected conditional survival time function $\mathbb{E}[T(a) \mid X = x]$ is identifiable. In the following, we focus on clinical trials where unconfoundedness is ensured by design and real-world data without hidden confounding. We later also provide an extension to real-world data with hidden confounding in Appendix E.2. (iv) Censoring overlap is additionally common in survival analysis [8, 86] and ensures a positive probability of being uncensored (i.e., experiencing the event before censoring) every covariate value.

3.1 Challenges from informative censoring

Estimating heterogeneous treatment effects in survival analysis is challenging because standard assumptions required for valid causal inference are often violated. In particular, most of the causal survival literature [15, 24, 25, 83, 91, 92] assumes non-informative censoring. While this allows for identification of the CATE, it rarely holds in medical practice. In such cases, censoring carries prognostic information, which, if ignored, leads to biased estimates. Below, we formalize the challenges from informative censoring and thereby motivate the need for our framework later.

Why are the standard assumptions from the survival literature [15, 24, 83] violated? These works typically make the assumption of non-informative censoring to identify the CATE $\tau_{a_1,a_2}(x)$ from censored data, which can be stated as

$$T \perp\!\!\!\perp C \mid X = x, A = a, \qquad \forall x \in \mathcal{X}, a \in \mathcal{A} \tag{2}$$

This assumption requires that the censoring time is independent of the patient's survival time, conditional on covariates and treatment. Under this assumption, the expected conditional survival time function among the censored patients, i.e., $\mathbb{E}[T \mid X = x, A = a, \Delta = 1]$, is identifiable. However, in practice, this assumption is often violated. For example, patients with a more severe disease (and thus shorter survival times T) may drop out earlier, for example, when transferred to palliative care or other facilities. Similarly, patients may be more likely to be lost to follow-up as death approaches, for instance, by withdrawing from the clinical trial [32, 82]. In these cases, the expected conditional potential survival function under censoring, $\mathbb{E}[T(a) \mid X = x, \Delta = 1]$, is *not* identified. Therefore, the causal estimand, the CATE, is also *not* identified as stated in the following lemma.

Lemma 3.2. *Let the* expected conditional survival function $\nu(\delta, x, a)$ *and the* censoring strength $\xi(x, a)$ *be defined as above. Under informative censoring, the true survival time is unobserved, and, therefore, the CATE is not identified; i.e.,*

$$\begin{aligned}\tau_{a_1,a_2}(x) = \nu(0, x, a_1)[1 - \xi(x, a_1)] + \mathbb{E}[T \mid X = x, A = a_1, \Delta = 1]\xi(x, a_1) \\ - \nu(0, x, a_2)[1 - \xi(x, a_2)] - \mathbb{E}[T \mid X = x, A = a_2, \Delta = 1]\xi(x, a_2).\end{aligned} \tag{3}$$

Proof. See Appendix D.1. □

Because these outcomes for the equations in blue color are never observed by definition, they cannot be recovered from the data without further assumptions. Hence, even with an infinitely large sample, the true values of these expectations, and therefore of the CATE, cannot be determined from the observed data distribution alone. As a result, unbiased point estimation of the CATE is impossible under informative censoring.

Therefore, we relax the assumption of non-informative censoring from Eq. 2 and adopt the approach of *partial identification*: instead of targeting a single point estimate, we aim to construct informative lower and upper bounds on the CATE. To achieve this, we build on the idea of worst-case survival bounds [3, 77, 98], but where we incorporate both censoring strength and domain knowledge (e.g., expected survival time after dropout) via sensitivity functions. This approach is beneficial in practice because we explicitly account for censoring bias while still providing clinically meaningful insights. For instance, the resulting bounds allow clinicians to determine whether the effect of treatment is positive for certain subgroups even in the presence of censoring, which implies a *robust benefit* from treatment.

4 Our approach: partial identification of the CATE under informative censoring

To address the above challenges, we now move away from point estimation to partial identification of the CATE, which allows us to obtain informative bounds in the presence of informative censoring. To

do so, we first notice that, as stated in Lemma 3.2, the key challenge for estimating the CATE lies in the expected conditional potential survival function under censoring, i.e., $\mathbb{E}[T(a) \mid X = x, \Delta = 1]$. As a remedy, our approach is therefore to first derive partial identification bounds for this quantity, which then yield bounds for CAPO $\tau_a(x)$. We can then naturally obtain bounds for CATE $\tau_{a_1,a_2}(x)$, as the CATE is defined as the difference between CAPOs, i.e., $\tau_{a_1}(x)$ and $\tau_{a_2}(x)$. Our bounds, which we derive below, build on Manski bounds [54], but we carefully adapt to our survival setting. In that sense, our bounds follow the idea of worst-case bounds [3, 77, 98] in survival analysis. Importantly, unlike standard Manski bounds [54], which assume the only source of missing data is treatment assignment, our survival setting introduces an additional source of partial observability through informative censoring, which makes partial identification in our setting non-trivial.

Next, we formally define the lower and upper bounds for the CAPO, which we denote by $\mu^-(x, a)$ and $\mu^+(x, a)$, respectively. The bounds thus capture the range of plausible values under our partial identification framework while allowing for informative censoring.

Lower bound: To construct a lower bound for the CAPO, we leverage the definition of $\tilde{T} = \min\{C, T\}$. By design, we have $T \geq \tilde{T}$ and $\mathbb{E}[T(a) \mid X = x, \Delta = 1] \geq \mathbb{E}[\tilde{T} \mid X = x, A = a, \Delta = 1]$ when conditioned on the censored events (i.e., $\Delta = 1$). We then replace T with $\tilde{T}$ in Eq. (3) and obtain a lower bound $\mu^-(x, a)$ as follows:

$$\begin{aligned}\mu(x, a) &\geq \nu(0, x, a)[1 - \xi(x, a)] + \mathbb{E}[\tilde{T} \mid \Delta = 1, X = x, A = a]\xi(x, a) \\ &= \mathbb{E}[\tilde{T} \mid X = x, A = a] := \mu^-(x, a). \end{aligned} \tag{4}$$

Upper bound: To construct an upper bound for the CAPO, we introduce $\gamma(x, a)$ as a sensitivity function, stemming from domain knowledge on the maximum survival time after dropout to construct informative bounds. Unlike many other applications of partial identification, our sensitivity function has a natural interpretation: it simply captures the maximum possible expected survival time a patient may have after censoring (we discuss how to specify this in clinical practice below).

Formally, we consider sensitivity function $\gamma(x, a)$ which satisfies

$$\mathbb{E}[T - \tilde{T} \mid X = x, A = a, \Delta = 1] \leq \gamma(x, a) \leq t_{\max} - \underbrace{\mathbb{E}[\tilde{T} \mid X = x, A = a, \Delta = 1]}_{:=\nu(1,x,a)}, \tag{5}$$

for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$. The left-hand side $\mathbb{E}[T - \tilde{T} \mid X = x, A = a, \Delta = 1]$ reflects the expected remaining survival time beyond the observed follow-up, while the right-hand site $t_{\max} - \nu(1, x, a)$ reflects the difference between the theoretical maximum lifespan and the observed follow-up. Depending on the setting, $\gamma(x, a)$ can incorporate domain knowledge (when specific information about post-dropout survival is available) or be set to $t_{\max} - \nu(1, x, a)$ in the absence of such information. Importantly, the latter ensures that we can still obtain informative bounds even when domain knowledge is not available.

This setting is realistic in many clinical and oncology applications, where prior information about post-dropout survival can be obtained from historical trials [61], disease staging systems [17], and established clinical benchmarks [93], such as 3-year DFS or 5-year OS. For example, in advanced lung cancer or pancreatic cancer, survival beyond a few months or years is exceedingly rare [34, 39, 61]. In this sense, $\gamma(x, a)$ serves as a sensitivity function: a user-specified input that encodes clinically reasonable assumptions about post-dropout survival and allows researchers to assess how conclusions change under those assumptions. Importantly, $\gamma(x, a)$ may vary across individuals with different covariates, such as baseline age, lifestyle factors, or comorbidities, thereby allowing patient-specific heterogeneity to be incorporated.

Depending on the availability of domain knowledge, we consider two ways of specifying upper bounds on the CAPOs. **Case** ① : when clinical knowledge about post-dropout survival is available for different covariates and treatments, one can set $\gamma(x, a)$ to a clinically plausible maximum post-dropout survival time and use it to construct an upper bound on the CAPOs. **Case** ② : when such domain knowledge is unavailable, one can instead set $\gamma(x, a) = t_{\max} - \nu(1, x, a)$ to obtain a conservative upper bound on the CAPOs. This choice makes no strong clinical assumptions about post-dropout survival while ensuring that the resulting bounds remain informative.

Upper bounds for post-dropout survival. We consider two cases:

• **Case** ① **(domain-knowledge upper bounds):** In the first case, when domain knowledge about $\gamma(x, a)$ is available, this information can be directly used as a sensitivity function to construct

informative upper bounds for the CAPO. For example, in some lung cancer trials [37], patients typically survive no longer than two to five months after dropping out, which suggests setting $\gamma(x, a) = 3$. The resulting upper bound is

$$\mu(x, a) \leq \nu(0, x, a)[1 - \xi(x, a)] + \nu(1, x, a)\xi(x, a) + \gamma(x, a)\xi(x, a) := \mu^+(x, a), \tag{6}$$

which is expressed as a weighted combination of the expected conditional survival time without censoring $\nu(0, x, a)$, the expected conditional survival time with censoring $\nu(1, x, a)$, and the post-dropout survival captured by $\gamma(x, a)$.

• **Case ② (conservative upper bounds):** In the second case, domain knowledge about post-dropout survival is unavailable—for example, due to relocation, withdrawal of consent, or complete loss to follow-up. In such situations, a conservative sensitivity function is used. Specifically, we take the upper bound of $\gamma(x, a)$ and set

$$\gamma(x, a) = t_{\max} - \nu(1, x, a) \tag{7}$$

where $t_{\max}$ is the maximum support of the distribution of $\tilde{T}$. Substituting this into Eq. (3) yields the conservative upper bound

$$\mu(x, a) \leq \nu(0, x, a)[1 - \xi(x, a)] + t_{\max}\xi(x, a) := \mu^+(x, a). \tag{8}$$

5 Property of the partial identification bounds for CATE

We now present our main result: our derivation of the partial identification bounds for CATE $\tau^-_{a_1,a_2}(x)$ and $\tau^+_{a_1,a_2}(x)$, which characterize the range of the CATE under informative censoring.

Theorem 5.1. *Under assumptions 3.1, the CATE is bounded via*

$$\tau^-_{a_1,a_2}(x) \leq \tau_{a_1,a_2}(x) \leq \tau^+_{a_1,a_2}(x), \tag{9}$$

where $\tau^+_{a_1,a_2}(x) = \mu^+(x, a_1) - \mu^-(x, a_2)$ and $\tau^-_{a_1,a_2}(x) = \mu^-(x, a_1) - \mu^+(x, a_2)$. Here, $\mu^-(x, a_1)$ and $\mu^-(x, a_2)$ are given by Eq. (4), and $\mu^+(x, a_1)$ and $\mu^+(x, a_2)$ are given by Eq. (6) or Eq. (8) depending on the choice of $\gamma(x, a)$.

The width of the bounds is an important property, as tighter bounds convey more information. Since the bounds on the CATE are derived from those on the CAPO, we study the theoretical properties of the CAPO bounds. In the following proposition, we show that the tightness of the bounds is determined by the censoring strength in the data and the post-dropout survival time function.

Proposition 5.2. *Let $\mu^-(x, a)$ and $\mu^+(x, a)$ be the lower bound and upper bound of the CAPO. The width of the bound for the CAPO is then*

$$\mu^+(x, a) - \mu^-(x, a) = \gamma(x, a)\xi(x, a), \tag{10}$$

where $\gamma(x, a)$ is the known sensitivity function, and where the censoring strength $\xi(x, a)$ is the probability of patients being censored given covariates and treatments. The bounds are informative under either of the following conditions: (a) when strong domain knowledge removes uncertainty about post-dropout survival; or (b) when $\xi(x, a) = 0$, which corresponds to the absence of censoring.

Proof. See Appendix D.3. □

The above proposition characterizes when our partial identification bounds become informative. Hence, our bounds are particularly well-suited for settings with low or moderate censoring: when censoring is low, the width of bounds remains tight even if $\gamma(x, a)$ is misspecified; when censoring is high, but domain knowledge is available to construct $\gamma(x, a)$, the bounds remain informative. Hence, this ensures that our framework provides valid and interpretable results under informative censoring, thereby helping clinicians identify patient subgroups that reliably benefit from treatment without relying on potentially biased survival models.

6 SurvB-learner: A meta-learner for estimating the bounds

We now develop a two-stage meta-learner, called SurvB-learner, for estimating bounds from Theorem 5.1. For simplicity, we first derive the two-stage meta-learner for the CAPOs; the corresponding bounds for the CATE can be obtained directly by taking the difference between two CAPOs (see Appendix C for details). Importantly, our two-stage meta-learner is flexible and can be instantiated with arbitrary machine learning methods (including, e.g., linear models or neural networks).

Why the simple "plug-in" learner is problematic: A straightforward approach to estimating the bounds is the *plug-in learner*. In this approach, one first estimates the nuisance functions $\hat{\nu}(\delta, x, a)$, $\hat{\xi}(x, a)$, and $\hat{\pi}_a(x)$, which correspond to $\nu(\delta, x, a)$, $\xi(x, a)$, and $\pi_a(x)$ for all $\delta \in \{0, 1\}$, $a \in \mathcal{A}$, $x \in \mathcal{X}$. Then, one can directly "plug in" the estimated nuisance functions into Eq. (4), Eq. (6), and Eq. (8) to obtain the bounds. However, while straightforward, this approach is known to suffer from plug-in bias [48], which limits statistical efficiency and suboptimal estimation performance.

To address the drawbacks of the plug-in learner, we propose a *two-stage meta-learner*, which, as we show later, is doubly robust and quasi-oracle efficient. To achieve this, we borrow the general principles of two-stage learners [51, 58]: first, to construct pseudo-outcomes that are, in expectation, equal to the target estimand; and, second, to use the pseudo-outcomes in a second-stage regression to estimate the target quantity, which, in our case, are $\mu^{\pm}(x, a)$. We now adapt it to our setting.

6.1 Overview of the SurvB-learner

At a high level, our SurvB-learner proceeds in two stages (see the pseudocode in Algorithm 1). (1) We estimate the nuisance functions with cross-fitting, where any suitable machine learning model can be applied, and construct pseudo-outcomes $\hat{\phi}^{\pm}(x, a)$. (2) We regress the estimated pseudo-outcomes on the covariates $X = x$, which yields a function to estimate the CATE bound. In the following, we show how the pseudo-outcomes are constructed (additional derivations are provided in Appendix D.4).

Lower bound: The pseudo-outcome for the lower bound is defined by

$$\begin{aligned}
&\hat{\phi}^{-}(x, a) \qquad (11)\\
&= \frac{1(A = a, \Delta = 0)}{\hat{\pi}_a(x)}\{\tilde{T} - \hat{\nu}(0, x, a)\} + \hat{\nu}(0, x, a)[1 - \hat{\xi}(x, a)] + \frac{\hat{\nu}(0, x, a)1(A = a)}{\hat{\pi}_a(x)}\{1(\Delta = 0) - [1 - \hat{\xi}(x, a)]\}\\
&+ \frac{1(A = a, \Delta = 1)}{\hat{\pi}_a(x)}\{\tilde{T} - \hat{\nu}(1, x, a)\} + \hat{\nu}(1, x, a)\hat{\xi}(x, a) + \frac{\hat{\nu}(1, x, a)1(A = a)}{\hat{\pi}_a(x)}\{1(\Delta = 1) - \hat{\xi}(x, a)\}.
\end{aligned}$$

Upper bound: For the pseudo-outcome of the upper bound, we again distinguish two cases as above. These correspond to having a sensitivity function $\gamma(x, a)$ with domain knowledge in **Case** ①, and a conservative upper bound in **Case** ②.

• **Case** ① (*domain-knowledge upper bounds).* In this case, the upper bound for a given $\gamma(x, a)$ is stated in Eq. (6). The corresponding pseudo-outcome is then given by

$$\begin{aligned}
&\hat{\phi}^{+}(x, a) \qquad (12)\\
&= \frac{1(A = a, \Delta = 0)}{\hat{\pi}_a(x)}\{\tilde{T} - \hat{\nu}(0, x, a)\} + \hat{\nu}(0, x, a)[1 - \hat{\xi}(x, a)] + \frac{\hat{\nu}(0, x, a)1(A = a)}{\hat{\pi}_a(x)}\{1(\Delta = 0) - [1 - \hat{\xi}(x, a)]\}\\
&+ \frac{1(A = a, \Delta = 1)}{\hat{\pi}_a(x)}\{\tilde{T} - \hat{\nu}(1, x, a)\} + \hat{\nu}(1, x, a)\hat{\xi}(x, a) + \frac{\hat{\nu}(1, x, a)1(A = a)}{\hat{\pi}_a(x)}\{1(\Delta = 1) - \hat{\xi}(x, a)\}\\
&+ \gamma(x, a)\frac{1(A = a)}{\hat{\pi}_a(x)}\left\{1(\Delta = 1) - \hat{\xi}(x, a)\right\} + \gamma(x, a)\hat{\xi}(x, a).
\end{aligned}$$

• **Case** ② (*conservative upper bounds).* Recall that the conservative upper bound is a special case of $\gamma(x, a) = t_{\max} - \nu(1, x, a)$, as defined in Eq. (8). The corresponding pseudo-outcome is given by

$$\begin{aligned}
&\hat{\phi}^{+}(x, a) \qquad (13)\\
&= \frac{1(A = a, \Delta = 0)}{\hat{\pi}_a(x)}\{\tilde{T} - \hat{\nu}(0, x, a)\} + \hat{\nu}(0, x, a)[1 - \hat{\xi}(x, a)] + \frac{\hat{\nu}(0, x, a)1(A = a)}{\hat{\pi}_a(x)}\{1(\Delta = 0) - [1 - \hat{\xi}(x, a)]\}\\
&+ t_{\max}\frac{1(A = a)}{\hat{\pi}_a(x)}\left\{1(\Delta = 1) - \hat{\xi}(x, a)\right\} + t_{\max}\hat{\xi}(x, a).
\end{aligned}$$

The full estimation procedure is stated in Algorithm 1.

6.2 Theoretical properties

Below, we derive several favorable theoretical properties for our SurvB-learner, including consistency, double robustness, and quasi-oracle efficiency.

Theorem 6.1 (Consistency and double robustness)**.** *Let $\hat{\nu}(\delta, x, a)$, $\hat{\xi}(x, a)$, and $\hat{\pi}_a(x)$, denote the estimators of $\nu(\delta, x, a)$, $\xi(x, a)$, and $\pi_a(x)$, respectively. Then, for all $x \in \mathcal{X}$, we have that $\mathbb{E}\left[\hat{\phi}^{+}(x, a) \mid X = x\right] = \mu^{+}(x, a)$ and $\mathbb{E}\left[\hat{\phi}^{-}(x, a) \mid X = x\right] = \mu^{-}(x, a)$ if at least one of the following conditions holds: (1) $\hat{\pi}_a(x) = \pi_a(x)$, or (2) $\hat{\nu}(\delta, x, a) = \nu(\delta, x, a)$ and $\hat{\xi}(x, a) = \xi(x, a)$.*

Proof. See Appendix D.5. □

Both consistency and double robustness are general properties of our SurvB-learner, regardless of the underlying machine learning model. Consistency means that, under correct specification of the nuisance components, our estimator converges to the true target quantity as the sample size grows. Double robustness means that the bounds remain consistent if either the propensity score function $\pi_a(x)$ is correctly specified, or if the combination of the expected conditional survival time function $\nu(\delta, x, a)$ and the censoring strength $\xi(x, a)$ are correctly specified. Hence, our SurvB-learner is robust against misspecification in one set of nuisance components. Importantly, the estimated CATE bounds also inherit double robustness, since they are constructed from the CAPO bounds. This property parallels the double robustness of the DR-learner in the standard CATE setting [47], but here we extend it to partial identification under informative censoring.

We further derive an asymptotic bound on the convergence rate of SurvB-learner under standard smoothness assumptions. For this, we consider s-smooth functions from the Hölder class $\mathcal{H}(s)$, which achieve Stone's minimax rate [79] of $n^{\frac{-2s}{2s+p}}$, where p is the dimension of $\mathcal{X}$.

Assumption 6.2 (Smoothness)**.** We assume that (1) the nuisance component $\nu(\delta, x, a)$ is α-smooth, $\xi(x, a)$ is β-smooth, and $\pi_a(x)$ is ζ-smooth; (2) all nuisance components are estimated with their respective minimax rate of $n^{\frac{-2k}{2k+p}}$, where $k \in \{\alpha, \beta, \zeta\}$; and (3) the oracle CAPO $\tau_a(\cdot)$ and the oracle CATE $\tau(\cdot)$ are η-smooth and that the initial CATE estimator $\hat{\tau}(x)$ converges with rate $r_\tau(n)$.

Assumption 6.3 (Boundedness)**.** We assume that there exist constants C, ε, $L > 0$ such that, for all $x \in \mathcal{X}$, it holds that: (1) $\mid \nu(\delta, x, a) \mid \leq C$, (2) $\varepsilon < \hat{\pi}_a(x) < 1 - \varepsilon$; and (3) $\mid \hat{\mu}^{\pm}(x, a) \mid \leq L$.

Assumptions 6.2 and 6.3 are standard in the literature and in line with previous works on the theoretical analysis of CATE point estimators [14, 47] and estimators for partial identification bounds [59]. Assumption 6.2 provides a strategy to quantify the difficulty of the underlying nonparametric regression problems through smoothness conditions, while Assumption 6.3 ensures that both the oracle bounds for the CATE and estimators are bounded. Based on this, we state the quasi-oracle efficiency of our SurvB-learner.

Theorem 6.4 (Quasi-oracle efficiency)**.** *Let $\mathcal{D}_l$ for $l \in \{1, 2, 3\}$ be independent samples of size n. Let $\hat{\nu}(\delta, x, a)$, $\hat{\xi}(x, a)$ be trained on $\mathcal{D}_1$, and $\hat{\pi}_a(x)$ be trained on $\mathcal{D}_2$. We denote $\hat{\phi}^{\pm}(x, a)$ as the pseudo-outcome and $\hat{\mu}^{\pm}(x, a) = \hat{\mathbb{E}}_n[\hat{\phi}^{\pm}(x, a) \mid X = x]$ as the pseudo-outcome regression on $\mathcal{D}_3$ for some generic estimator $\hat{\mathbb{E}}_n[\cdot \mid X = x]$ of $\mathbb{E}_n[\cdot \mid X = x]$. Suppose that the second-stage estimator $\hat{\mathbb{E}}_n$ achieves the minimax rate $n^{\frac{-2\eta}{2\eta+p}}$ and satisfies the stability assumption from Kennedy [47]. Under the assumptions of smoothness, the oracle risk has the following upper bound:*

$$\mathbb{E}\left[(\hat{\mu}^{\pm}(x, a) - \mu_a^{\pm}(x))^2\right] \lesssim n^{\frac{-2\eta}{2\eta+p}} + n^{\left(\frac{-2\zeta}{2\zeta+p} - \frac{2\beta}{2\beta+p}\right)} + n^{\left(\frac{-2\zeta}{2\zeta+p} - \frac{2\alpha}{2\alpha+p}\right)}. \tag{14}$$

Proof. See Appendix D.6. In the proof, we even provide a more general bound, which depends on the pointwise mean-squared errors of the nuisance parameters estimators. □

The above theorem gives an upper bound on the risk of the SurvB-learner in comparison to the theoretical optimal convergence rate that would be achieved if knowing the oracle nuisance function. Hence, quasi-oracle efficiency means that the SurvB-learner asymptotically achieves the same convergence rate as if the ground-truth nuisance functions were known. Eq. (14) illustrates this property by shown that, even if the propensity score is highly complex and its estimator converges slowly, our SurvB-learner still converges quickly as long as the other nuisance estimators (i.e. the combination of expected conditional survival time function $\nu(\delta, x, a)$ and censoring strength $\xi(x, a)$) converge sufficiently fast. To derive the above bound, we leverage the cross-fitting approach from Kennedy [47], which was initially used to analyze the DR-learner for CATE estimation, where it allowed for the derivation of robust convergence rates. This technique has since been adapted to other meta-learners (e.g., [14]), but, notably, not to partial identification in the causal survival setting.

7 Experiments

Further, we evaluate the effectiveness of the bounds generated by our SurvB-learner through experiments on synthetic datasets (we provide the results of real-world dataset in Appendix I). Synthetic datasets are commonly used to evaluate causal inference methods [83, 24, 15], since they have the advantage of providing access to the ground-truth CATE and thus allow for direct comparison against oracle estimates. We additionally benchmark against standard *point estimator* that assumes non-informative censoring in Appendix G.1.

Data: We simulate samples from several scenarios with different underlying functions and censoring probabilities. Here, we consider several scenarios where we vary the censoring probability across datasets (e.g., $\xi = 0.2$, 0.4, or 0.6). Details of the data-generating mechanisms are provided in Appendix F. Altogether, we generate over 60 different datasets under varying scenarios. We then compare our SurvB-learner against the oracle CATE and oracle bounds calculated using the ground-truth nuisance estimators. We evaluate our SurvB-learner for both **Case** ① with domain-knowledge bounds and **Case** ② with conservative bounds.

In our experiments, we employ the random forest [6] for nuisance estimation in both the plug-in learner and our SurvB-learner, which ensures that performance differences arise solely from the meta-learners rather than from the underlying model choice. We apply three-fold cross-fitting for our SurvB-learner and use the random forest regressors with default hyperparameter settings in the second stage regression. For more implementation details, we refer to Appendix H.

Bound Type	Censoring strength	① Domain knowledge		② Conservative	
		Plug-in learner	SurvB-learner	Plug-in learner	SurvB-learner
Exponential function	0.2	3.244 ± 0.152	$\mathbf{2.541 \pm 0.138}$	2.120 ± 0.994	$\mathbf{1.477 \pm 1.006}$
	0.4	4.265 ± 0.129	$\mathbf{3.429 \pm 0.071}$	2.572 ± 1.571	$\mathbf{2.006 \pm 1.396}$
	0.6	5.133 ± 0.120	$\mathbf{3.995 \pm 0.066}$	2.969 ± 2.008	$\mathbf{2.305 \pm 1.610}$
Sin function	0.2	1.400 ± 0.067	$\mathbf{0.799 \pm 0.020}$	1.104 ± 0.140	$\mathbf{0.576 \pm 0.114}$
	0.4	1.602 ± 0.095	$\mathbf{1.064 \pm 0.010}$	1.127 ± 0.240	$\mathbf{0.758 \pm 0.158}$
	0.6	1.782 ± 0.082	$\mathbf{1.243 \pm 0.016}$	1.199 ± 0.299	$\mathbf{0.868 \pm 0.177}$
Logistic-sin function	0.2	1.646 ± 0.052	$\mathbf{1.112 \pm 0.047}$	1.273 ± 0.285	$\mathbf{0.747 \pm 0.306}$
	0.4	2.060 ± 0.092	$\mathbf{1.507 \pm 0.057}$	1.435 ± 0.478	$\mathbf{1.003 \pm 0.399}$
	0.6	2.362 ± 0.051	$\mathbf{1.739 \pm 0.025}$	1.552 ± 0.635	$\mathbf{1.145 \pm 0.465}$

* Smaller is better. Best value in bold.

Table 1: **Results for experiments with synthetic data**. The experiments serve two purposes: (i) to show that our SurvB-learner obtains values close to the oracle bounds, and (ii) to demonstrate that the plug-in learner is not robust and therefore inferior. Experimental details are in Appendix H. We report the mean $\pm$ standard deviation of the RMSE compared to the oracle bound over 5 random runs for different synthetic datasets. Overall, we generate over 60 different datasets under varying scenarios. Our SurvB-learner outperforms the plug-in learner by a clear margin.

Result: We report the results in terms of the root mean squared error (RMSE) with respect to the oracle bounds (Table 1). Our SurvB-learner achieves the smallest average RMSE and the lowest variance, indicating that it captures the underlying patterns in the data well. On average, the RMSE from our SurvB-learner is 1.43 times smaller than that of the plug-in learner. This confirms our theoretical motivation: the plug-in learner is inferior due to plug-in bias, while our SurvB-learner, which is specifically designed to debias the plug-in estimator, performs best.

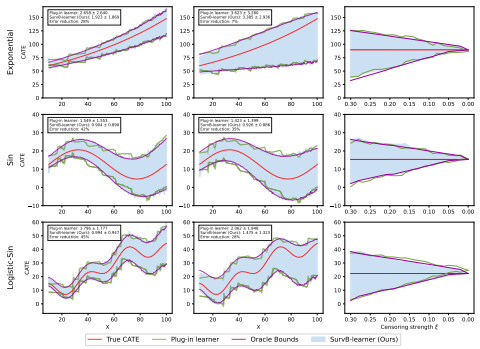

Figure 2: **Results for experiments with synthetic data**. Comparison of estimation methods for bounds based on synthetic datasets. Shown are the results *domain-knowledge bounds* (left) and for *conservative bounds* (center). **Right:** Estimated bound width across different censoring strengths for the conservative bounds. The width should go to zero as the censoring strength goes to zero. Our SurvB-learner (shown by the blue shaded area in the background) is close to the oracle bounds and often overlaps with them. In contrast, the plug-in learner is unstable ("wiggly"), because of which the plug-in learner is thus not robust, consistent with theory.

Fig. 2 provides additional insights by visualizing the oracle bounds in comparison to both the plug-in learner and our SurvB-learner. We make two key observations: (i) Our SurvB-learner reliably learns both valid domain-knowledge bounds and valid conservative bounds (see the left and center columns in Fig. 2). Our SurvB-learner aligns more closely with the oracle bounds than the plug-in learner across different datasets and censoring strengths. Oftentimes, the plug-in learner even reports bounds that are too narrow and thus not faithful. (ii) We report the widths of the conservative bound (see right column in Fig. 2). Therein, the widths of the conservative bounds shrink as censoring strength decreases toward zero. This aligns with our expectation that censoring strength primarily determines the bound width.

7.1 Empirical study: ADJUVANT dataset

We demonstrate our SurvB-learner on the ADJUVANT dataset [53, 99], a real-world clinical trial in non-small cell lung cancer. Here, we apply conservative censoring bounds to partially-identify the CATE. Details about the results, data and implementation are provided in Appendix I. $\Rightarrow$ Our SurvB-learner can reliably identify patient subgroups that benefit from treatment even in small-sample settings, which demonstrates the clinical utility of our SurvB-learner.

Conclusion: We propose an assumption-lean framework that estimates CATE bounds via our SurvB-learner to identify subgroups benefiting from treatment under informative censoring, with **future work** on learning sensitivity functions from external data and extending the framework with stronger inference tools.

Acknowledgement

Our research is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Research, Technology and Space. This work has also been supported by the German Federal Ministry of Research, Technology and Space (Grant: 03LWH0181B).

References

[1] Y. Bai and Y. Cui. Partial causal identification for right censored data with noncompliance. *Journal of Nonparametric Statistics*, 2025.

[2] G. A. Barnard. Statistical inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(2):115–149, 1949.

[3] A. P. Basu and J. P. Klein. Some recent results in competing risks theory. In *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, pages 216–229. Institute of Mathematical Statistics, Hayward, CA, 1982. ISBN 978-0-940600-02-7.

[4] N. Bo and Y. Ding. Estimating interpretable heterogeneous treatment effect with causal subgroup discovery in survival outcomes. *arXiv preprint*, arXiv:2409.19241, 2025.

[5] M. Bonvini, E. Kennedy, V. Ventura, and L. Wasserman. Sensitivity analysis for marginal structural models. *arXiv preprint*, arXiv:2210.04681, 2022.

[6] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[7] N. E. Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review / Revue Internationale de Statistique*, 43(1):45–57, 1975.

[8] W. Cai and M. J. van der Laan. One-step targeted maximum likelihood for time-to-event outcomes. *arXiv preprint*, arXiv:1802.09479, 2019.

[9] E. Candès, L. Lei, and Z. Ren. Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):24–45, 2023.

[10] C. Cheng, F. Li, L. E. Thomas, and F. F. Li. Addressing extreme propensity scores in estimating counterfactual survival functions via the overlap weights. *American Journal of Epidemiology*, 191(6):1140–1151, 2022.

[11] K. Colangelo and Y.-Y. Lee. Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint*, arXiv:2004.03036, 2023.

[12] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[13] Y. Cui, M. R. Kosorok, E. Sverdrup, S. Wager, and R. 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.

[14] A. Curth and M. van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *AISTATS*, 2021.

[15] A. Curth, C. Lee, and M. van der Schaar. Survite: Learning heterogeneous treatment effects from time-to-event data. In *NeurIPS*, 2021.

[16] H. Davidov, S. Feldman, G. Shamai, R. Kimmel, and Y. Romano. Conformalized survival analysis for general right-censored data. In *ICLR*, 2025.

[17] F. C. Detterbeck, D. J. Boffa, A. W. Kim, and L. T. Tanoue. The eighth edition lung cancer stage classification. *Chest*, 151(1):193–203, 2017.

[18] J. Dorn and K. Guo. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, 118(544):2645–2657, 2023.

[19] J. Dorn, G. Kevin, and N. Kallus. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *Journal of the American Statistical Association*, 120 (549):331–342, 2025.

[20] G. Duarte, N. Finkelstein, D. Knox, J. Mummolo, and I. Shpitser. An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association*, 119(547): 1778–1793, 2024.

[21] K. Fizazi, N. Tran, L. Fein, N. Matsubara, A. Rodriguez-Antolin, B. Y. Alekseev, M. Özgüroğlu, D. Ye, S. Feyerabend, A. Protheroe, P. De Porre, T. Kheoh, Y. C. Park, M. B. Todd, K. N. Chi, and LATITUDE Investigators. Abiraterone plus prednisone in metastatic, castration-sensitive prostate cancer. *The New England Journal of Medicine*, 377(4):352–360, 2017.

[22] D. Frauen, V. Melnychuk, and S. Feuerriegel. Sharp bounds for generalized causal sensitivity analysis. In *NeurIPS*, 2023.

[23] D. Frauen, F. Imrie, A. Curth, V. Melnychuk, S. Feuerriegel, and M. van der Schaar. A neural framework for generalized causal sensitivity analysis. In *ICLR*, 2024.

[24] D. Frauen, M. Schröder, K. Hess, and S. Feuerriegel. Orthogonal survival learners for estimating heterogeneous treatment effects from time-to-event data. In *NeurIPS*, 2025.

[25] Z. Gao and T. Hastie. Estimating heterogeneous treatment effects for general responses. *Biometrics*, 81(4):ujaf162, 2025.

[26] Y. Gui, R. Hore, Z. Ren, and R. F. Barber. Conformalized survival analysis with adaptive cut-offs. *Biometrika*, 111(2):459–477, 2024.

[27] F. Gunsilius. A path-sampling method to partially identify causal effects in instrumental variable models. *arXiv preprint*, arXiv:1910.09502, 2020.

[28] T. R. Gupta, D. Schwartz, R. Saha, P. Wen, R. Rahman, and L. Trippa. Informative censoring in externally controlled clinical trials: a potential source of bias. *ESMO Open*, 10(1):104094, 2025.

[29] F. E. Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer Series in Statistics. Springer International Publishing, 2015.

[30] N. C. Henderson, T. A. Louis, G. L. Rosner, and R. Varadhan. Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. *Biostatistics*, 21(1):50–68, 2020.

[31] S. Heng and D. S. Small. Sharpening the rosenbaum sensitivity bounds to address concerns about interactions between observed and unobserved covariates. *Statistica Sinica*, 31:2331–2353, 2021.

[32] M. A. Hernán, S. Hernández-Díaz, and J. M. Robins. A structural approach to selection bias. *Epidemiology*, 15(5):615–625, 2004.

[33] L. Hu, J. Ji, and F. Li. Estimating heterogeneous survival treatment effect in observational data using machine learning. *Statistics in Medicine*, 40(21):4691–4713, 2021.

[34] Y. Hu, S. Liu, L. Wang, Y. Liu, D. Zhang, and Y. Zhao. Treatment-free survival after discontinuation of immune checkpoint inhibitors in mNSCLC: a systematic review and meta-analysis. *Frontiers in Immunology*, 14, 2023.

[35] D. Hui, I. Glitza, G. Chisholm, S. Yennu, and E. Bruera. Attrition rates, reasons, and predictive factors in supportive care and palliative oncology clinical trials. *Cancer*, 119(5):1098–1105, 2013.

[36] H. Ichimura and W. K. Newey. The influence function of semiparametric estimators. *arXiv preprint*, arXiv:1508.01378, 2021.

[37] H. Imai, K. Mori, K. Wakuda, A. Ono, H. Akamatsu, T. Shukuya, T. Taira, H. Kenmotsu, T. Naito, K. Kaira, H. Murakami, M. Endo, T. Nakajima, N. Yamamoto, and T. Takahashi. Progression-free survival, post-progression survival, and tumor response as surrogate markers for overall survival in patients with extensive small cell lung cancer. *Annals of Thoracic Medicine*, 10(1):61–66, 2015.

[38] G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86(1):4–29, 2004.

[39] Y. Inoue, Y. Kitahara, M. Karayama, K. Asada, K. Nishimoto, S. Matsuura, D. Hashimoto, M. Fujii, T. Matsui, N. Inami, M. Toyoshima, H. Matsuda, M. Ikeda, M. Niwa, Y. Kaida, M. Sato, Y. Ito, H. Yasui, Y. Suzuki, H. Hozumi, K. Furuhashi, N. Enomoto, T. Fujisawa, N. Inui, and T. Suda. Post-discontinuation survival in patients with advanced NSCLC receiving immune checkpoint inhibitors: A pooled analysis of prospective cohort studies. *JTO Clinical and Research Reports*, 6(8):100847, 2025.

[40] A. Jesson, S. Mindermann, Y. Gal, and U. Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In *ICML*, 2021.

[41] A. Jesson, A. R. Douglas, P. Manshausen, M. Solal, N. Meinshausen, P. Stier, Y. Gal, and U. Shalit. Scalable sensitivity and uncertainty analyses for causal-effect estimates of continuous-valued interventions. In *NeurIPS*, 2022.

[42] Y. Jin, Z. Ren, and Z. Zhou. Sensitivity analysis under the f-sensitivity models: a distributional robustness perspective. *arXiv preprint*, arXiv:2203.04373, 2022.

[43] Y. Jin, Z. Ren, and E. J. Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6): e2214889120, 2023.

[44] N. Kallus, X. Mao, and A. Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In *AISTATS*, 2019.

[45] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[46] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, 2018.

[47] E. H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2), 2023.

[48] E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint*, arXiv:2203.06469, 2023.

[49] N. Kilbertus, M. J. Kusner, and R. Silva. A class of algorithms for general instrumental variable models. In *Advances in Neural Information Processing Systems*, volume 33, pages 20108–20119, 2020.

[50] J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Statistics for Biology and Health. Springer, New York, 2003.

[51] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Meta-learners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

[52] K. M. Leung, R. M. Elashoff, and A. A. Afifi. Censoring issues in survival analysis. *Annual Review of Public Health*, 18:83–104, 1997.

[53] S.-Y. Liu, H. Bao, Q. Wang, W.-M. Mao, Y. Chen, X. Tong, S.-T. Xu, L. Wu, Y.-C. Wei, Y.-Y. Liu, C. Chen, Y. Cheng, R. Yin, F. Yang, S.-X. Ren, X.-F. Li, J. Li, C. Huang, Z.-D. Liu, S. Xu, K.-N. Chen, S.-D. Xu, L.-X. Liu, P. Yu, B.-H. Wang, H.-T. Ma, H.-H. Yan, S. Dong, X.-C. Zhang, J. Su, J.-J. Yang, X.-N. Yang, Q. Zhou, X. Wu, Y. Shao, W.-Z. Zhong, and Y.-L. Wu. Genomic signatures define three subtypes of EGFR-mutant stage II–III non-small-cell lung cancer with distinct adjuvant therapy outcomes. *Nature Communications*, 12(1):6450, 2021.

[54] C. F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80 (2):319–323, 1990.

[55] H. Mao, L. Li, W. Yang, and Y. Shen. On the propensity score weighting analysis with survival outcome: Estimands, estimation, and inference. *Statistics in Medicine*, 37(26):3745–3763, 2018.

[56] Y. Matsuyama and T. Yamaguchi. Estimation of the marginal survival time in the presence of dependent competing risks using inverse probability of censoring weighted (IPCW) methods. *Pharmaceutical Statistics*, 7(3):202–214, 2008.

[57] S. A. Murphy, A. J. Rossini, and A. W. van der Vaart. Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, 92(439):968–976, 1997.

[58] X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

[59] M. Oprescu, J. Dorn, M. Ghoummaid, A. Jesson, N. Kallus, and U. Shalit. B-learner: Quasi-oracle bounds on heterogeneous causal effects under hidden confounding. In *ICML*, 2023.

[60] K. Padh, J. Zeitler, D. Watson, M. Kusner, R. Silva, and N. Kilbertus. Stochastic causal programming for bounding treatment effects. In *CLeaR*, 2023.

[61] F. Pecci, R. Thummalapalli, S. L. Alden, B. Ricciuti, J. V. Alessi, A. Elkrief, H. Rizvi, X. Wang, M. Jeng, J. V. Egger, V. R. Vaz, A. Barrichello, G. Lamberti, A. Di Federico, V. Santo, G. Rossato de Almeida, M. Gandhi, P. Clark, M. Nishino, B. E. Johnson, M. Hellmann, A. J. Schoenfeld, and M. M. Awad. Factors associated with disease progression after discontinuation of immune checkpoint inhibitors for immune-related toxicity in patients with advanced non–small cell lung cancer. *Clinical Cancer Research*, 31(12):2413–2425, 2025.

[62] P. E. Perez-Cruz, O. Shamieh, C. E. Paiva, J. H. Kwon, M. A. Muckaden, E. Bruera, and D. Hui. Factors associated with attrition in a multicenter longitudinal observational study of patients with advanced cancer. *Journal of Pain and Symptom Management*, 55(3):938–945, 2018.

[63] J. M. Robins and D. M. Finkelstein. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics*, 56(3):779–788, 2000.

[64] J. M. Robins and A. Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. *AIDS epidemiology*, pages 297–331, 1992.

[65] P. R. Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26, 1987.

[66] P. R. Rosenbaum and D. B. Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(2):212–218, 1983.

[67] D. Rubin and M. J. van der Laan. A doubly robust censoring unbiased transformation. *The International Journal of Biostatistics*, 3(1):Article 4, 2007.

[68] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

[69] M. Rubinstein, D. Agniel, L. Han, M. Horvitz-Lennon, and S.-L. Normand. Bounding causal effects with an unknown mixture of informative and non-informative censoring. *arXiv preprint*, arXiv:2411.16902, 2024.

[70] S. Sakaguchi. Partial identification and inference in duration models with endogenous censoring. *Journal of Applied Econometrics*, 39(2):308–326, 2024.

[71] D. E. Schaubel and G. Wei. Double inverse-weighted estimation of cumulative treatment effects under nonproportional hazards and dependent censoring. *Biometrics*, 67(1):29–38, 2011.

[72] S. Schrod, A. Schäfer, S. Solbrig, R. Lohmayer, W. Gronwald, P. J. Oefner, T. Beißbarth, R. Spang, H. U. Zacharias, and M. Altenbuchinger. BITES: balanced individual treatment effect for survival data. *Bioinformatics*, 38(Supplement_1):i60–i67, 2022.

[73] J. Schweisthal, D. Frauen, M. Van Der Schaar, and S. Feuerriegel. Meta-learners for partially-identified treatment effects across multiple environments. In *ICML*, 2024.

[74] J. Schweisthal, D. Frauen, M. Schröder, K. Hess, N. Kilbertus, and S. Feuerriegel. Learning representations of instruments for partial identification of treatment effects. In *ICML*, 2025.

[75] U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*, 2017.

[76] J. Shand, E. Stovold, L. Goulding, and K. Cheema. Cancer care treatment attrition in adults: Measurement approaches and inequities in patient dropout rates: a rapid review. *BMC Cancer*, 24(1):1345, 2024.

[77] E. V. Slud and L. V. Rubinstein. Dependent competing risks and summary survival curves. *Biometrika*, 70(3):643–649, 1983.

[78] D. Soriano, E. Ben-Michael, P. J. Bickel, A. Feller, and S. D. Pimentel. Interpretable sensitivity analysis for balancing weights. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(4):707–721, 2023.

[79] C. J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360, 1980.

[80] S. Tabib and D. Larocque. Non-parametric individual treatment effect estimation for survival data with random forests. *Bioinformatics (Oxford, England)*, 36(2):629–636, 2020.

[81] Z. Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

[82] A. J. Templeton, E. Amir, and I. F. Tannock. Informative censoring — a neglected cause of bias in oncology trials. *Nature Reviews Clinical Oncology*, 17(6):327–328, 2020.

[83] M. J. Van Der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer, New York, NY, 2003. ISBN 978-0-387-21700-0.

[84] C. Voinot, C. Berenfeld, I. Mayer, B. Sebastien, and J. Josse. Treatment effect estimation in causal survival analysis: Practical recommendations. *arXiv preprint*, arXiv:2501.05836, 2025.

[85] L. J. Wei. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*, 11(14-15):1871–1879, 1992.

[86] T. Westling, A. Luedtke, P. B. Gilbert, and M. Carone. Inference for treatment-specific survival curves using machine learning. *Journal of the American Statistical Association*, 119(546): 1541–1553, 2024.

[87] S. Wiegrebe, P. Kopper, R. Sonabend, B. Bischl, and A. Bender. Deep learning for survival analysis: a review. *Artificial Intelligence Review*, 57(3):65, 2024.

[88] I. Willems, J. Beyhum, and I. V. Keilegom. Bounds for the regression parameters in dependently censored survival models. *arXiv preprint*, arXiv:2503.11210, 2025.

[89] K. Xia, K.-Z. Lee, Y. Bengio, and E. Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. In *NeurIPS*, 2021.

[90] K. Xia, Y. Pan, and E. Bareinboim. Neural causal models for counterfactual identification and estimation. In *ICLR*, 2023.

[91] S. Xu, R. Cobzaru, S. N. Finkelstein, R. E. Welsch, K. Ng, and Z. Shahn. Estimating heterogeneous treatment effects on survival outcomes using counterfactual censoring unbiased transformations. *arXiv preprint*, arXiv:2401.11263, 2024.

[92] Y. Xu, N. Ignatiadis, E. Sverdrup, S. Fleming, S. Wager, and N. Shah. Treatment heterogeneity for survival outcomes. *arXiv preprint*, arXiv:2207.07758, 2022.

[93] J. Yin, M. E. Salem, J. G. Dixon, Z. Jin, R. Cohen, A. DeGramont, E. Van Cutsem, J. Taieb, S. R. Alberts, N. Wolmark, H.-J. Schmoll, L. B. Saltz, T. J. George, R. R. M. Goldberg, R. Kerr, S. Lonardi, T. Yoshino, G. Yothers, A. Grothey, T. Andre, and Q. Shi. Reevaluating disease-free survival as an endpoint vs overall survival in stage III adjuvant colon cancer trials. *JNCI: Journal of the National Cancer Institute*, 114(1):60–67, 2022.

[94] A. Ying. Proximal survival analysis to handle dependent right censoring. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1414–1434, 2024.

[95] H. Zhang and B. H. Singer. *Recursive Partitioning and Applications*. Springer Science & Business Media, 2010. ISBN 978-1-4419-6824-1.

[96] W. Zhang, T. D. Le, L. Liu, Z.-H. Zhou, and J. Li. Mining heterogeneous causal effects for personalized cancer treatment. *Bioinformatics*, 33(15):2372–2378, 2017.

[97] Q. Zhao, D. S. Small, and B. B. Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(4):735–761, 2019.

[98] M. Zheng and J. P. Klein. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1):127–138, 1995.

[99] W.-Z. Zhong, Q. Wang, W.-M. Mao, S.-T. Xu, L. Wu, Y. Shen, Y.-Y. Liu, C. Chen, Y. Cheng, L. Xu, J. Wang, K. Fei, X.-F. Li, J. Li, C. Huang, Z.-D. Liu, S. Xu, K.-N. Chen, S.-D. Xu, L.-X. Liu, P. Yu, B.-H. Wang, H.-T. Ma, H.-H. Yan, X.-N. Yang, Q. Zhou, Y.-L. Wu, S.-Y. Wang, J. Hu, W. Liu, W. Li, and J.-H. Shi. Gefitinib versus vinorelbine plus cisplatin as adjuvant treatment for stage II–IIIA (N1–N2) EGFR-mutant NSCLC (ADJUVANT/ctong1104): a randomised, open-label, phase 3 study. *The Lancet Oncology*, 19(1):139–148, 2018.

A Framework of our work

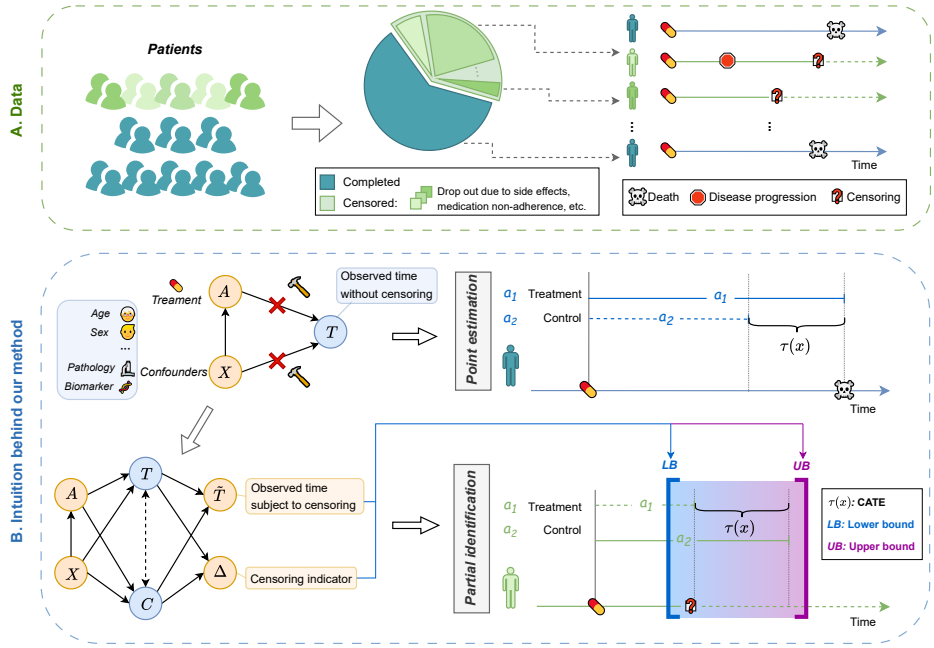

Figure 3: **Overview of our framework.** **(A)** Data: A key challenge in survival analysis is *censoring*, where follow-up information about patient outcomes is incomplete. In practice, this can be due to various reasons such as patients dropping out of clinical studies for side effects [28, 87]. As a result, the exact event time (e.g., of death or disease progression) for censored patients is unknown. **(B)** Intuition behind our framework: Under informative censoring (i.e., when dropout depends on survival time), unbiased point estimation of the CATE $\tau_{a_1,a_2}(x)$ is generally not possible. We therefore reframe the estimation task using partial identification, which allows us to construct informative bounds (i.e., *LB* and *UB*) for the CATE, while explicitly accounting for censoring bias. **(C) Output example:** We propose a two-stage meta-learner called **SurvB-learner** to efficiently estimate these bounds, which provide intuitive insights into how unknown censoring may affect CATE estimates across patient subgroups.

B Extended related work

In the following, we provide an overview of existing works that are broadly relevant to our work on partial identification of CATEs with censored datasets. To this end, we review three streams of work: (1) partial identification for CATE estimation, (2) methods dealing with informative censoring in survival analysis, and (3) survival analysis in a causal inference context.

B.1 Partial identification for CATE estimation

There is a rich body of literature on the partial identification of causal quantities from uncensored data. Existing works can be broadly categorized into two steps, beginning with methods for general partial identification and followed by those employing specific sensitivity models.

The aim of partial identification is to compute bounds for the causal quantity of interest whenever point identification is not possible [54]. Several works for partial identification have developed methods under different settings and thus different structural assumptions, such as discrete variables [20], instrumental variables [27, 49, 73, 74]. More recently, neural network–based approaches have been applied to partial identification [60, 89, 90]. However, works on partial identification for causal quantities in survival settings are comparatively scarce.

Alternatively, a large class of partial identification methods employs sensitivity models to impose assumptions on the strength of hidden confounding. These are typically classified by the underlying sensitivity model. Popular sensitivity models include Rosenbaum's sensitivity model [66, 65, 31], the marginal sensitivity model (MSM) [59, 81, 44, 97, 40, 18, 22, 78, 19], and the f-sensitivity model [42]. There are further sensitivity models suitable for continuous treatment settings, which are often based on the MSM [22, 5, 41] and the f-sensitivity model [43]. Frauen et al. [23] provide a unified framework that is compatible with multiple sensitivity models, including Rosenbaum's sensitivity model, the MSM, and the f-sensitivity model. However, all of them are dealing with unobserved confoundedness in the datasets and, therefore, are *not* directly applicable to censored datasets, where bounds do *not* come from confounding bias but censoring bias.

B.2 Classical survival methods

In medical practice, survival times are not always observed due to censoring [52]. To account for this, several tools have been developed, including the Kaplan-Meier estimator [45] and the Cox proportional hazards model [12, 7]. Other common approaches include accelerated failure time (AFT) models [85] and proportional odds models [57, 29], which combine parametric and semi-parametric specifications. Together, these methods form the foundation of standard survival analysis, but they typically assume that censoring is *non-informative* (while our focus is on *informative* censoring).

B.3 Survival analysis in causal inference contexts

Efficient average treatment effect estimation for time-to-event data: Most work on semiparametric inference for time-to-event data has focused on average treatment effects (ATEs). For example, Rubin and van der Laan [67] proposed so-called doubly robust censoring unbiased transformations for semiparametric efficient inference on the ATE under censoring. Furthermore, various works proposed one-step Targeted Maximum Likelihood Estimators (TMLE) for estimating average causal quantities in survival settings [83, 8, 86]. Complementary to these, Cheng et al. [10] proposed inverse probability of censoring weighted (IPCW) estimators for weighted average survival functions, though these rely on correctly specified parametric nuisance models.

Early developments in causal survival analysis have focused on addressing the bias induced by *informative* censoring. Methods such as those in [64, 63, 56] developed inverse probability of censoring weighting (IPCW) techniques to obtain point estimates under specific model, e.g., Kaplan-Meier esitmator, proportional hazard model and AFT models. However, these approaches are not framed around current potential outcomes frameworks and therefore do not target causal estimands such as the CATE. Moreover, none of them are model-agnostic sensitivity tools: their bounds are tied to particular survival models.

Building upon this gap, recent work extends survival analysis to causal inference, where the goal is to estimate treatment effects, such as the CATE, under censoring. Here, methods fall broadly into

two categories: model-based learners and meta-learners. The former, model-based learners, include tree-based methods [96, 30, 80, 13, 33] and neural-network-based methods [72, 46, 14]. In contrast, few works have proposed (orthogonal) meta-learners specifically designed for censored time-to-event data [25, 91, 24, 92]. However, these methods focus on *point* estimation and typically require *censoring-independent assumptions*, which limit their use in real-world applications. Another line of research develops inferential methods for censored data based on conformal inference [9, 26, 16], but these approaches aim to produce prediction intervals for individual treatment effects rather than bounds on CATEs due to censoring.

Some recent studies have aimed to relax the standard assumption of non-informative censoring. For example, Bai and Cui [1] developed a partial identification method for the ATE using instrumental variables. Rubinstein et al. [69] proposed a method for bounding the ATE under a mixture of informative and non-informative censoring, which relies on the ratio of informative censoring in the dataset to bound the ATE rather than focusing solely on informative censoring. Our method develops bounds for CATE under informative censoring. Ying [94] proposed a proxy-based causal inference method. However, these approaches focus on the ATE and *not* the CATE under informative censoring.

Research gap: To the best of our knowledge, no existing work provides partial identification bounds for the CATE in survival analysis under informative censoring.

C SurvB-learner for CATE

In our main paper, we have stated our SurvB-learner for the CAPOs. In this section, we now state our SurvB-learner for the CATE. For this, we follow Theorem 5.1 and take the difference between the lower bounds and upper bounds of CAPOs. As in the main paper, we distinguish two cases: **Case** (1) with domain knowledge, and **Case** (2) with a conservative upper bound.

• **Case** (1) (*domain-knowledge upper bounds).* Here, we construct upper bounds for a given $\gamma(x, a)$. Then, the pseudo-outcome of the upper bound is given by

$$
\begin{aligned}
&\hat{\phi}^{+}_{a_1,a_2}(x) \\
=&\hat{\phi}^{+}(x,a_1) - \hat{\phi}^{-}(x,a_2) \\
=&\frac{1(A=a_1,\Delta=0)}{\hat{\pi}_{a_1}(x)}\{\tilde{T}-\hat{\nu}(0,x,a_1)\} + \frac{\hat{\nu}(0,x,a_1)1(A=a_1)}{\hat{\pi}_{a_1}(x)}\{1(\Delta=0)-[1-\hat{\xi}(x,a_1)]\} + \hat{\nu}(0,x,a_1)[1-\hat{\xi}(x,a_1)] \\
&+ \frac{1(A=a_1,\Delta=1)}{\hat{\pi}_{a_1}(x)}\{\tilde{T}-\hat{\nu}(1,x,a_1)\} + \frac{\hat{\nu}(1,x,a_1)1(A=a_1)}{\hat{\pi}_{a_1}(x)}\{1(\Delta=1)-\hat{\xi}(x,a_1)\} + \hat{\nu}(1,x,a_1)\hat{\xi}(x,a_1) \\
&+ \gamma(x,a_1)\frac{1(A=a_1)}{\hat{\pi}_{a_1}(x)}\left\{1(\Delta=1)-\hat{\xi}(x,a_1)\right\} + \gamma(x,a_1)\hat{\xi}(x,a_1) \\
&- \frac{1(A=a_2,\Delta=0)}{\hat{\pi}_{a_2}(x)}\{\tilde{T}-\hat{\nu}(0,x,a_2)\} - \frac{\hat{\nu}(0,x,a_2)1(A=a_2)}{\hat{\pi}_{a_2}(x)}\{1(\Delta=0)-[1-\hat{\xi}(x,a_2)]\} - \hat{\nu}(0,x,a_2)[1-\hat{\xi}(x,a_2)] \\
&- \frac{1(A=a_2,\Delta=1)}{\hat{\pi}_{a_2}(x)}\{\tilde{T}-\hat{\nu}(1,x,a_2)\} - \frac{\hat{\nu}(1,x,a_2)1(A=a_2)}{\hat{\pi}_{a_2}(x)}\{1(\Delta=1)-\hat{\xi}(x,a_2)\} - \hat{\nu}(1,x,a_2)\hat{\xi}(x,a_2).
\end{aligned}
\tag{15}
$$

The pseudo-outcome of the lower bound is given by

$$
\begin{aligned}
&\hat{\phi}^{-}_{a_1,a_2}(x) \\
=&\hat{\phi}^{-}(x,a_1) - \hat{\phi}^{+}(x,a_2) \\
=&\frac{1(A=a_1,\Delta=0)}{\hat{\pi}_{a_1}(x)}\{\tilde{T}-\hat{\nu}(0,x,a_1)\} + \frac{\hat{\nu}(0,x,a_1)1(A=a_1)}{\hat{\pi}_{a_1}(x)}\{1(\Delta=0)-[1-\hat{\xi}(x,a_1)]\} + \hat{\nu}(0,x,a_1)[1-\hat{\xi}(x,a_1)] \\
&+ \frac{1(A=a_1,\Delta=1)}{\hat{\pi}_{a_1}(x)}\{\tilde{T}-\hat{\nu}(1,x,a_1)\} + \frac{\hat{\nu}(1,x,a_1)1(A=a_1)}{\hat{\pi}_{a_1}(x)}\{1(\Delta=1)-\hat{\xi}(x,a_1)\} + \hat{\nu}(1,x,a_1)\hat{\xi}(x,a_1) \\
&- \frac{1(A=a_2,\Delta=0)}{\hat{\pi}_{a_2}(x)}\{\tilde{T}-\hat{\nu}(0,x,a_2)\} - \frac{\hat{\nu}(0,x,a_2)1(A=a_2)}{\hat{\pi}_{a_2}(x)}\{1(\Delta=0)-[1-\hat{\xi}(x,a_2)]\} - \hat{\nu}(0,x,a_2)[1-\hat{\xi}(x,a_2)] \\
&- \frac{1(A=a_2,\Delta=1)}{\hat{\pi}_{a_2}(x)}\{\tilde{T}-\hat{\nu}(1,x,a_2)\} - \frac{\hat{\nu}(1,x,a_2)1(A=a_2)}{\hat{\pi}_{a_2}(x)}\{1(\Delta=1)-\hat{\xi}(x,a_2)\} - \hat{\nu}(1,x,a_2)\hat{\xi}(x,a_2) \\
&- \gamma(x,a_2)\frac{1(A=a_2)}{\hat{\pi}_{a_2}(x)}\left\{1(\Delta=1)-\hat{\xi}(x,a_2)\right\} - \gamma(x,a_2)\hat{\xi}(x,a_2).
\end{aligned}
\tag{16}
$$

• **Case** (2) (*conservative upper bounds).* The conservative upper bound is a special case of $\gamma(x, a) = t_{\max} - \mathbb{E}[\tilde{T} \mid \Delta = 1, X = x, A = a]$. Here, the corresponding pseudo-outcome is given by

$$
\begin{aligned}
&\hat{\phi}^{+}_{a_1,a_2}(x) \\
=&\hat{\phi}^{+}(x,a_1) - \hat{\phi}^{-}(x,a_2) \\
=&\frac{1(A=a_1,\Delta=0)}{\hat{\pi}_{a_1}(x)}\{\tilde{T}-\hat{\nu}(0,x,a_1)\} + \frac{\hat{\nu}(0,x,a_1)1(A=a_1)}{\hat{\pi}_{a_1}(x)}\{1(\Delta=0)-[1-\hat{\xi}(x,a_1)]\} + \hat{\nu}(0,x,a_1)[1-\hat{\xi}(x,a_1)] \\
&+ t_{\max}\frac{1(A=a_1)}{\hat{\pi}_{a_1}(x)}\left\{1(\Delta=1)-\hat{\xi}(x,a_1)\right\} + t_{\max}\hat{\xi}(x,a_1) \\
&- \frac{1(A=a_2,\Delta=0)}{\hat{\pi}_{a_2}(x)}\{\tilde{T}-\hat{\nu}(0,x,a_2)\} - \frac{\hat{\nu}(0,x,a_2)1(A=a_2)}{\hat{\pi}_{a_2}(x)}\{1(\Delta=0)-[1-\hat{\xi}(x,a_2)]\} - \hat{\nu}(0,x,a_2)[1-\hat{\xi}(x,a_2)] \\
&- \frac{1(A=a_2,\Delta=1)}{\hat{\pi}_{a_2}(x)}\{\tilde{T}-\hat{\nu}(1,x,a_2)\} - \frac{\hat{\nu}(1,x,a_2)1(A=a_2)}{\hat{\pi}_{a_2}(x)}\{1(\Delta=1)-\hat{\xi}(x,a_2)\} - \hat{\nu}(1,x,a_2)\hat{\xi}(x,a_2).
\end{aligned}
\tag{17}
$$

The pseudo-outcome of the lower bound is given by

$$
\begin{aligned}
&\hat{\phi}^{-}_{a_1,a_2}(x) \\
=&\hat{\phi}^{-}(x,a_1) - \hat{\phi}^{+}(x,a_2) \\
=&\frac{1(A=a_1,\Delta=0)}{\hat{\pi}_{a_1}(x)}\{\tilde{T}-\hat{\nu}(0,x,a_1)\} + \frac{\hat{\nu}(0,x,a_1)1(A=a_1)}{\hat{\pi}_{a_1}(x)}\{1(\Delta=0)-[1-\hat{\xi}(x,a_1)]\} + \hat{\nu}(0,x,a_1)[1-\hat{\xi}(x,a_1)] \\
&+ \frac{1(A=a_1,\Delta=1)}{\hat{\pi}_{a_1}(x)}\{\tilde{T}-\hat{\nu}(1,x,a_1)\} + \frac{\hat{\nu}(1,x,a_1)1(A=a_1)}{\hat{\pi}_{a_1}(x)}\{1(\Delta=1)-\hat{\xi}(x,a_1)\} + \hat{\nu}(1,x,a)\hat{\xi}(x,a_1) \\
&- \frac{1(A=a_2,\Delta=0)}{\hat{\pi}_{a_2}(x)}\{\tilde{T}-\hat{\nu}(0,x,a_2)\} - \frac{\hat{\nu}(0,x,a_2)1(A=a_2)}{\hat{\pi}_{a_2}(x)}\{1(\Delta=0)-[1-\hat{\xi}(x,a_2)]\} - \hat{\nu}(0,x,a_2)[1-\hat{\xi}(x,a_2)] \\
&- t_{\max}\frac{1(A=a_2)}{\hat{\pi}_{a_2}(x)}\left\{1(\Delta=1)-\hat{\xi}(x,a_2)\right\} - t_{\max}\hat{\xi}(x,a_2).
\end{aligned}
\tag{18}
$$

Algorithm 1 SurvB-learner

Input: Observational data $\mathcal{D} = \{(x_i, a_i, \tilde{t}_i, \delta_i)_{i=1}^n\}$, domain knowledge $\gamma(x,a)$
Ensure: Estimated bounds $\hat{\mu}^-(x,a)$, $\hat{\mu}^+(x,a)$

/* Stage 1: Nuisance estimation with cross-fitting */
$\hat{\nu}(\delta,x,a) \leftarrow \hat{\mathbb{E}}[\tilde{T} \mid \Delta=\delta, X=x, A=a]$
$\hat{\xi}(x,a) \leftarrow \hat{\mathbb{P}}[\Delta=1 \mid X=x, A=a]$
$\hat{\pi}_a(x) \leftarrow \hat{\mathbb{P}}[A=a \mid X=x]$
$\hat{\phi}^{\pm}(x,a) \leftarrow$ Eq. (11), (12), (13) based on $\{\hat{\pi}_a, \hat{\xi}, \hat{\nu}\}$
/* Stage 2: Pseudo-outcome regression */
$\hat{\mu}^-(x,a) \leftarrow \hat{\mathbb{E}}[\hat{\phi}^- \mid X=x]$
$\hat{\mu}^+(x,a) \leftarrow \hat{\mathbb{E}}[\hat{\phi}^+ \mid X=x]$
return $\hat{\mu}^-(x,a),\ \hat{\mu}^+(x,a)$

D Proofs

D.1 Proof of Lemma 3.2

Proof. The result follows from the law of total expectation [2] and the identifiability assumptions, i.e.,

$$\begin{aligned}
&\mathbb{E}[T(a) \mid X = x] \\
=&\mathbb{E}[T \mid X = x, A = a] \\
=&\mathbb{E}[T \mid X = x, A = a, \Delta = 0] P(\Delta = 0 \mid X = x, A = a) \\
&+ \mathbb{E}[T \mid X = x, A = a, \Delta = 1] P(\Delta = 1 \mid X = x, A = a).
\end{aligned} \tag{19}$$

Then, we have

$$\begin{aligned}
&\mathbb{E}[T(a_1) - T(a_2) \mid X = x] \\
=&\mathbb{E}[T(a_1) \mid X = x] - \mathbb{E}[T(a_2) \mid X = x] \\
=&\mathbb{E}[T \mid X = x, A = a_1] - \mathbb{E}[T \mid X = x, A = a_2] \\
=&\nu(0, x, a_1)[1 - \xi(x, a_1)] + \mathbb{E}[T \mid X = x, A = a_1, \Delta = 1]\xi(x, a_1) \\
&- \nu(0, x, a_2)[1 - \xi(x, a_2)] - \mathbb{E}[T \mid X = x, A = a_2, \Delta = 1]\xi(x, a_2).
\end{aligned} \tag{20}$$

□

D.2 Proof of Theorem 5.1

Proof. Note that

$$\tau_{a_1,a_2}(x) = \mu(x, a_1) - \mu(x, a_2). \tag{21}$$

Further, we have that

$$\mu(x, a) = \nu(0, x, a)[1 - \xi(x, a)] + \mathbb{E}[T \mid X = x, A = a, \Delta = 1]\xi(x, a). \tag{22}$$

while $\mathbb{E}[\tilde{T} \mid \Delta = 1, X = x, A = a] \leq \mathbb{E}[T \mid \Delta = 1, X = x, A = a] \leq \mathbb{E}[\tilde{T} \mid \Delta = 1, X = x, A = a] + \gamma(x, a)$ by the definition of $\gamma(x, a)$. Therefore, by the definitions of $\mu^-(x, a)$ and $\mu^+(x, a)$, we have

$$\mu^-(x, a) \leq \mu(x, a) \leq \mu^+(x, a). \tag{23}$$

Hence, taking the minimum and maximum of $\mu(x, a_1)$ and $\mu(x, a_2)$ yields the result:

$$\mu^-(x, a_1) - \mu^+(x, a_2) \leq \tau_{a_1,a_2}(x) \leq \mu^+(x, a_1) - \mu^-(x, a_2). \tag{24}$$

□

D.3 Proof of Proposition 5.2

Proof. First, we recall the definition of the domain knowledge upper bounds for two CAPOs with different treatments a_1 and a_2:

$$\begin{aligned}
\mu^-(x,a_1) =& \nu(0,x,a_1)(1-\xi(x,a_1)) + \nu(1,x,a_1)\xi(x,a_1),\\
\mu^+(x,a_1) =& \nu(0,x,a_1)(1-\xi(x,a_1)) + \nu(1,x,a_1)\xi(x,a_1) + \gamma(x,a_1)\xi(x,a_1),\\
\mu^-(x,a_2) =& \nu(0,x,a_2)(1-\xi(x,a_2)) + \nu(1,x,a_2)\xi(x,a_2),\\
\mu^+(x,a_2) =& \nu(0,x,a_2)(1-\xi(x,a_2)) + \nu(1,x,a_2)\xi(x,a_2) + \gamma(x,a_2)\xi(x,a_2).
\end{aligned} \tag{25}$$

Then, we could see that for treatment $a \in \mathcal{A}$, the width of CAPO bounds is

$$\mu^+(x,a) - \mu^-(x,a) = \gamma(x,a)\xi(x,a) \tag{26}$$

Then, we follow Theorem 5.1 to derive ① domain knowledge bounds of CATE $\tau_{a_1,a_2}(x)$ via

$$\begin{aligned}
&\tau^-_{a_1,a_2}(x)\\
=&\mu^-(x,a_1) - \mu^+(x,a_2)\\
=&\nu(0,x,a_1)(1-\xi(x,a_1)) + \nu(1,x,a_1)\xi(x,a_1) - \nu(0,x,a_2)(1-\xi(x,a_2)) - \nu(1,x,a_2)\xi(x,a_2)\\
&- \gamma(x,a_2)\xi(x,a_2),\\
&\tau^+_{a_1,a_2}(x)\\
=&\mu^+(x,a_1) - \mu^-(x,a_2)\\
=&\nu(0,x,a_1)(1-\xi(x,a_1)) + \nu(1,x,a_1)\xi(x,a_1) - \nu(0,x,a_2)(1-\xi(x,a_2)) - \nu(1,x,a_2)\xi(x,a_2)\\
&+ \gamma(x,a_1)\xi(x,a_1).
\end{aligned} \tag{27}$$

By the definition of $\gamma(x,a)$ in Eq. (5) and Theorem 5.1, the true CATE is guaranteed to lie within these bounds, i.e., $\tau^-_{a_1,a_2}(x) \le \tau_{a_1,a_2}(x) \le \tau^+_{a_1,a_2}(x)$. The width of the bounds is then

$$\tau^+_{a_1,a_2}(x) - \tau^-_{a_1,a_2}(x) = \gamma(x,a_1)\xi(x,a_1) + \gamma(x,a_2)\xi(x,a_2). \tag{28}$$

Here, we yield for ② conservative bounds of CATE $\tau_{a_1,a_2}(x)$ via

$$\begin{aligned}
\mu^-(x,a_1) =& \nu(0,x,a_1)(1-\xi(x,a_1)) + \nu(1,x,a_1)\xi(x,a_1),\\
\mu^+(x,a_1) =& \nu(0,x,a_1)(1-\xi(x,a_1)) + t_{\max}\xi(x,a_1)\\
\mu^-(x,a_2) =& \nu(0,x,a_2)(1-\xi(x,a_2)) + \nu(1,x,a_2)\xi(x,a_2),\\
\mu^+(x,a_2) =& \nu(0,x,a_2)(1-\xi(x,a_2)) + t_{\max}\xi(x,a_2).
\end{aligned} \tag{29}$$

Next, we follow Theorem 5.1 to derive the domain knowledge bounds of the CATE $\tau_{a_1,a_2}(x)$. We obtain

$$\begin{aligned}
&\tau^-_{a_1,a_2}(x)\\
=&\mu^-(x,a_1) - \mu^+(x,a_2)\\
=&\nu(0,x,a_1)(1-\xi(x,a_1)) + \nu(1,x,a_1)\xi(x,a_1) - \nu(0,x,a_2)(1-\xi(x,a_2)) - t_{\max}\xi(x,a_2),\\
&\tau^+_{a_1,a_2}(x)\\
=&\mu^+(x,a_1) - \mu^-(x,a_2)\\
=&\nu(0,x,a_1)(1-\xi(x,a_1)) + t_{\max}\xi(x,a_1) - \nu(0,x,a_2)(1-\xi(x,a_2)) - \nu(1,x,a_2)\xi(x,a_2)
\end{aligned} \tag{30}$$

By the definition of $\gamma(x,a) = t_{\max} - \nu(1,x,a)$ in Eq. (5) and Theorem 5.1, the true CATE is guaranteed to lie within these bounds, i.e., $\tau^-_{a_1,a_2}(x) \le \tau_{a_1,a_2}(x) \le \tau^+_{a_1,a_2}(x)$. The width of the bounds is then

$$\tau^+_{a_1,a_2}(x) - \tau^-_{a_1,a_2}(x) = [t_{\max} - \nu(1,x,a_1)]\xi(x,a_1) + [t_{\max} - \nu(1,x,a_2)]\xi(x,a_2). \tag{31}$$

□

D.4 Derivation of meta-learners

First, we derive the efficient influence function (EIF) score for the **lower bound**:

$$
\begin{aligned}
&\mathbb{EIF}(\psi(\mathbb{P})) \\
=&\mathbb{EIF}\left(\int_x \mathbb{E}\left[\nu(0,x,a)\left(1-\xi(x,a)\right)+\nu(1,x,a)\xi(x,a)\right]p(x)\,dx\right) \\
=&\frac{1(A=a,\Delta=0)}{\hat{\pi}_a(x)}\{\tilde{T}-\hat{\nu}(0,x,a)\}+\frac{\hat{\nu}(0,x,a)1(A=a)}{\hat{\pi}_a(x)}\{1(\Delta=0)-[1-\hat{\xi}(x,a)]\}+\hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)] \\
&+\frac{1(A=a,\Delta=1)}{\hat{\pi}_a(x)}\{\tilde{T}-\hat{\nu}(1,x,a)\}+\frac{\hat{\nu}(1,x,a)1(A=a)}{\hat{\pi}_a(x)}\{1(\Delta=1)-\hat{\xi}(x,a)\}+\hat{\nu}(1,x,a)\hat{\xi}(x,a)-\psi(a).
\end{aligned}
\tag{32}
$$

Upper bound: For the upper bound, we distinguish the two cases from the main paper as follows.

Case ①:(*domain-knowledge upper bounds*). We treat the domain function $\gamma(x,a)$ as a constant function and thus yield

$$
\begin{aligned}
&\mathbb{EIF}(\psi(\mathbb{P})) \\
=&\mathbb{EIF}\left(\int_x \mathbf{E}\left[\nu(0,x,a)\left(1-\xi(x,a)\right)+\nu(1,x,a)\xi(x,a)+\gamma(x,a)\xi(x,a)\right]p(x)\,dx\right) \\
=&\frac{1(A=a,\Delta=0)}{\hat{\pi}_a(x)}\{\tilde{T}-\hat{\nu}(0,x,a)\}+\frac{\hat{\nu}(0,x,a)1(A=a)}{\hat{\pi}_a(x)}\{1(\Delta=0)-[1-\hat{\xi}(x,a)]\}+\hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)] \\
&+\frac{1(A=a,\Delta=1)}{\hat{\pi}_a(x)}\{\tilde{T}-\hat{\nu}(1,x,a)\}+\frac{\hat{\nu}(1,x,a)1(A=a)}{\hat{\pi}_a(x)}\{1(\Delta=1)-\hat{\xi}(x,a)\}+\hat{\nu}(1,x,a)\hat{\xi}(x,a) \\
&+\gamma(x,a)\frac{1(A=a)}{\hat{\pi}_a(x)}\{1(\Delta=1)-\xi(X,a)\}+\gamma(x,a)\xi(X,a)-\psi(a).
\end{aligned}
\tag{33}
$$

Case ②:(*conservative upper bounds*). We set $\gamma(x,a)$ as our known sensitivity function. We yield

$$
\begin{aligned}
&\mathbb{EIF}(\psi(\mathbb{P})) \\
=&\mathbb{EIF}\left(\int_x \nu(0,x,a)\left(1-\xi(x,a)\right)+t_{\max}\xi(x,a)p(x)\,dx\right) \\
=&\int_x \mathbb{EIF}\left[\nu(0,x,a)\left(1-\xi(x,a)\right)+t_{\max}\xi(x,a)p(x)\,dx\right] \\
=&\int_x \mathbb{EIF}\left[\nu(0,x,a)\left(1-\xi(x,a)\right)p(x)\right]+\mathbb{EIF}\left[t_{\max}\xi(x,a)p(x)\right]dx \\
=&\int_x \mathbb{EIF}\left[\nu(0,x,a)\left(1-\xi(x,a)\right)p(x)\right]+\mathbb{EIF}\left[t_{\max}\xi(x,a)p(x)\right]dx \\
=&\underbrace{\int_x \mathbb{EIF}\left[\nu(0,x,a)\left(1-\xi(x,a)\right)p(x)\right]dx}_{*}+\underbrace{\int_x \mathbb{EIF}\left[t_{\max}\xi(x,a)p(x)\right]dx}_{**}.
\end{aligned}
\tag{34}
$$

We proceed with the derivation by separating the above expression into two parts:

$$
\begin{aligned}
&(*)\\
=&\int_x \mathbb{EIF}\left[\nu(0,x,a)\right]\left[1-\xi(x,a)\right]p(x)+\nu(0,x,a)\mathbb{EIF}\left[1-\xi(x,a)\right]p(x)+\nu(0,x,a)\left[1-\xi(x,a)\right]\mathbb{EIF}\left[p(x)\right]dx\\
=&\int_x \frac{1\{X=x,A=a,\Delta=0\}}{p(x,a)[1-\xi(x,a)]}[\tilde{T}-\nu(0,x,a)]\left[1-\xi(x,a)\right]p(x)\\
&+\nu(0,x,a)\frac{1\{X=x,A=a\}}{p(x,a)}[1(\Delta=0)-1+\xi(x,a)]p(x)\\
&+\nu(0,x,a)\left[1-\xi(x,a)\right]\left[1\{X=x\}-p(x)\right]dx\\
=&\int_x \frac{1(X=x,A=a,\Delta=0)}{p(a\mid x)}\left[\tilde{T}-\nu(0,x,a)\right]+\frac{1(X=x,A=a)}{p(a\mid x)}\nu(0,x,a)\left[1(\Delta=0)-1+\xi(x,a)\right]\\
&+\nu(0,x,a)[1-\xi(x,a)]\left[1\{X=x\}-p(x)\right]dx\\
=&\frac{1(A=a,\Delta=0)}{p(a\mid X)}\left[\tilde{T}-\nu(0,x,a)\right]+\frac{1(A=a)}{p(a\mid X)}\nu(0,X,a)\{1(\Delta=0)-[1-\xi(X,a)]\}\\
&+\nu(0,X,a)[1-\xi(X,a)]-\psi_*(a),
\end{aligned}
\tag{35}
$$

and

$$
\begin{aligned}
&(**)\\
=&\int_x \mathbb{EIF}\left[t_{\max}\xi(x,a)p(x)\right]dx\\
=&t_{\max}\int_x \mathbb{EIF}\left[\xi(x,a)p(x)\right]dx\\
=&t_{\max}\int_x \mathbb{EIF}\left[\xi(x,a)]p(x)+\xi(x,a)\mathbb{EIF}\left[p(x)\right]dx\right.\\
=&t_{\max}\int_x \frac{1(X=x,A=a)}{p(X=x,A=a)}\left\{1(\Delta=1)-\xi(x,a)\right\}p(x)+\xi(x,a)\left[1\{X=x\}-p(x)\right]dx\\
=&t_{\max}\frac{1(A=a)}{p(a\mid X)}\left\{1(\Delta=1)-\xi(X,a)\right\}+t_{\max}\xi(X,a)-\psi_{**}(a).
\end{aligned}
\tag{36}
$$

Finally, our estimator for the conservative upper bound is

$$
\begin{aligned}
&\hat{\phi}^+(x,a)\\
=&\frac{1(A=a,\Delta=0)}{\hat{\pi}_a(x)}\{\tilde{T}-\hat{\nu}(0,x,a)\}+\frac{\hat{\nu}(0,x,a)1(A=a)}{\hat{\pi}_a(x)}\{1(\Delta=0)-[1-\hat{\xi}(x,a)]\}\\
&+\hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)]+t_{\max}\frac{1(A=a)}{\hat{\pi}_a(x)}\left\{1(\Delta=1)-\hat{\xi}(x,a)\right\}+t_{\max}\hat{\xi}(x,a).
\end{aligned}
\tag{37}
$$

D.5 Proof of Theorem 6.1

Proof. We proceed by calculating $\mathbb{E}\left[\hat{\phi}^{+} \mid X=x\right]$ and $\mathbb{E}\left[\hat{\phi}^{-} \mid X=x\right]$ for each pseudo-outcome $\hat{\phi}^{+}$ and $\hat{\phi}^{-}$, which corresponds to an oracle second stage regression.

We start with **the lower bounds**, which use pseudo-outcomes defined in Eq. (11). We can then write the equation as

$$
\begin{aligned}
&\hat{\phi}^{-}(x,a)\\
=&\frac{1(A=a,\Delta=0)}{\hat{\pi}_a(x)}\{\tilde{T}-\hat{\nu}(0,x,a)\}+\frac{\hat{\nu}(0,x,a)1(A=a)}{\hat{\pi}_a(x)}\{1(\Delta=0)-[1-\hat{\xi}(x,a)]\}\\
&+\frac{1(A=a,\Delta=1)}{\hat{\pi}_a(x)}\{\tilde{T}-\hat{\nu}(1,x,a)\}+\frac{\hat{\nu}(1,x,a)1(A=a)}{\hat{\pi}_a(x)}\{1(\Delta=1)-\hat{\xi}(x,a)\}\\
&+\hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)]+\hat{\nu}(1,x,a)\hat{\xi}(x,a)\\
=&\frac{1(A=a,\Delta=0)}{\hat{\pi}_a(x)}\tilde{T}-\frac{1(A=a)}{\hat{\pi}_a(x)}\hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)]+\hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)]\\
&+\frac{1(A=a,\Delta=1)}{\hat{\pi}_a(x)}\tilde{T}-\frac{1(A=a)}{\hat{\pi}_a(x)}\hat{\nu}(1,x,a)\hat{\xi}(x,a)+\hat{\nu}(1,x,a)\hat{\xi}(x,a).
\end{aligned}
\tag{38}
$$

Hence, by calculating the conditional expectation, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\hat{\phi}^{-}(x,a) \mid X=x\right]\\
=&\mathbb{E}\left[\frac{1(A=a,\Delta=0)}{\hat{\pi}_a(x)}\tilde{T} \mid X=x\right]-\mathbb{E}\left[\frac{1(A=a)}{\hat{\pi}_a(x)}\hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)] \mid X=x\right]\\
&+\mathbb{E}\left[\hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)] \mid X=x\right]+\mathbb{E}\left[\frac{1(A=a,\Delta=1)}{\hat{\pi}_a(x)}\tilde{T} \mid X=x\right]\\
&-\mathbb{E}\left[\frac{1(A=a)}{\hat{\pi}_a(x)}\hat{\nu}(1,x,a)\hat{\xi}(x,a) \mid X=x\right]+\mathbb{E}\left[\hat{\nu}(1,x,a)\hat{\xi}(x,a) \mid X=x\right]\\
=&\frac{\pi_a(x)}{\hat{\pi}_a(x)}[1-\xi(x,a)]\nu(0,x,a)-\frac{\pi_a(x)}{\hat{\pi}_a(x)}[1-\hat{\xi}(x,a)]\hat{\nu}(0,x,a)+\hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)]\\
&+\frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a)\nu(1,x,a)-\frac{\pi_a(x)}{\hat{\pi}_a(x)}\hat{\xi}(x,a)\hat{\nu}(1,x,a)+\hat{\nu}(1,x,a)\hat{\xi}(x,a)
\end{aligned}
\tag{39}
$$

Using $\hat{\nu}(\delta,x,a)=\nu(\delta,x,a)$, $\hat{\xi}(x,a)=\xi(x,a)$ and $\hat{\pi}_a(x)=\pi_a(x)$ implies

$$
\begin{aligned}
&\mathbb{E}\left[\hat{\phi}^{-}(x,a) \mid X=x\right]\\
=&\frac{\pi_a(x)}{\hat{\pi}_a(x)}[1-\xi(x,a)]\nu(0,x,a)-\frac{\pi_a(x)}{\hat{\pi}_a(x)}[1-\hat{\xi}(x,a)]\hat{\nu}(0,x,a)+\hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)]\\
&+\frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a)\nu(1,x,a)-\frac{\pi_a(x)}{\hat{\pi}_a(x)}\hat{\xi}(x,a)\hat{\nu}(1,x,a)+\hat{\nu}(1,x,a)\hat{\xi}(x,a)\\
=&[1-\xi(x,a)]\nu(0,x,a)-[1-\xi(x,a)]\nu(0,x,a)+[1-\xi(x,a)]\nu(0,x,a)\\
&+\xi(x,a)\nu(1,x,a)-\xi(x,a)\nu(1,x,a)+\xi(x,a)\nu(1,x,a)\\
=&[1-\xi(x,a)]\nu(0,x,a)+\xi(x,a)\nu(1,x,a)\\
=&\mu^{-}(x,a),
\end{aligned}
\tag{40}
$$

which proves consistency.

We distinguish two cases. (1) Under $\hat{\nu}(\delta, x, a) = \nu(\delta, x, a)$ and $\hat{\xi}(x, a) = \xi(x, a)$, this reduces to

$$\begin{aligned}
&\mathbb{E}\left[\hat{\phi}^-(x,a) \mid X = x\right] \\
=&\frac{\pi_a(x)}{\hat{\pi}_a(x)}[1-\xi(x,a)]\nu(0,x,a) - \frac{\pi_a(x)}{\hat{\pi}_a(x)}[1-\xi(x,a)]\nu(0,x,a) + \nu(0,x,a)[1-\xi(x,a)] \\
&+ \frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a)\nu(1,x,a) - \frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a)\nu(0,x,a) + \nu(1,x,a)\xi(x,a) \\
=&[1-\xi(x,a)]\nu(0,x,a) + \xi(x,a)\nu(1,x,a) \\
=&\mu^-(x,a).
\end{aligned} \tag{41}$$

(2) Under $\hat{\pi}_a(x) = \pi_a(x)$, this reduces to

$$\begin{aligned}
&\mathbb{E}\left[\hat{\phi}^-(x,a) \mid X = x\right] \\
=&\frac{\pi_a(x)}{\pi_a(x)}[1-\xi(x,a)]\nu(0,x,a) - \frac{\pi_a(x)}{\pi_a(x)}[1-\hat{\xi}(x,a)]\hat{\nu}(0,x,a) + \hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)] \\
&+ \frac{\pi_a(x)}{\pi_a(x)}\xi(x,a)\nu(1,x,a) - \frac{\pi_a(x)}{\pi_a(x)}\hat{\xi}(x,a)\hat{\nu}(0,x,a) + \hat{\nu}(1,x,a)\hat{\xi}(x,a) \\
=&[1-\xi(x,a)]\nu(0,x,a) + \xi(x,a)\nu(1,x,a) \\
=&\mu^-(x,a).
\end{aligned} \tag{42}$$

Together, this proves the double robustness.

Next, we move on to the upper bounds and prove them again in two cases. The pseudo-outcome of **Case** 1(*domain knowledge upper bounds*). defined in Eq. (12) can be simplified to

$$\begin{aligned}
&\hat{\phi}^+(x,a) \\
=&\frac{1(A=a,\Delta=0)}{\hat{\pi}_a(x)}\{\tilde{T} - \hat{\nu}(0,x,a)\} + \frac{\hat{\nu}(0,x,a)1(A=a)}{\hat{\pi}_a(x)}\{1(\Delta=0) - [1-\hat{\xi}(x,a)]\} \\
&+ \frac{1(A=a,\Delta=1)}{\hat{\pi}_a(x)}\{\tilde{T} - \hat{\nu}(1,x,a)\} + \frac{\hat{\nu}(1,x,a)1(A=a)}{\hat{\pi}_a(x)}\{1(\Delta=1) - \hat{\xi}(x,a)\} \\
&+ \hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)] + \hat{\nu}(1,x,a)\hat{\xi}(x,a) \\
&+ \gamma(x,a)\frac{1(A=a)}{\hat{\pi}_a(x)}\left\{1(\Delta=1) - \hat{\xi}(x,a)\right\} + \gamma(x,a)\hat{\xi}(x,a) \\
=&\frac{1(A=a,\Delta=0)}{\hat{\pi}_a(x)}\tilde{T} - \frac{1(A=a)}{\hat{\pi}_a(x)}\hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)] + \hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)] \\
&+ \frac{1(A=a,\Delta=1)}{\hat{\pi}_a(x)}\tilde{T} - \frac{1(A=a)}{\hat{\pi}_a(x)}\hat{\nu}(1,x,a)\hat{\xi}(x,a) + \hat{\nu}(1,x,a)\hat{\xi}(x,a) \\
&+ \gamma(x,a)\frac{1(A=a,\Delta=1)}{\hat{\pi}_a(x)} - \gamma(x,a)\frac{1(A=a)}{\hat{\pi}_a(x)}\hat{\xi}(x,a) + \gamma(x,a)\hat{\xi}(x,a).
\end{aligned} \tag{43}$$

The proof works analogously to the lower bound. That is, we calculate the conditional expectation over the pseudo outcome:

$$
\begin{aligned}
&\mathbb{E}\left[\hat{\phi}^+(x,a) \mid X=x\right] \\
=&\mathbb{E}\left[\frac{1(A=a,\Delta=0)}{\hat{\pi}_a(x)}\tilde{T} \mid X=x\right] - \mathbb{E}\left[\frac{1(A=a)}{\hat{\pi}_a(x)}\hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)] \mid X=x\right] \\
&+\mathbb{E}\left[\hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)] \mid X=x\right] + \mathbb{E}\left[\frac{1(A=a,\Delta=1)}{\hat{\pi}_a(x)}\tilde{T} \mid X=x\right] \\
&-\mathbb{E}\left[\frac{1(A=a)}{\hat{\pi}_a(x)}\hat{\nu}(1,x,a)\hat{\xi}(x,a) \mid X=x\right] + \mathbb{E}\left[\hat{\nu}(1,x,a)\hat{\xi}(x,a) \mid X=x\right] \\
&+\mathbb{E}\left[\gamma(x,a)\frac{1(A=a,\Delta=1)}{\hat{\pi}_a(x)} \mid X=x\right] - \mathbb{E}\left[\gamma(x,a)\frac{1(A=a)}{\hat{\pi}_a(x)}\hat{\xi}(x,a) \mid X=x\right] \\
&+\mathbb{E}\left[\gamma(x,a)\hat{\xi}(x,a) \mid X=x\right] \\
=&\frac{\pi_a(x)}{\hat{\pi}_a(x)}[1-\xi(x,a)]\nu(0,x,a) - \frac{\pi_a(x)}{\hat{\pi}_a(x)}[1-\hat{\xi}(x,a)]\hat{\nu}(0,x,a) + \hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)] \\
&+\frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a)\nu(1,x,a) - \frac{\pi_a(x)}{\hat{\pi}_a(x)}\hat{\xi}(x,a)\hat{\nu}(1,x,a) + \hat{\nu}(1,x,a)\hat{\xi}(x,a) \\
&+\gamma(x,a)\frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a) - \gamma(x,a)\frac{\pi_a(x)}{\hat{\pi}_a(x)}\hat{\xi}(x,a) + \gamma(x,a)\hat{\xi}(x,a).
\end{aligned} \tag{44}
$$

Hence, $\hat{\nu}(\delta,x,a)=\nu(\delta,x,a)$, $\hat{\xi}(x,a)=\xi(x,a)$ and $\hat{\pi}_a(x)=\pi_a(x)$ implies

$$
\begin{aligned}
&\mathbb{E}\left[\hat{\phi}^+(x,a) \mid X=x\right] \\
=&\frac{\pi_a(x)}{\hat{\pi}_a(x)}[1-\xi(x,a)]\nu(0,x,a) - \frac{\pi_a(x)}{\hat{\pi}_a(x)}[1-\hat{\xi}(x,a)]\hat{\nu}(0,x,a) + \hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)] \\
&+\frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a)\nu(1,x,a) - \frac{\pi_a(x)}{\hat{\pi}_a(x)}\hat{\xi}(x,a)\hat{\nu}(1,x,a) + \hat{\nu}(1,x,a)\hat{\xi}(x,a) \\
&+\gamma(x,a)\frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a) - \gamma(x,a)\frac{\pi_a(x)}{\hat{\pi}_a(x)}\hat{\xi}(x,a) + \gamma(x,a)\hat{\xi}(x,a) \\
=&[1-\xi(x,a)]\nu(0,x,a) + \xi(x,a)\nu(1,x,a) + \gamma(x,a)\xi(x,a) \\
=&\mu^+(x,a).
\end{aligned} \tag{45}
$$

which proves the consistency.

Also, (1) under $\hat{\nu}(\delta,x,a)=\nu(\delta,x,a)$ and $\hat{\xi}(x,a)=\xi(x,a)$, this reduces to

$$
\begin{aligned}
&\mathbb{E}\left[\hat{\phi}^+(x,a) \mid X=x\right] \\
=&\frac{\pi_a(x)}{\hat{\pi}_a(x)}[1-\xi(x,a)]\nu(0,x,a) - \frac{\pi_a(x)}{\hat{\pi}_a(x)}[1-\hat{\xi}(x,a)]\hat{\nu}(0,x,a) + \hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)] \\
&+\frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a)\nu(1,x,a) - \frac{\pi_a(x)}{\hat{\pi}_a(x)}\hat{\xi}(x,a)\hat{\nu}(0,x,a) + \hat{\nu}(1,x,a)\hat{\xi}(x,a) \\
&+\gamma(x,a)\frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a) - \gamma(x,a)\frac{\pi_a(x)}{\hat{\pi}_a(x)}\hat{\xi}(x,a) + \gamma(x,a)\hat{\xi}(x,a) \\
=&\frac{\pi_a(x)}{\hat{\pi}_a(x)}[1-\xi(x,a)]\nu(0,x,a) - \frac{\pi_a(x)}{\hat{\pi}_a(x)}[1-\xi(x,a)]\nu(0,x,a) + \nu(0,x,a)[1-\xi(x,a)] \\
&+\frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a)\nu(1,x,a) - \frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a)\nu(0,x,a) + \nu(1,x,a)\xi(x,a) \\
&+\gamma(x,a)\frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a) - \gamma(x,a)\frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a) + \gamma(x,a)\xi(x,a) \\
=&[1-\xi(x,a)]\nu(0,x,a) + \xi(x,a)\nu(1,x,a) + \gamma(x,a)\xi(x,a) \\
=&\mu^+(x,a).
\end{aligned} \tag{46}
$$

(2) Under $\hat{\pi}_a(x) = \pi_a(x)$, this reduces to

$$
\begin{aligned}
&\mathbb{E}\left[\hat{\phi}^+(x,a) \mid X = x\right] \\
=&\frac{\pi_a(x)}{\hat{\pi}_a(x)}[1-\xi(x,a)]\nu(0,x,a) - \frac{\pi_a(x)}{\hat{\pi}_a(x)}[1-\hat{\xi}(x,a)]\hat{\nu}(0,x,a) + \hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)] \\
&+ \frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a)\nu(1,x,a) - \frac{\pi_a(x)}{\hat{\pi}_a(x)}\hat{\xi}(x,a)\hat{\nu}(0,x,a) + \hat{\nu}(1,x,a)\hat{\xi}(x,a) \\
&+ \gamma(x,a)\frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a) - \gamma(x,a)\frac{\pi_a(x)}{\hat{\pi}_a(x)}\hat{\xi}(x,a) + \gamma(x,a)\hat{\xi}(x,a) \\
=&\frac{\pi_a(x)}{\pi_a(x)}[1-\xi(x,a)]\nu(0,x,a) - \frac{\pi_a(x)}{\pi_a(x)}[1-\hat{\xi}(x,a)]\hat{\nu}(0,x,a) + \hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)] \\
&+ \frac{\pi_a(x)}{\pi_a(x)}\xi(x,a)\nu(1,x,a) - \frac{\pi_a(x)}{\pi_a(x)}\hat{\xi}(x,a)\hat{\nu}(1,x,a) + \hat{\nu}(1,x,a)\hat{\xi}(x,a) \\
&+ \gamma(x,a)\frac{\pi_a(x)}{\pi_a(x)}\xi(x,a) - \gamma(x,a)\frac{\pi_a(x)}{\pi_a(x)}\hat{\xi}(x,a) + \gamma(x,a)\hat{\xi}(x,a) \\
=&[1-\xi(x,a)]\nu(0,x,a) + \xi(x,a)\nu(1,x,a) + \gamma(x,a)\xi(x,a) \\
=&\mu^+(x,a).
\end{aligned}
\tag{47}
$$

Finally, the pseudo-outcomes of **Case** ② (*conservative upper bounds*). is defined in Eq. (13). We simplify the equation to

$$
\begin{aligned}
&\hat{\phi}^+(x,a) \\
=&\frac{1(A=a,\Delta=0)}{\hat{\pi}_a(x)}\{\tilde{T} - \hat{\nu}(0,x,a)\} + \frac{\hat{\nu}(0,x,a)1(A=a)}{\hat{\pi}_a(x)}\{1(\Delta=0) - [1-\hat{\xi}(x,a)]\} \\
&+ \hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)] + t_{\max}\frac{1(A=a)}{\hat{\pi}_a(x)}\left\{1(\Delta=1) - \hat{\xi}(x,a)\right\} + t_{\max}\hat{\xi}(x,a) \\
=&\frac{1(A=a,\Delta=0)}{\hat{\pi}_a(x)}\tilde{T} - \frac{1(A=a)}{\hat{\pi}_a(x)}\hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)] + \hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)] \\
&+ t_{\max}\frac{1(A=a,\Delta=1)}{\hat{\pi}_a(x)} - t_{\max}\frac{1(A=a)}{\hat{\pi}_a(x)}\hat{\xi}(x,a) + t_{\max}\hat{\xi}(x,a).
\end{aligned}
\tag{48}
$$

Again, by taking expectation conditional on $X = x$, $A = a$, we obtain

$$
\begin{aligned}
&\mathbb{E}[\hat{\phi}^+(x,a) \mid X = x] \\
=&\frac{\pi_a(x)}{\hat{\pi}_a(x)}[1-\xi(x,a)]\nu(0,x,a) - \frac{\pi_a(x)}{\hat{\pi}_a(x)}[1-\hat{\xi}(x,a)]\hat{\nu}(0,x,a) + \hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)] \\
&+ t_{\max}\frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a) - t_{\max}\frac{\pi_a(x)}{\hat{\pi}_a(x)}\hat{\xi}(x,a) + t_{\max}\hat{\xi}(x,a).
\end{aligned}
\tag{49}
$$

Hence, using the fact that $\hat{\nu}(\delta,x,a) = \nu(\delta,x,a)$, $\hat{\xi}(x,a) = \xi(x,a)$ and $\hat{\pi}_a(x) = \pi_a(x)$ implies

$$
\begin{aligned}
&\mathbb{E}\left[\hat{\phi}^+(x,a) \mid X = x\right] \\
=&[1-\xi(x,a)]\nu(0,x,a) - [1-\xi(x,a)]\nu(0,x,a) + \nu(0,x,a)[1-\xi(x,a)] \\
&+ t_{\max}\xi(x,a) - t_{\max}\xi(x,a) + t_{\max}\xi(x,a) \\
=&[1-\xi(x,a)]\nu(0,x,a) + t_{\max}\xi(x,a) \\
=&\mu^+(x,a),
\end{aligned}
\tag{50}
$$

which proves consistency.

Again, (1) under $\hat{\nu}(\delta, x, a) = \nu(\delta, x, a)$ and $\hat{\xi}(x, a) = \xi(x, a)$, this reduces to

$$
\begin{aligned}
&\mathbb{E}[\hat{\phi}^+(x,a) \mid X = x] \\
=&\frac{\pi_a(x)}{\hat{\pi}_a(x)}[1 - \xi(x,a)]\nu(0,x,a) - \frac{\pi_a(x)}{\hat{\pi}_a(x)}[1 - \hat{\xi}(x,a)]\hat{\nu}(0,x,a) + \hat{\nu}(0,x,a)[1 - \hat{\xi}(x,a)] \\
&+ t_{\max}\frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a) - t_{\max}\frac{\pi_a(x)}{\hat{\pi}_a(x)}\hat{\xi}(x,a) + t_{\max}\hat{\xi}(x,a) \\
=&\frac{\pi_a(x)}{\hat{\pi}_a(x)}[1 - \xi(x,a)]\nu(0,x,a) - \frac{\pi_a(x)}{\hat{\pi}_a(x)}[1 - \xi(x,a)]\nu(0,x,a) + \nu(0,x,a)[1 - \xi(x,a)] \\
&+ t_{\max}\frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a) - t_{\max}\frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a) + t_{\max}\xi(x,a) \\
=&\hat{\nu}(0,x,a)[1 - \hat{\xi}(x,a) + t_{\max}\xi(x,a) \\
=&\mu^+(x,a),
\end{aligned}
\tag{51}
$$

(2) Under $\hat{\pi}_a(x) = \pi_a(x)$, this reduces to

$$
\begin{aligned}
&\mathbb{E}[\hat{\phi}^+(x,a) \mid X = x] \\
=&\frac{\pi_a(x)}{\hat{\pi}_a(x)}[1 - \xi(x,a)]\nu(0,x,a) - \frac{\pi_a(x)}{\hat{\pi}_a(x)}[1 - \hat{\xi}(x,a)]\hat{\nu}(0,x,a) + \hat{\nu}(0,x,a)[1 - \hat{\xi}(x,a)] \\
&+ t_{\max}\frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a) - t_{\max}\frac{\pi_a(x)}{\hat{\pi}_a(x)}\hat{\xi}(x,a) + t_{\max}\hat{\xi}(x,a) \\
=&\frac{\pi_a(x)}{\pi_a(x)}[1 - \xi(x,a)]\nu(0,x,a) - \frac{\pi_a(x)}{\pi_a(x)}[1 - \hat{\xi}(x,a)]\hat{\nu}(0,x,a) + \hat{\nu}(0,x,a)[1 - \hat{\xi}(x,a)] \\
&+ t_{\max}\frac{\pi_a(x)}{\pi_a(x)}\xi(x,a) - t_{\max}\frac{\pi_a(x)}{\pi_a(x)}\hat{\xi}(x,a) + t_{\max}\hat{\xi}(x,a) \\
=&\nu(0,x,a)[1 - \xi(x,a) + t_{\max}\xi(x,a) \\
=&\mu^+(x,a).
\end{aligned}
\tag{52}
$$

This proves double robustness again. □

D.6 Proof of Theorem 6.4

Lemma D.1. *Consider the setting described in Theorem 6.4. Then, we have*

$$\begin{aligned}&\mathbb{E}\left[(\hat{\mu}^+(x,a)-\mu^+(x,a))^2 \mid X=x\right]\\&\lesssim \mathcal{R}(x)+\mathbb{E}\left[(\hat{\pi}_a(x)-\pi_a(x))^2\right]\left(\mathbb{E}\left[(\hat{\nu}(0,x,a)-\nu(0,x,a))^2\right]+\mathbb{E}\left[(\hat{\nu}(1,x,a)-\nu(1,x,a))^2\right]\right.\\&\left.+\mathbb{E}\left[\left(\hat{\xi}(x,a)-\xi(x,a)\right)^2\right]\right).\end{aligned} \tag{53}$$

Proof. Let $\mu^+(x,a)$ be the corresponding oracle to $\hat{\mu}^+(x,a)$. Further ,we define $\tilde{\mu}^+(x,a) = \hat{\mathbb{E}}_n[\phi^+(x,a) \mid X=x]$. Using the assumption, we can apply Proposition 1 of [47] and obtain

$$\mathbb{E}\left[(\hat{\mu}^+(x,a)-\mu^+(x,a))^2 \mid X=x\right] \lesssim \mathcal{R}(x,a)+\mathbb{E}\left[\hat{r}(x,a)^2\right], \tag{54}$$

where $\mathcal{R}(x,a)=\mathbb{E}\left[(\mu^+(x,a)-\tilde{\mu}^+(x,a))^2\right]$ is the oracle risk of the second stage regression. We can apply Eq. (49) in the proof of Theorem 6.1 Section D.5 to obtain the following.

Case (1): domain-knowledge upper bound.

$$\begin{aligned}&\hat{r}(x,a)\\=&\frac{\pi_a(x)}{\hat{\pi}_a(x)}[1-\xi(x,a)]\nu(0,x,a)-\frac{\pi_a(x)}{\hat{\pi}_a(x)}[1-\hat{\xi}(x,a)]\hat{\nu}(0,x,a)+\hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)]\\&+\frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a)\nu(1,x,a)-\frac{\pi_a(x)}{\hat{\pi}_a(x)}\hat{\xi}(x,a)\hat{\nu}(0,x,a)+\hat{\nu}(1,x,a)\hat{\xi}(x,a)\\&+\gamma(x,a)\frac{\pi_a(x)}{\hat{\pi}_a(x)}\xi(x,a)-\gamma(x,a)\frac{\pi_a(x)}{\hat{\pi}_a(x)}\hat{\xi}(x,a)+\gamma(x,a)\hat{\xi}(x,a)-\mu^+(x,a)\\=&\left\{\frac{\pi_a(x)}{\hat{\pi}_a(x)}-1\right\}\left\{\nu(0,x,a)\left[\hat{\xi}(x,a)-\xi(x,a)\right]+[1-\hat{\xi}(x,a)]\left[\nu(0,x,a)-\hat{\nu}(0,x,a)\right]\right\}\\&+\left\{\frac{\pi_a(x)}{\hat{\pi}_a(x)}-1\right\}\left\{\nu(1,x,a)\left[\xi(x,a)-\hat{\xi}(x,a)\right]+\hat{\xi}(x,a)\left[\nu(1,x,a)-\hat{\nu}(1,x,a)\right]\right\}\\&+\left\{\frac{\pi_a(x)}{\hat{\pi}_a(x)}-1\right\}\gamma(x,a)\left[\xi(x,a)-\hat{\xi}(x,a)\right]\\=&\left\{\frac{\pi_a(x)}{\hat{\pi}_a(x)}-1\right\}\left\{[\nu(1,x,a)-\gamma(x,a)-\nu(0,x,a)]\left[\hat{\xi}(x,a)-\xi(x,a)\right]\right.\\&\left.+[1-\hat{\xi}(x,a)]\left[\nu(0,x,a)-\hat{\nu}(0,x,a)\right]+\hat{\xi}(x,a)[\hat{\nu}(1,x,a)-\nu(1,x,a)]\right\}.\end{aligned} \tag{55}$$

Applying the inequality $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ together with Assumption 6.3 from and the fact that $\pi_a(x) \le 1$ yields

$$\begin{aligned}\hat{r}(x,a)^2=&\left\{\frac{\pi_a(x)}{\hat{\pi}_a(x)}-1\right\}^2\left\{[\nu(1,x,a)+\gamma(x,a)-\nu(0,x,a)]\left[\hat{\xi}(x,a)-\xi(x,a)\right]\right.\\&\left.+[1-\hat{\xi}(x,a)]\left[\nu(0,x,a)-\hat{\nu}(0,x,a)\right]+\hat{\xi}(x,a)[\hat{\nu}(1,x,a)-\nu(1,x,a)]\right\}^2\\\le&3\left\{\frac{\pi_a(x)}{\hat{\pi}_a(x)}-1\right\}^2\left\{[\nu(1,x,a)+\gamma(x,a)-\nu(0,x,a)]^2\left[\hat{\xi}(x,a)-\xi(x,a)\right]^2\right.\\&\left.+[1-\hat{\xi}(x,a)]^2\left[\nu(0,x,a)-\hat{\nu}(0,x,a)\right]^2+\hat{\xi}(x,a)^2[\hat{\nu}(1,x,a)-\nu(1,x,a)]^2\right\}\\=&3\left\{\frac{\pi_a(x)}{\hat{\pi}_a(x)}-1\right\}^2[\nu(1,x,a)+\gamma(x,a)-\nu(0,x,a)]^2\left[\hat{\xi}(x,a)-\xi(x,a)\right]^2\\&+3\left\{\frac{\pi_a(x)}{\hat{\pi}_a(x)}-1\right\}^2[1-\hat{\xi}(x,a)]^2\left[\nu(0,x,a)-\hat{\nu}(0,x,a)\right]^2\\&+3\left\{\frac{\pi_a(x)}{\hat{\pi}_a(x)}-1\right\}^2\hat{\xi}(x,a)^2\left[\hat{\nu}(1,x,a)-\nu(1,x,a)\right]^2.\end{aligned} \tag{56}$$

By using Assumption 6.3, which gives $\hat{\pi}_a(x) > \varepsilon$ and thus $\left\{\frac{\pi_a(x)}{\hat{\pi}_a(x)} - 1\right\}^2 = \frac{(\pi_a(x)-\hat{\pi}_a(x))^2}{\hat{\pi}_a(x)^2} \leq \frac{1}{\varepsilon^2}(\pi_a(x) - \hat{\pi}_a(x))^2$, together with the boundedness of ν, ξ, and γ, we obtain

$$\begin{aligned}\hat{r}(x,a)^2 \leq & \frac{3(2C + t_{\max})^2}{\varepsilon^2} \left[\hat{\pi}_a(x) - \pi_a(x)\right]^2 \left[\hat{\xi}(x,a) - \xi(x,a)\right]^2 \\ & + \frac{3}{\varepsilon^2} \left[\hat{\pi}_a(x) - \pi_a(x)\right]^2 \left[\nu(0,x,a) - \hat{\nu}(0,x,a)\right]^2 \\ & + \frac{3}{\varepsilon^2} \left[\hat{\pi}_a(x) - \pi_a(x)\right]^2 \left[\hat{\nu}(1,x,a) - \nu(1,x,a)\right]^2 .\end{aligned} \tag{57}$$

Applying expectations on both sides yields the result, because $\hat{\pi}_a(x) \perp\!\!\!\perp (\hat{\nu}(\delta,x,a), \hat{\xi}(x,a))$ due to sample splitting.

□

E Extensions

E.1 Setting with continuous treatments

In many applications, treatments are not discrete but continuous (e.g., dosage levels, duration of therapy). Below, we extend our SurvB-learner to settings with continuous treatments where $A \in \mathcal{A} \subseteq \mathbb{R}$.

The key technical issue is the following. In the discrete case, we use the indicator $\mathbb{1}(A = a)$ weighted by the propensity score $\pi_a(x)$. For continuous treatments, the treatment variable A has a density with respect to the Lebesgue measure, so the point-mass indicator $\mathbb{1}(A = a)$ has probability zero and dividing by the generalized propensity score $f_{A|X}(a \mid x)$ no longer yields a well-behaved influence function. Our solution is to replace $\mathbb{1}(A = a)$ with a smooth, symmetric kernel function $K_h(A - a)$, following standard practice in semiparametric estimation for continuous treatment [36, 11]. This kernel smoothing localizes the pseudo-outcome around the target treatment value a and ensures stable estimation.

Our solution to overcome this is the following. We approximate the Dirac delta $\delta(\cdot)$ with a smooth, symmetric kernel function $K_h(t)$. Thereby, we follow standard practice in semiparametric estimation for continuous treatments [36, 11]. This kernel smoothing makes the localization well-behaved and ensures stable estimation.

With the above modification, our partial identification framework, including the domain-knowledge upper bounds, extends naturally to continuous treatments, as shown in Eq. (58). Formally, the pseudo-outcome under a continuous treatment setting is defined via

$$\begin{aligned}
&\hat{\phi}^{+}(x,a) \\
=&\frac{1(\Delta=0)K_h(A-a)}{\hat{f}_{A|X}(a \mid x)}\{\tilde{T}-\hat{\nu}(0,x,a)\}+\frac{\hat{\nu}(0,x,a)K_h(A-a)}{\hat{f}_{A|X}(a \mid x)}\{1(\Delta=0)-[1-\hat{\xi}(x,a)]\}+\hat{\nu}(0,x,a)[1-\hat{\xi}(x,a)] \\
&+\frac{1(\Delta=1)K_h(A-a)}{\hat{f}_{A|X}(a \mid x)}\{\tilde{T}-\hat{\nu}(1,x,a)\}+\frac{\hat{\nu}(1,x,a)K_h(A-a)}{\hat{f}_{A|X}(a \mid x)}\{1(\Delta=1)-\hat{\xi}(x,a)\}+\hat{\nu}(1,x,a)\hat{\xi}(x,a) \\
&+\gamma(x,a)\frac{K_h(A-a)}{\hat{f}_{A|X}(a \mid x)}\Big\{1(\Delta=1)-\hat{\xi}(x,a)\Big\}+\gamma(x,a)\hat{\xi}(x,a),
\end{aligned} \tag{58}$$

where $\hat{f}_{A|X}(a \mid x)$ is an estimator of $f_{A|X}(a \mid x)$, which is the generalized propensity score and the $K_h(A-a) = k(\frac{A-a}{h})/h$ is the kernel with standard second-order kernel function $k(\cdot)$ and bandwidth h. As a result, our SurvB-learner meta-learner for the continuous setting continues to enjoy the desirable properties established in the discrete setting – most notably, consistency and double robustness – which ensures that the bounds are reliably estimated even under model misspecification.

E.2 Hidden confounding

In this section, we provide an extension for observational data (e.g., patient registries, electronic health records) that can be subject to hidden confounding. To this end, we consider data with hidden confounding U, where the full population is $(X, A, T, C, U) \sim \mathbb{P}$. For simplicity, we consider the binary treatment setting in the following derivation, i.e., $A \in \{0, 1\}$. Given the setting, we are interested in the lower bound in Eq. (4) and the upper bound in Eq. (6) of the form

$$\begin{aligned} \mathcal{Q}_{\mathrm{LB}}(x, a, \mathcal{M}) &= \mathbb{E}_{\mathbb{P}(\cdot \mid x, \mathrm{do}(A=a))}[\tilde{T}], \\ \mathcal{Q}_{\mathrm{UB}}(x, a, \mathcal{M}) &= \mathbb{E}_{\mathbb{P}(\cdot \mid x, \mathrm{do}(A=a))}[\tilde{T}] + \gamma(x, a) \cdot \mathbb{P}(\Delta = 1 \mid x, \mathrm{do}(A = a)). \end{aligned} \tag{59}$$

Definition E.1. (SCM) Let $\mathcal{V} = \{V_1, \ldots, V_n\}$ denote a set of observed endogenous variables, let $\mathcal{U} \sim P_{\mathcal{U}}$ denote a set of unobserved exogenous variables, and let $\mathcal{F} = \{f_{V_1}, \ldots, f_{V_n}\}$ with $f_{V_i} : \mathrm{Pa}(V_i) \subseteq \mathcal{V} \cup \mathcal{U} \to V_i$. The tuple $(\mathcal{V}, \mathcal{U}, \mathcal{F}, P_{\mathcal{U}})$ is called a structural causal model (SCM).

We assume a directed graph $\mathcal{G}_{\mathcal{C}}$ induced by SCM $\mathcal{C}$ to be acyclic.

Definition E.2. (Generalized marginal sensitivity model (GMSM)). Let $\mathbf{V} = \{\mathbf{X}, \mathbf{A}, \mathbf{T}, \mathbf{C}\}$. Let $\mathbf{X}$, $\mathbf{A}$, $\mathbf{T}$, and $\mathbf{C}$ denote a set of observed endogenous variables, $\mathbf{U}$ a set of unobserved exogenous variables, and $\mathcal{G}$ a causal directed acyclic graph (DAG) on $\mathbf{V} \bigcup \mathbf{U}$. For an observational distribution $P_{\mathbf{V}}$ on $\mathbf{V}$ and a family $\mathcal{P}$ of joint probability distribution on $\mathbf{V} \bigcup \mathbf{U}$ that satisfy

$$\frac{1}{(1-\Gamma)\mathbb{P}(a \mid x) + \Gamma} \leq \frac{P(U = u \mid x, a)}{P(U = u \mid x, \mathrm{do}(A = a))} \leq \frac{1}{(1-\Gamma^{-1})\mathbb{P}(a \mid x) + \Gamma^{-1}}. \tag{60}$$

Using the definition above, we aim to obtain bounds $\mathcal{Q}^-(x, a, \mathcal{S}) \leq \mathcal{Q}^+(x, a, \mathcal{S})$ so that

$$\begin{aligned} \mathcal{Q}^-_{\mathrm{LB}}(x, a, \mathcal{S}) &= \inf_{\mathcal{M} \in \mathcal{C}(\mathcal{S})} \mathcal{Q}_{\mathrm{LB}}(x, a, \mathcal{M}) \text{ and } \mathcal{Q}^+_{\mathrm{LB}}(x, a, \mathcal{S}) = \sup_{\mathcal{M} \in \mathcal{C}(\mathcal{S})} \mathcal{Q}_{\mathrm{LB}}(x, a, \mathcal{M}) \\ \mathcal{Q}^-_{\mathrm{UB}}(x, a, \mathcal{S}) &= \inf_{\mathcal{M} \in \mathcal{C}(\mathcal{S})} \mathcal{Q}_{\mathrm{UB}}(x, a, \mathcal{M}) \text{ and } \mathcal{Q}^+_{\mathrm{UB}}(x, a, \mathcal{S}) = \sup_{\mathcal{M} \in \mathcal{C}(\mathcal{S})} \mathcal{Q}_{\mathrm{UB}}(x, a, \mathcal{M}). \end{aligned} \tag{61}$$

Assumption E.3. We assume the conditional unconfoundedness, which is $(T(a), C(a)) \perp\!\!\!\perp A \mid X, U$, $\forall a$.

Theorem E.4. *Let $\mathcal{S}$ be a GMSM with bounds $s^-_W = \frac{1}{(1-\Gamma)\mathbb{P}(a|x)+\Gamma}$ and $s^+_W = \frac{1}{(1-\Gamma^{-1})\mathbb{P}(a|x)+\Gamma^{-1}}$, for $W \in \{T, C\}$. We define $c^+_W = \frac{(1-s^-_W)s^+_W}{s^+_W - s^-_W}$. Since $W \in \{T, C\} \in \mathbb{R}$ is continuous, we define the probability density function*

$$\mathbb{P}_+(w \mid x, a) = \begin{cases} \dfrac{1}{s^+_W}\, \mathbb{P}(w \mid x, a), & \text{if } F(w) \leq c^+_W, \\ \dfrac{1}{s^-_W}\, \mathbb{P}(w \mid x, a), & \text{if } F(w) > c^+_W. \end{cases} \tag{62}$$

where $F_+(\cdot)$ denote the conditional CDF corresponding to $\mathbb{P}_+(w \mid x, a)$. Further let $F_-(\cdot)$ be the conditional CDF corresponding to $\mathbb{P}_-(w \mid x, a)$, which is defined by swapping signs (in c_W and s_W). For any SCM $\mathcal{M}$, we denote the CDF corresponding to $\mathbb{P}_{\mathcal{M}}(\cdot \mid x, \mathrm{do}(A = a))$ by $F_{\mathcal{M}}(\cdot)$. Then, for all w, we have

$$F_+(w) \leq \inf_{\mathcal{M} \in \mathcal{C}(\mathcal{S})} F_{\mathcal{M}}(w) \text{ and } F_-(w) \geq \sup_{\mathcal{M} \in \mathcal{C}(\mathcal{S})} F_{\mathcal{M}}(w). \tag{63}$$

Assume now that $\mathbb{P}(u \mid x, \mathrm{do}(A = a)) = \mathbb{P}(u \mid x)$. Then, the bounds are sharp (i.e., equality holds in Eq. (63) whenever A is discrete and it holds that $1/s^+_W \geq \mathbb{P}(a \mid x)$.

Proof. We adopt the proof of Theorem 1 from Frauen et al. [22]. □

We now leverage Theorem E.4 to obtain explicit solutions of our lower bounds and upper bounds.

Corollary E.5. *(Bounds with hidden confounding) When the decision function $\mathcal{D}$ induced by the GMSM $\mathcal{S}$ is monotone, we obtain sharp bounds, i.e.,*

$$\begin{aligned} \mathcal{Q}^-_{\mathrm{LB}} &= \mathbb{E}_{\mathbb{P}_-}[\tilde{T} \mid x, a], \\ \mathcal{Q}^+_{\mathrm{UB}} &= \mathbb{E}_{\mathbb{P}_+}[\tilde{T} \mid x, a] + \gamma(x, a) \cdot \mathbb{P}_+(\Delta = 1 \mid x, a). \end{aligned} \tag{64}$$

Proof. We adopt the proof of Corollary 1 from Frauen et al. [22]. □

F Details for the synthetic datasets

Data-generating process: For our synthetic dataset, we simulate an observed confounder $X \sim \mathrm{Uniform}[10, 100]$ to mimic the patients' age, and a latent frailty variable $U \sim \mathcal{N}(0, 1)$, which represents unobserved disease severity and is not available to the learner. We define the propensity score in the randomized control trial as $\pi_a(x) = \mathbb{P}(A = 1 \mid X = x) = 0.5$, and in observational dataset experiments as $\pi_a(x) = \sigma\big(\frac{x-45}{45}\big)$. Next, we specify the treatment effect function $\tau(x, a)$:

- Exponential function: $\tau(x, a) = 20 \cdot \exp(a + 0.01x) + \varepsilon$
- Sin function: $\tau(x, a) = \big(\sin\big(\frac{x-10}{90} \cdot 2\pi\big) + 1.2\big) \cdot 10 \cdot a + x + \varepsilon$
- Logistic-sin function: $\tau(x, a) = \frac{30}{1+\exp(-0.1(x-50))} + 5\sin(0.2x) + 10 + \varepsilon$

where $\varepsilon \sim \mathcal{N}(0, 0.1)$. The survival time is

$$T = \tau(x, a) \cdot a + \tfrac{1}{3}\big(\sin(12x) + x\big) + \tfrac{1}{60}\cos(20x) - \beta_T \cdot U + \varepsilon, \tag{65}$$

where $\beta_T = 1.0$, so that sicker patients (larger U) have shorter survival times. The censoring time is generated as

$$C = c_0 \cdot \exp\big(-\beta_C \cdot U + \eta\big), \quad \eta \sim \mathcal{N}(0, 0.1), \tag{66}$$

where $\beta_C = 1.0$ and c_0 is chosen to achieve the target censoring rate $\xi \in \{0.2, 0.4, 0.6\}$. The observed outcome and censoring indicator are then

$$\tilde{T} = \min\{T, C\}, \qquad \Delta = \mathbb{1}(C \le T). \tag{67}$$

By construction, U simultaneously shortens T and C, inducing $C \not\perp T \mid X, A$, which constitutes genuine informative censoring. We generate 2000 samples for each setting and random seed.

G Additional experimental results

G.1 Additional experiments with point estimators

We implement the method from Rubin and van der Laan [67], which extends the doubly robust censoring unbiased transformation to CATE estimation via a second-stage regression on the corresponding pseudo-outcomes. This baseline shares the same target estimand as our framework but assumes non-informative censoring, i.e., $C \perp T \mid X, A$, which is violated in our setting. We therefore use it as a point-estimation baseline to empirically demonstrate the bias induced by ignoring informative censoring.

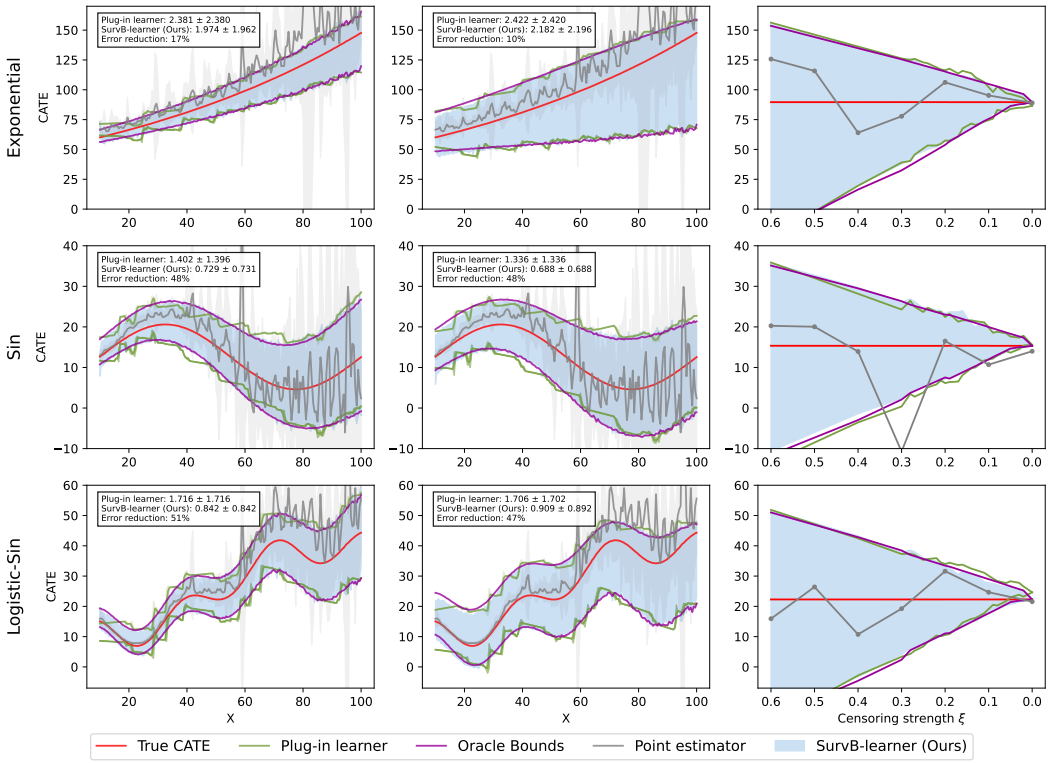

Figure 4: **Results for experiments with synthetic data including the point estimation baseline**.

Results are shown in Fig. 4. The comparison exposes a fundamental limitation of point estimation under informative censoring. As $\xi(x, a)$ grows (decided by the DGP in appendix F), the point estimator increasingly deviates from the true CATE and eventually falls entirely outside our bounds, confirming that ignoring informative censoring produces invalid estimates in high-censoring regimes.

When $\xi(x, a)$ is small, the point estimator performs comparably to our SurvB-learner, consistent with our theoretical result that the width of CAPO bounds $\gamma(x, a)\xi(x, a) \to 0$ as $\xi \to 0$, and our bounds recover the point estimation. And in this regime, even conservative bounds without domain knowledge are informative. Together, these observations confirm that our SurvB-learner is strictly more general: it degrades to point estimation when censoring is weak, while remaining valid when censoring is strong and point estimation fails.

G.2 Empirical coverage and smoothness of estimated bounds

We further evaluate our SurvB-learner against the plug-in learner using empirical coverage and total variation (TV). Coverage measures the proportion of evaluation points where the estimated bounds

contain the true CATE:

$$\widehat{\text{Coverage}} = \frac{1}{RM} \sum_{r=1}^{R} \sum_{j=1}^{M} \mathbb{1}\left\{\hat{\tau}_r^-(x_j) \leq \tau_r(x_j) \leq \hat{\tau}_r^+(x_j)\right\}, \tag{68}$$

where R is the number of simulation seeds and M is the number of evaluation points.

Bounds	Function	SurvB-learner	Plug-in learner
Domain knowledge	exponential	**100.00%**	97.45%
	sin	**100.00%**	99.75%
	logistic-sin	**100.00%**	97.15%
Conservative	exponential	**100.00%**	100.00%
	sin	**100.00%**	99.95%
	logistic-sin	**100.00%**	99.95%

Table 2: **Empirical coverage of estimated CATE bounds.** Higher is better.

Table 2 shows that our SurvB-learner achieves 100% empirical coverage across all settings, whereas the plug-in learner fails to achieve complete coverage in several cases.

We further measure the smoothness of estimated bounds using total variation:

$$\text{TV}(f) = \sum_{j=1}^{M-1} |f(x_{j+1}) - f(x_j)|, \tag{69}$$

where $\{x_1, \ldots, x_M\}$ is a fixed ordered evaluation grid. For each simulation seed, we sum the TV of the lower and upper bound functions and report the mean and standard deviation across seeds. Values closer to the oracle TV indicate better recovery of the underlying variation.

Bounds	Function	Oracle TV	SurvB-learner TV	Plug-in TV
Domain knowledge	exponential	78.2 ± 0.5	$\mathbf{84.0 \pm 2.7}$	89.4 ± 4.0
	sin	31.2 ± 0.1	$\mathbf{33.1 \pm 0.7}$	39.2 ± 1.6
	logistic-sin	60.2 ± 0.4	$\mathbf{61.0 \pm 1.8}$	63.8 ± 2.1
Conservative	exponential	49.4 ± 0.6	$\mathbf{65.1 \pm 2.6}$	80.8 ± 5.5
	sin	26.7 ± 0.1	$\mathbf{28.1 \pm 0.7}$	37.6 ± 2.5
	logistic-sin	50.7 ± 0.2	$\mathbf{52.7 \pm 1.2}$	57.1 ± 1.6

Table 3: **Total variation of estimated CATE bounds.** Mean $\pm$ standard deviation across simulation seeds. Values closer to the oracle TV are preferred.

As shown in Table 3, our SurvB-learner consistently achieves TV closer to the oracle bounds and lower variability across simulation seeds than the plug-in learner. Together, these results demonstrate better empirical coverage and more stable bound estimation, consistent with Fig. 2.

G.3 Sensitivity analysis on the specification of the post-dropout survival function

We investigate the sensitivity of our bounds to the specification of $\gamma(x, a)$. In our synthetic DGPs, the true expected post-dropout survival time is available as $\gamma^\star(x, a) = \mathbb{E}[T - \tilde{T} \mid X = x, A = a, \Delta = 1]$. We set $\gamma_\rho(x, a) = \rho\gamma^\star(x, a)$ for both treatment arms and vary ρ, including values that underestimate ($\rho < 1$) or overestimate ($\rho > 1$) the true post-dropout survival. We additionally consider the conservative specification. Oracle nuisance functions are used to isolate the effect of misspecifying $\gamma(x, a)$.

Table 4 shows that underestimating $\gamma(x, a)$ produces narrower bounds but reduces coverage, while $\rho \geq 1$ achieves full coverage in our experiments at a linear cost in bound width. This follows from Proposition 5.2, where the CATE bound width is the sum of $\gamma(x, a)\xi(x, a)$ across both arms. Hence, $\gamma(x, a)$ should be chosen as a plausible upper bound rather than a best estimate of post-dropout survival. When domain knowledge is unavailable, the conservative specification provides a valid alternative.

Metric	$\rho = 0.5$	$\rho = 0.8$	$\rho = 1.0$	$\rho = 1.2$	$\rho = 1.5$	$\rho = 1.7$	Conservative
Coverage	0.5874	0.8252	1.0000	1.0000	1.0000	1.0000	1.0000
Average width	7.700	12.320	15.400	18.480	23.100	26.180	35.344

Table 4: **Sensitivity analysis of $\gamma(x, a)$ on the Sin DGP.** Coverage and average CATE bound width for different scaling factors ρ. Values $\rho < 1$ violate Eq. (5).

G.4 Informativeness of conservative bounds under varying censoring strengths

We examine whether conservative bounds remain informative by stratifying evaluation points in the Sin DGP according to the arm-average censoring strength $\bar{\xi}(x) = [\xi(x, a_1) + \xi(x, a_2)]/2$. We report the average conservative bound width and empirical coverage within each stratum.

Metric	$\xi < 0.05$	$0.05 \leq \xi < 0.2$	$0.2 \leq \xi < 0.3$	$\xi > 0.3$
Conservative width	7.156	27.141	44.428	52.233
Conservative coverage	1.0000	1.0000	1.0000	1.0000

Table 5: Conservative CATE bound width and coverage across censoring-strength strata.

Table 5 shows that conservative bounds maintain full coverage while their width increases with censoring strength. This follows from

$$\tau^{+}_{a_1,a_2}(x) - \tau^{-}_{a_1,a_2}(x) = \sum_{a \in \{a_1,a_2\}} [t_{\max} - \nu(1, x, a)]\xi(x, a). \tag{70}$$

As censoring vanishes in both arms, the bound width approaches zero. Thus, conservative bounds remain relatively narrow in low-censoring subgroups, while wider bounds reflect greater identification uncertainty in heavily censored subgroups.

G.5 Comparison with existing causal survival estimators under informative censoring

We compare our SurvB-learner with three point-estimation methods from SurvHTE-Bench: the doubly robust (DR) point estimator, integrated causal survival forest (Integrated-CSF), and hazard-EIF meta-learner. Unlike our SurvB-learner, which estimates CATE bounds under informative censoring, these baselines target point estimates under non-informative censoring. We evaluate all methods on three synthetic DGPs with informative censoring and report ATE estimates with 95% intervals for the baselines and domain-knowledge ATE bounds for our SurvB-learner and the plug-in learner.

Method	sin	logistic-sin	exponential
True ATE	12.60	24.10	97.52
DR point estimator	11.64 ± 0.09	39.50 ± 6.12	101.18 ± 3.12
Integrated-CSF	12.29 ± 0.02	27.38 ± 0.04	99.20 ± 0.08
Hazard-EIF meta-learner	12.20 ± 0.02	27.40 ± 0.50	99.43 ± 0.34
SurvB-learner (LB, UB)	$[5.94 \pm 0.36,\ 21.03 \pm 0.38]$	$[18.97 \pm 0.37,\ 34.07 \pm 0.39]$	$[83.20 \pm 1.50,\ 108.35 \pm 1.66]$
Plug-in learner (LB, UB)	$[6.00 \pm 0.33,\ 21.09 \pm 0.32]$	$[19.06 \pm 0.36,\ 34.15 \pm 0.30]$	$[83.49 \pm 1.42,\ 108.64 \pm 1.38]$

Table 6: **Comparison with existing causal survival estimators under informative censoring.** Point estimates are reported with 95% intervals, while bounds are reported with uncertainty margins for each endpoint.

Table 6 shows that the reported 95% intervals of all three point estimation baselines fail to contain the true ATE across the DGPs, whereas both our SurvB-learner and the plug-in learner produce bounds containing the truth. This highlights the distinction between sampling uncertainty and identification uncertainty: narrow confidence intervals around point estimates do not account for potential bias from informative censoring, whereas our bounds explicitly account for this uncertainty through $\gamma(x, a)$.

H Implementation details

H.1 Experiments with synthetic data

For the synthetic experiments, we build on the implementation framework commonly used to evaluate meta-learners for point-identified CATE estimation [24]. We employ the random forests [6] for nuisance estimation in both the plug-in learner and our SurvB-learner. This ensures that performance differences arise solely from the meta-learners rather than from model choice. All learners were implemented in Python 3.11.12 using scikit-learn 1.5.2.

Random forests were fitted with the default hyperparameter settings and without additional tuning (i.e, n_estimators $= 100$, max_depth $=$ None, and min_samples_leaf $= 2$). For the implementation of the plug-in learner, we also use a random forest with default settings to estimate the nuisance functions, namely, $\nu(\delta, x, a)$ and $\xi(x, a)$. Then, we follow the Eq. (4), Eq. (6), and Eq. (8) in Section 4 directly to compute the bounds. For the implementation of our SurvB-learner, we apply three-fold cross-fitting, and also apply the random forests using default settings in the second stage regression.

For Domain knowledge bounds, we set $\gamma(X, a)$ is uniformaly over covariates and treatments, which make sense, since for synthetic data we are comapre the accuracy of SurvB-learner, instead of complexity. For exponential function, $\gamma = 50$; sin function, $\gamma = 30$, logistic-sin function, $\gamma = 30$.

Regarding the RMSE metric: We define the RMSE in Table 1 as the sum of the upper and lower bounds against their respective ground-truth values μ^+ and μ^-:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_i^{\pm} - \mu_i^{\pm}\right)^2}. \tag{71}$$

That is, the reported RMSE measures how accurately the estimated bounds recover the true bounds, not the deviation from a single point estimate.

H.2 ADJUVANT dataset

The ADJUVANT trial was designed to enroll patients with EGFR-mutant stage II–IIIA non-small cell lung cancer (NSCLC). However, in the available dataset, one patient was recorded as clinical stage I, and we therefore excluded this patient from our analysis.

Given the limited sample size of the ADJUVANT cohort ($N = 170$), we adopt three-fold cross-fitting. In the first stage, we estimate nuisance functions using random forests, which are nonparametric ensembles of decision trees [6]. In the second stage, we regress pseudo-outcomes on covariates and treatment with random forest regressors using the default settings to obtain the bound estimates. Following the ADJUVANT trial [99, 53], we set $t_{\max} = 50$ months for DFS and OS time.

For the data-driven analysis in Figure 5 **(C)**, we adapt the idea of regression tree partitioning [95], where we replace the standard squared-error splitting rule by a customized criterion that maximizes the percentage of lower bounds (LBs) above zero within each leaf state in Eq. 72. Hence, we have

$$\begin{aligned} p_L &= \frac{1}{|M_L|}\sum_{i\in M_L}\mathbf{1}\{\hat{\tau}^-_{\text{gefitinib, chemo}}(x_i) > 0\}, \\ p_R &= \frac{1}{|M_R|}\sum_{i\in M_R}\mathbf{1}\{\hat{\tau}^-_{\text{gefitinib, chemo}}(x_i) > 0\}, \\ \text{Gain}(M_L, M_R) &= \max\{p_L,\ p_R\}, \end{aligned} \tag{72}$$

where M_L, M_R are the set of samples that fall into the left/right leaf node under a candidate split. We set the maximum tree depth to 2 and require at least 2 samples per leaf node.

I Real-world dataset

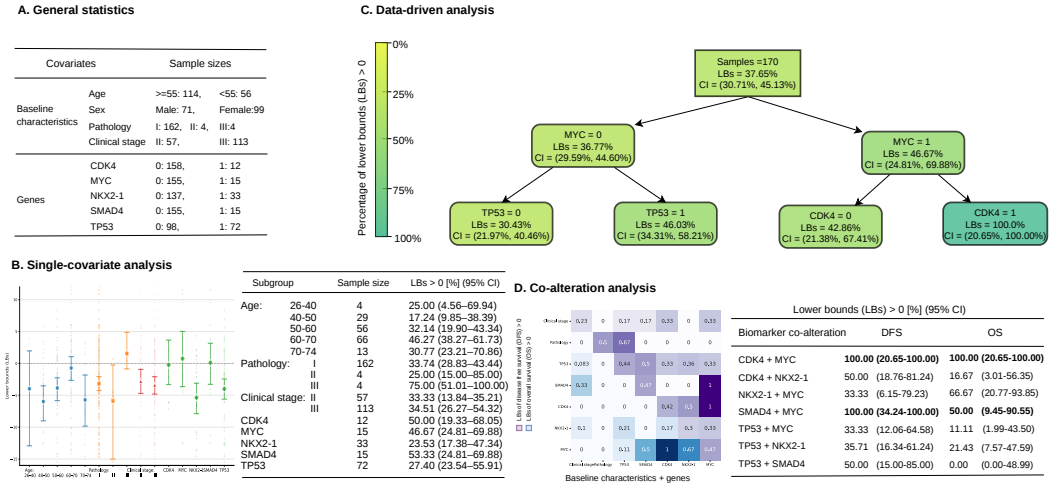

Figure 5: **Findings for the ADJUVANT trial.** The ADJUVANT trial [99, 53] assesses the effect of adjuvant gefitinib in $N = 170$ patients with EGFR-mutant stage II–IIIA non-small cell lung cancer (NSCLC), yet the original study was subject to large censoring (37.06% for DFS and 63.53% for OS). Hence, our aim is to identify subgroups of patients that have a robust treatment benefit despite potential informative censoring. To do so, we report the proportion of patients where the lower bounds (LBs) of the CATE are strictly greater than zero (i.e., "LBs ≥ 0"), expressed as percentages (%). So, a value close to 100% indicates that nearly all patients in this subgroup benefit from treatment, with a CATE above zero despite potential informative censoring. **(A)** Patient characteristics: The table on the left reports the baseline characteristics and genetic alterations of the patient population. **(B)** Single-covariate analysis: The forest plots show the mean and standard deviation of the estimated lower bounds of the CATE within subgroups defined by baseline characteristics or genetic covariates by bootstrap. For each subgroup on the x-axis, we performed 2000 bootstrap replications. The background scatter points represent the CATE lower bound estimates without bootstrapping. **(C)** Data-driven analysis: Using a tailored recursive partition algorithm, we select subgroups with the highest percentage of LBs above zero. Method details are in Appendix H. **(D)** Co-alteration analysis: The heatmap (left) reports the probabilities that the conservative lower bounds are above zero ("LBs ≥ 0") for pairs of genetic biomarkers, while the table (right, sorted alphabetically) summarizes these probabilities for both disease-free survival (DFS) and overall survival (OS) for selected genetic co-alterations. Hence, our framework provides robust evidence of treatment benefit for certain subgroups, even for the efficacy outcome (OS), despite the original study not being able to confirm this, and in the presence of informative censoring that could otherwise bias the results.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We state clearly in the introduction and abstract our contributions.

   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: We mentioned limitation in the end of main paper.

   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: We provide the proofs in the appendix.

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: We have implementation details in the appendix for the data generation process and models to use.

   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: All the details about the data generation process are in the appendix. And we provide an anonymous GitHub link in the paper.

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: It is all included in the implementation detail section.

   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: We use the RMSE to report errors with standard errors.

   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: [Yes]

   Justification: We put the computing source requirement in the implementation details.

   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: We use synthetic data and a publicly available clinical trial dataset (ADJUVANT).

   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: Our framework supports subgroup-level treatment effect quantification under informative censoring, which is directly relevant to personalized treatment decision-making in clinical practice.

    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: Our method is not related with pre-trained language models, image generators, or scraped datasets.

    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: All the codes and public clinical datasets are properly cited.

    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: [N/A]

    Justification: We do not introduce new assets. And the code will be made public upon acceptance.

    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: We do not include with 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: We do not include human subjects.

    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: LLM is only used for writing and formatting purposes.

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.