History-Aware Conformal Prediction Sets for Censored Time-to-Event Outcomes

Yuyao Wang
Department of Biostatistics
Brown University
Providence, RI 02912
yuyao_wang1@brown.edu

Alexander W. Levis
Department of Biostatistics, Epidemiology, and Informatics
University of Pennsylvania
Philadelphia, PA 19104
alexander.levis@pennmedicine.upenn.edu

Shu Yang
Department of Statistics
North Carolina State University
Raleigh, NC 27695
syang24@ncsu.edu

Larry Han
Department of Biostatistics
Brown University
Providence, RI 02912
larry_han@brown.edu

Abstract

Existing conformal prediction methods for time-to-event outcomes leverage only baseline covariates, producing prediction intervals that are insufficiently informative to facilitate decision making. We propose History-Aware Prediction Sets (HAPS), a conformal framework that constructs prediction sets for individual event times using covariate histories observed up to a decision time, targeting coverage among individuals who have survived to this time. HAPS handles right censoring adjusted for time-varying confounders via inverse probability of censoring weighting. When the censoring weights are consistently estimated, it achieves PAAC (probably asymptotically approximately correct) coverage among survivors. We further propose two doubly robust extensions of HAPS to weaken reliance on consistent estimation of the censoring distribution. In simulations, HAPS and its extensions reduce median prediction interval length by up to 75% relative to baseline comparators while maintaining close to nominal coverage. On two public benchmark data sets, HAPS reduces the median interval length by up to 60% for predictions at year 5, compared to the baseline comparators.

1 Introduction

Many real-world decisions involve predicting individual event times for those who are event-free at a prediction time τ (i.e., with event time $T > \tau$), using covariate histories observed up to τ. Modern machine learning methods can flexibly incorporate high-dimensional covariates to produce accurate point predictions of event times. However, in high-stakes settings, point predictions alone are insufficient; rigorous uncertainty quantification is needed for reliable decision-making (Banerji et al., 2023).

Conformal prediction (Vovk et al., 2005) is a general and model-agnostic framework for uncertainty quantification that can be applied with off-the-shelf prediction algorithms and has recently been extended to time-to-event outcomes under right censoring. Existing work has focused on constructing lower prediction bounds (LPBs) (Candès et al., 2023; Gui et al., 2024; Sesia and Svetnik, 2025; Davidov et al., 2025; Farina et al., 2026; Si and Qiu, 2025), with more recent efforts constructing mixed-type and two-sided prediction intervals (Holmes and Marandon, 2024; Qin et al., 2025; Yi et al., 2025). However, these methods are static in the sense that they perform calibration at the

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

time origin using only baseline covariates, focusing on marginal coverage guarantees over the entire population. Existing methods do not address the dynamic task of constructing prediction intervals at a decision time τ for individuals who have survived to τ, using their covariate history. The dynamic setting is more difficult because of the updated prediction target and the calibration problem. First, the appropriate notion of coverage must be updated to condition on survivors at τ, which we refer to as *survivor-conditional coverage*. Second, right censoring coarsens not only the event time but also the covariate trajectory updated over time.

We propose a conformal framework that constructs two-sided prediction intervals for individual event times among survivors at a prespecified decision time τ, based on time-varying covariate history observed up to τ. Methodologically, we formulate dynamic conformal calibration as estimation of an optimal scalar parameter that indexes a nested family of candidate sets at time τ. Under right censoring, we derive an observed-data estimating equation for the calibration parameter using inverse probability of censoring weighting (IPCW) (Robins and Rotnitzky, 1992; Rotnitzky and Robins, 2005), which maps the censoring-free calibration condition to observed data without requiring correct modeling of the event-time distribution. We then integrate these estimating equations with split conformal prediction and establish asymptotic survivor-conditional coverage at time τ. We further develop two doubly robust extensions of HAPS: one based on a doubly robust post-processing procedure and another based on augmented IPCW (AIPCW), to reduce reliance on consistent estimation of the censoring distribution.

Empirically, our simulations show that, in the presence of time-varying confounders for censoring, baseline conformal methods that use only baseline covariates to adjust for censoring can have biased coverage. By leveraging time-varying covariates, HAPS maintains close to nominal coverage while producing substantially narrower two-sided prediction intervals than baseline methods. On two public benchmark data sets from the `survival` R package (details provided in Section 6), HAPS reduces the median interval length by 9–60% for predictions at year 5, relative to baseline methods.

1.1 Related work

Conformal prediction for right censored data. Early developments on conformal prediction for time-to-event outcomes constructed LPBs under *type-I censoring* (Candès et al., 2023; Gui et al., 2024). Subsequent work has addressed the more common right censoring setting, by censoring imputation approaches that reduce the problem to the type-I censoring setting (Sesia and Svetnik, 2025), weighting approaches to handle right censoring (Davidov et al., 2025), and doubly robust methods based on semiparametric efficiency theory for right censoring (Farina et al., 2026; Si and Qiu, 2025). Recently, Holmes and Marandon (2024) proposed mixed-type prediction sets based on censoring status prediction, and two-sided prediction intervals have been proposed using inverse probability of censoring weighting (Yi et al., 2025) and bootstrap-based methods under working models (Qin et al., 2025). A detailed summary of the above literature is provided in Appendix A.

Despite these advances, existing conformal survival methods share a common limitation in that the target is a static prediction problem defined at the time origin using only baseline covariates. In many settings, baseline covariates only explain a small proportion of variation for time-to-event outcomes, which can be distal, leading to prediction sets that are too wide to be useful in practice (Henderson et al., 2001). Figure 1 illustrates this limitation using two benchmark simulation setups from Gui et al. (2024) and Farina et al. (2026). Here, even when oracle quantiles of the conditional event time distribution are used, the resulting 90% LPBs are close to zero.

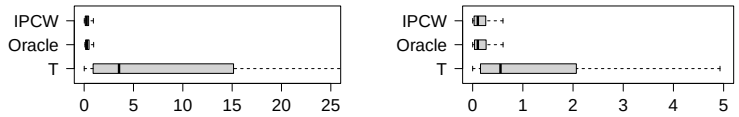


Figure 1: Intrinsic limitation of static approaches to construct LPBs. Boxplots of 90% LPBs from the IPCW approach in Farina et al. (2026) with random survival forests, the oracle 0.1 quantile of the conditional event time distribution given baseline covariate, and the true individual event times T, under two benchmark simulation settings: (left) simulation setup 1 in Gui et al. (2024), and (right) simulation setup 1 in Farina et al. (2026).

Dynamic survival prediction with time-varying covariates. Independent of conformal prediction, dynamic survival prediction using time-varying covariates (see Rizopoulos et al. (2017) for a review) has been an important area of research in survival analysis. Recent advances in dynamic survival prediction algorithms include approaches based on super learner (Tanner et al., 2021; Rizopoulos and Taylor, 2024), random survival forests (Pickett et al., 2021), neural networks (Lee et al., 2019; Lin and Luo, 2022; Rhodes et al., 2023), and support vector machines (Xie et al., 2024). However, these methods are designed to provide accurate point predictions and lack rigorous uncertainty quantification, limiting their uptake in practice.

Relationship with adaptive and sequential conformal prediction. Updating prediction sets as new information arrives is conceptually related to adaptive or sequential conformal prediction for time series and online learning (Gibbs and Candes, 2021; Gibbs and Candès, 2024; Batra et al., 2023; Xu and Xie, 2023b,a). However, our setting differs in that prediction sets evolve because the survivor population at τ and the available covariate history change over time, not because new outcomes are observed. Accordingly, our goal is to achieve valid coverage among survivors at each prediction time τ, rather than over a stream of outcomes as in time series or online learning settings.

2 Problem setup and assumptions

Let T denote the event time of interest and C the right censoring time. Let Z_t denote the covariates measured at time $t \geq 0$, and let $\bar{Z}_t = \{Z_s : 0 \leq s \leq t\}$ denote the covariate history up to time t. In particular, Z_0 represents baseline covariates measured at the time origin. We observe $X = \min(T, C)$ and the event indicator $\Delta = \mathbb{1}(T < C)$, together with the covariate history observed up to time X. Thus, the observed data for a single individual is $O = (X, \Delta, \bar{Z}_X)$. We assume an i.i.d. sample $\mathcal{D} = \{O_i\}_{i=1}^n$ drawn from the joint distribution of O.

Prediction target and coverage criterion. For a selected prediction time $\tau \geq 0$, our goal is to construct a prediction set for the event time T_{n+1} of a new test individual, drawn from the same distribution as the training sample, who has survived to time τ, based on the covariate history $\bar{Z}_{\tau,n+1}$ observed up to that time. Specifically, given a desired coverage level $1 - \alpha \in (0, 1)$, we seek a prediction set $\hat{\mathcal{C}}(\bar{Z}_{\tau,n+1})$ satisfying the survivor-conditional coverage guarantee:

$$\mathbb{P}\Big\{T_{n+1} \in \hat{\mathcal{C}}(\bar{Z}_{\tau,n+1}) \mid T_{n+1} > \tau\Big\} \geq 1 - \alpha. \tag{1}$$

The prediction set $\hat{\mathcal{C}}(\bar{Z}_{\tau,n+1})$ can be one-sided (where the upper or lower limit is infinity) or two-sided, depending on the downstream application context. LPBs provide conservative guarantees on how long an individual is likely to remain event-free, while upper prediction bounds (UPBs) characterize the most optimistic yet plausible event timing; for example, whether the potential benefit of an intensive treatment is large enough to justify its cost or burden. In the following, we focus on constructing two-sided prediction intervals, with the understanding that the framework can also be applied to construct one-sided prediction bounds.

Remark 2.1. In contrast to existing conformal survival methods that rely only on baseline covariates Z_0, our prediction set $\hat{\mathcal{C}}(\bar{Z}_{\tau,n+1})$ explicitly incorporates the covariate history up to the prediction time τ. Allowing τ to vary allows prediction sets to be updated as new covariate information arrives, resulting in a dynamic procedure that refines uncertainty quantification over time.

We now state the assumptions under which our procedure is developed. Let $\lambda_C(t| \cdot) = \lim_{h \to 0^+} \mathbb{P}(t \leq C < t + h \mid \cdot, C \geq t)/h$ denote the conditional hazard function of the censoring time.

Assumption 2.2 (Conditional independent censoring given covariate history)**.** For all $t \geq 0$, $\lambda_C(t|\bar{Z}_T, T) = \lambda_C(t|\bar{Z}_t)$.

Assumption 2.2 states that, among individuals who are still at risk (i.e., have not experienced the event or been censored), the instantaneous risk of censoring may depend on past covariate history but not on future covariates or the event time. This assumption is commonly considered for censoring in the survival literature with time-varying covariates (e.g., Rotnitzky and Robins, 2005; Lok et al., 2018; Yang et al., 2020).

Assumption 2.3 (Positivity)**.** (i) $\mathbb{P}(T > \tau) > 0$; (ii) there exists $\eta > 0$ such that $\mathbb{P}(C > T \mid T, \bar{Z}_T) > \eta$ almost surely.

Assumption 2.3(i) requires that survival beyond τ is possible; in practice, one can choose τ so that enough individuals remain at risk at τ. Assumption 2.3(ii) requires that, for any realized covariate history, the event time is observed with non-negligible probability. This condition ensures identifiability of the upper tail of the event-time distribution and is standard in time-to-event analyses under right-censoring.

3 Dynamic conformal prediction

We first develop the dynamic conformal prediction framework in the absence of right censoring to highlight the main calibration idea. We will then show how to extend the formulation to the observed data with right censoring.

3.1 Under no right censoring

For a fixed prediction time $\tau \geq 0$, let $\bar{\mathcal{Z}}_\tau$ denote the support of $\bar{Z}_\tau$ conditional on $T > \tau$. Suppose we are given a prediction algorithm $\mathcal{A}$ for the conditional distribution of T given $(\bar{Z}_\tau, T > \tau)$. We consider a class of nested candidate prediction sets $\{\mathcal{C}_{\tau,\theta}(\bar{z}_\tau; \mathcal{A}) : \theta \in \Theta\}$, indexed by a scalar parameter $\theta \in \Theta \subseteq \mathbb{R}$ that is increasing in θ, where the index space Θ controls how rich the candidate class is. We will discuss below how Θ is chosen with an illustrative example on the candidate class.

Our goal is to obtain the prediction set with the smallest size in this class that achieves the desired survivor-conditional coverage. Formally, let θ^* be the smallest $\theta \in \Theta$ such that

$$\mathbb{P}\left\{ T \in \mathcal{C}_{\tau,\theta}(\bar{Z}_\tau; \mathcal{A}) \;\middle|\; T > \tau \right\} \geq 1 - \alpha. \tag{2}$$

Then the corresponding prediction set $\mathcal{C}_{\tau,\theta^*}$ is the smallest set in the candidate class that satisfies the coverage requirement among survivors at time τ. Existence of θ^* follows from the monotonicity of the left-hand side of (2) in θ, provided the class contains sufficiently large sets. Note that θ^* implicitly depends on the prediction algorithm $\mathcal{A}$, the form of the candidate class $\{\mathcal{C}_{\tau,\theta}\}_{\theta\in\Theta}$, and the conditional distribution of $(T, \bar{Z}_\tau) \mid (T > \tau)$.

Construction of the candidate class. For time-to-event prediction, estimates for the conditional quantile of T can be obtained using off-the-shelf software, so a natural way to construct the candidate class $\{\mathcal{C}_{\tau,\theta}\}_{\theta\in\Theta}$ is through quantile-based intervals. In particular, when placing equal emphasis on lower- and upper-tail miscoverage, one may consider

$$\mathcal{C}_{\tau,\theta}(\bar{z}_\tau; \mathcal{A}) = (\hat{q}_{1-\theta}(\bar{z}_\tau), \hat{q}_\theta(\bar{z}_\tau)), \tag{3}$$

where $\theta \in [0.5, 1]$, and $\hat{q}_\theta(\bar{z}_\tau)$ denotes the estimated θ-quantile of $T \mid (\bar{Z}_\tau = \bar{z}_\tau, T > \tau)$ according to $\mathcal{A}$. This candidate class is used in our simulations and data applications. In cases where lower- and upper-tail miscoverage are not equally undesirable, asymmetric quantile pairs can be used instead. Additionally, the choice of quantile levels determines the location of the nested candidate intervals and thus affects the length of the final prediction interval. To further improve efficiency, the position of the candidate intervals can be selected in a data-adaptive manner. More generally, one can form the candidate class based on conformity scores, with details provided in Appendix B.

Reformulation of the calibration condition. Note that the calibration parameter θ^* is unknown and needs to be estimated from data. To motivate our estimator and its extension in right censored settings, we rewrite the coverage condition (2) in terms of an estimating inequality. Note that (2) can be equivalently written as:

$$\mathbb{E}\left[\mathbb{1}\{T \in \mathcal{C}_{\tau,\theta}(\bar{Z}_\tau; \mathcal{A})\} \middle| T > \tau \right] \geq 1 - \alpha, \tag{4}$$

which can be rewritten in unconditional form as $\mathbb{E}\{U^*(\theta, \mathcal{A})\} \geq 0$, where

$$U^*(\theta; \mathcal{A}) = \mathbb{1}(T > \tau) \left[\mathbb{1}\{T \in \mathcal{C}_{\tau,\theta}(\bar{Z}_\tau; \mathcal{A})\} - (1 - \alpha) \right]. \tag{5}$$

Therefore, θ^* can be characterized as

$$\theta^* = \inf\{\theta \in \Theta : \mathbb{E}\{U^*(\theta; \mathcal{A})\} \geq 0\}. \tag{6}$$

When equality holds at θ^*, the function $U^*(\theta; \mathcal{A})$ serves as a nonparametric estimating function for θ^* in the censoring-free data.

3.2 Under right censoring

We now turn to the setting under right censoring. In this case, neither the event time T nor the covariate history $\bar{Z}_\tau$ is always fully observed, so the censoring-free calibration procedure developed in Section 3.1 no longer applies directly. Nevertheless, the key insight from the previous subsection still applies: if we can learn the calibration parameter θ^*, then a valid prediction set can be constructed as $\hat{\mathcal{C}} = \mathcal{C}_{\tau,\hat{\theta}}$. We first consider mapping the censoring-free data estimating function of θ^* to the observed data using IPCW and then consider two doubly robust variants.

Identification of θ^*. Let $G(t|\bar{Z}_t) = \mathbb{P}(C > t \mid \bar{Z}_t)$ denote the conditional survival probability of censoring given covariate history. Under assumption 2.2, $G(t|\bar{Z}_t) = \exp\{-\int_0^t \lambda_C(u|\bar{Z}_u)\, du\}$. Applying IPCW (Robins and Rotnitzky, 1992; Rotnitzky and Robins, 2005) to the estimating function of θ^* in (5) results in the following observed data estimating functions:

$$U(\theta; G, \mathcal{A}) = \frac{\Delta \mathbb{1}(X > \tau)}{G(X|\bar{Z}_X)} \left[\mathbb{1}\{X \in \mathcal{C}_{\tau,\theta}(\bar{Z}_\tau; \mathcal{A})\} - (1-\alpha)\right], \tag{7}$$

Under Assumption 2.2, $G(T|\bar{Z}_T) = \mathbb{P}(C > T \mid T, \bar{Z}_T)$, and the IPCW factor $\Delta/G(X|\bar{Z}_X) = \mathbb{1}(T < C)/G(T|\bar{Z}_T) = \mathbb{1}(T < C)/\mathbb{P}(T < C \mid T, \bar{Z}_T)$ reweights uncensored individuals to represent the hypothetical population under no right censoring, thereby correcting the bias induced by censoring.

Lemma 3.1. *Under Assumptions 2.2 and 2.3,* $\mathbb{E}\{U(\theta; G, \mathcal{A})\} = \mathbb{E}\{U^*(\theta; \mathcal{A})\}$ *for all* $\theta \in \Theta$. *Therefore,* $\theta^* = \inf\{\theta \in \Theta : \mathbb{E}\{U(\theta; G, \mathcal{A})\} \geq 0\}$.

Lemma 3.1 shows that the IPCW estimating function preserves the calibration condition from the censoring-free setting. In particular, θ^* can be identified as the minimum value of θ for which the weighted estimating equation is nonnegative in expectation. A proof is provided in Appendix D.

Estimation of θ^* and the prediction set. We next integrate the calibration condition of Lemma 3.1 to construct prediction sets that target the survivor-conditional coverage guarantee in (2). For illustration, we focus on split conformal prediction, but the same framework can be applied to other conformal prediction algorithms (see Angelopoulos et al. (2024) for a comprehensive review). The resulting dynamic split conformal prediction procedure is summarized in Algorithm 1.

Algorithm 1: History-Aware Prediction Sets (HAPS)

Input: observed data $\mathcal{D} = \{O_i\}_{i=1}^n$; coverage level $1-\alpha$; prediction time τ; prediction model $\mathcal{A}$ for $T \mid (\bar{Z}_\tau, T > \tau)$; a rule of constructing the candidate class after fitting the prediction model; censoring model for G; covariate history $\bar{Z}_{\tau,n+1}$ from a new test individual.

1. Randomly split $\mathcal{D}$ into a training set $\mathcal{D}_{\text{tr}}$ and a calibration set $\mathcal{D}_{\text{cal}}$ with almost equal size.
2. Fit the prediction model $\hat{\mathcal{A}}$ on $\mathcal{D}_{\text{tr}}$, and construct nested candidate sets $\{\mathcal{C}_{\tau,\theta}\}_{\theta\in\Theta}$.
3. On $\mathcal{D}_{\text{tr}}$, estimate the nuisance parameters G, and denote the estimate as $\hat{G}$.
4. On $\mathcal{D}_{\text{cal}}$, compute $\hat{\theta} = \inf\left\{\theta \in \Theta : \sum_{i\in\mathcal{D}_{\text{cal}}} U_i(\theta; \hat{G}, \hat{\mathcal{A}}) \geq 0\right\}$.

Output: prediction set $\hat{\mathcal{C}}(\bar{Z}_{\tau,n+1}) = \mathcal{C}_{\tau,\hat{\theta}}(\bar{Z}_{\tau,n+1}; \hat{\mathcal{A}})$.

In practice, $\hat{\theta}$ in Algorithm 1 can be computed using a grid search on Θ restricted to a discrete and finite set. For example, with the quantile-based candidate class in (3), we take Θ to be a uniform grid on $[0.5, 1]$ with increment 0.01 in both the simulation and data applications. The nuisance parameter G can be estimated using standard survival modeling techniques that accommodate dependent right censoring with time-varying covariates, with detailed discussion provided in Appendix C.1.

3.3 Survivor-conditional coverage

We now present the theoretical guarantee for HAPS. In particular, we show that the constructed prediction sets achieve survivor-conditional coverage at a given prediction time τ, up to terms that vanish as the calibration sample size increases and the estimator for the censoring distribution is consistent.

We consider the following L_2-norm of the censoring distribution estimation error: $\|\hat{G} - G\|_2 = \mathbb{E}_{X,\bar{Z}_X}\{|\hat{G}(X \mid \bar{Z}_X) - G(X \mid \bar{Z}_X)|^2\}^{1/2}$, where the expectation is taken with respect to $(X, \bar{Z}_X)$ conditional on the data used to estimate $\hat{G}$. Let $s_\tau = \mathbb{P}(T > \tau)$ denote the marginal survival probability at time τ, and $|\mathcal{D}_{\text{cal}}|$ denote the calibration sample size.

Theorem 3.2. *Under Assumptions 2.2 and 2.3, and assume that $\hat{G}(X_{n+1}|\bar{Z}_{X_{n+1},n+1}) > \eta$ almost surely, and that the class of candidate sets contains large enough sets so that Algorithm 1 admits a feasible solution, then for any $\epsilon \in (0,1)$, there exists a constant $\kappa > 0$ such that, with probability at least $1-\epsilon$ over $\mathcal{D}$,*

$$\mathbb{P}(T_{n+1} \in \hat{\mathcal{C}}(\bar{Z}_{\tau,n+1}) \mid T_{n+1} > \tau, \mathcal{D}) \geq 1 - \alpha - s_\tau^{-1}\eta^{-2}\|\hat{G} - G\|_2 - s_\tau^{-1}\eta^{-1}\left(\sqrt{\frac{1}{2}\log\frac{1}{\epsilon}} + \kappa\right)\frac{1}{\sqrt{|\mathcal{D}_{cal}|}}. \tag{8}$$

Theorem 3.2 shows that HAPS achieves *probably asymptotically approximately correct (PAAC)* coverage among survivors at time τ when G is consistently estimated. We adopt the term "PAAC" from Qiu et al. (2023), which is referred to as "asymptotic PAC" in Farina et al. (2026). The gap between the lower bound in (8) and the nominal level $1-\alpha$ decomposes into two components: (i) an estimation error term due to learning the censoring distribution G, and (ii) a calibration error term of order $|\mathcal{D}_{\text{cal}}|^{-1/2}$ from split conformal calibration. Consequently, when $\hat{G}$ is consistent and the calibration sample size grows, the coverage lower bound converges to the nominal level $1-\alpha$. The proof of Theorem 3.2 is provided in Appendix E.

When a (semi)-parametric model is used to estimate G, consistency of $\hat{G}$ can be achieved under correct model specification. In practice, one may also use flexible nonparametric or machine learning methods to estimate G, which can reduce reliance on parametric model assumptions. To further reduce sensitivity to potential misspecification of the censoring model, we next present two doubly robust extensions of HAPS.

4 Doubly robust extensions of HAPS

Doubly robust post-processing. Recall that $\hat{q}_\theta(\bar{z}_\tau)$ denotes the estimated θ-quantile of $T \mid (\bar{Z}_\tau = \bar{z}_\tau, T > \tau)$ under the prediction algorithm $\mathcal{A}$. Given the prediction interval $\hat{\mathcal{C}}(\bar{Z}_{\tau,n+1})$ from Algorithm 1, we consider the post-processing set

$$\hat{\mathcal{C}}_{\text{DR}}(\bar{Z}_{\tau,n+1}) = \hat{\mathcal{C}}(\bar{Z}_{\tau,n+1}) \cup \left(\hat{q}_{\alpha/2}(\bar{Z}_{\tau,n+1}),\ \hat{q}_{1-\alpha/2}(\bar{Z}_{\tau,n+1})\right).$$

We call this extension HAPS-DR. The prediction set $\hat{\mathcal{C}}_{\text{DR}}$ is doubly robust in the sense that it achieves PAAC coverage among survivors when either $\hat{G}$ is consistent and achieves asymptotic coverage among survivors when the quantile estimators $\hat{q}_{\alpha/2}$ and $\hat{q}_{1-\alpha/2}$ are consistent. When $\hat{G}$ is consistent, PAAC coverage follows from Theorem 3.2, since $\hat{\mathcal{C}}_{\text{DR}}$ contains the original HAPS prediction set $\hat{\mathcal{C}}$. On the other hand, when $\hat{q}_{\alpha/2}$ and $\hat{q}_{1-\alpha/2}$ consistently estimate the corresponding conditional quantiles of $T \mid (\bar{Z}_\tau, T > \tau)$, the quantile interval $\left(\hat{q}_{\alpha/2}(\bar{Z}_{\tau,n+1}),\ \hat{q}_{1-\alpha/2}(\bar{Z}_{\tau,n+1})\right)$ achieves asymptotic $1-\alpha$ coverage among survivors, and hence so does its union with $\hat{\mathcal{C}}$.

AIPCW. Unlike HAPS-DR, this extension applies the general AIPCW approach to derive a doubly robust estimating function for θ^*, which leads to prediction intervals with doubly robust coverage. AIPCW (Robins and Rotnitzky, 1992; Rotnitzky and Robins, 2005) is a general approach for doubly robust censoring adjustment by augmenting IPCW estimating equations with outcome modeling.

Before presenting the AIPCW estimating function, we first introduce the counting process and martingale notation for C. Let $N_C(t) = \mathbb{1}(X \leq t, \Delta = 0)$, and $M_C(t) = N_C(t) - \int_0^t \mathbb{1}(X \geq u)\lambda_C(u|\bar{Z}_u)du$. The AIPCW estimating function for θ^* is:

$$U_{\text{AIPCW}}(\theta; G, h_\tau, \mathcal{A}) = U(\theta; G, \mathcal{A}) + \int_0^X h_\tau(u, \bar{Z}_u; \theta, \mathcal{A})\frac{dM_C(u)}{G(u|\bar{Z}_u)}, \tag{9}$$

where $h_\tau(u, \bar{Z}_u; \theta, \mathcal{A}) = \mathbb{E}\left(\mathbb{1}(T > \tau)\left[\mathbb{1}\{T \in \mathcal{C}_{\tau,\theta}(\bar{Z}_\tau; \mathcal{A})\} - (1-\alpha)\right] \middle| \bar{Z}_u, T \geq u\right)$.

Compared to (7), U_{AIPCW} involves an additional nuisance function h_τ, which is a conditional expectation involving the joint distribution of event time and covariate history. Estimating this conditional expectation can be challenging when covariate histories are observed in continuous time. In Appendix C.2, we discuss these challenges and provide nonparametric identification and estimation of h_τ. We construct the prediction set by augmenting Step 3 of Algorithm 1 with an estimator of h_τ, and replacing U by U_{AIPCW} in Step 4. We call this extension HAPS-A.

HAPS-A is is rate doubly robust due to the mixed-bias property (Rotnitzky et al., 2021; Luo et al., 2025): the effect of nuisance estimation on the coverage error enters only through the product of the estimation errors for G and h_τ. Thus, G and h_τ may be estimated using flexible machine learning methods, which often converge more slowly than $n^{-1/2}$ individually, while still allowing the coverage error to vanish at a $n^{-1/2}$ rate, provided standard regularity conditions hold and that the product error rate of G and h_τ is faster than $n^{-1/2}$.

We now formally state the coverage guarantee for HAPS-A. Recall that $\lambda_C(t \mid \bar{Z}_t)$ denotes the conditional hazard function of C given covariate history, and let $\Lambda_C(t \mid \bar{Z}_t) = \int_0^t \lambda_C(s \mid \bar{Z}_s)\, ds$ denote the conditional cumulative hazard function of C. To simplify the presentation and build intuition about the product errors, we assume that $\hat{G}$ admits a nonnegative censoring hazard function $\hat{\lambda}_C$, with corresponding cumulative hazard $\hat{\Lambda}_C(t \mid \bar{Z}_t) = \int_0^t \hat{\lambda}_C(u \mid \bar{Z}_u)\, du = -\log \hat{G}(t \mid \bar{Z}_t)$. The coverage result extends to more general settings using integral products between the estimation errors of h_τ and Λ_C. Integral products are encountered in rate double robustness for time-to-event data where both sets of nuisance parameters involve functions of time (Ying, 2023; Wang et al., 2024a; Westling et al., 2024; Luo et al., 2025).

Theorem 4.1 (HAPS-A coverage)**.** *Under Assumptions 2.2 and 2.3 hold, and assume that* $\hat{G}(X_{n+1} \mid \bar{Z}_{X_{n+1},n+1}) > \eta$ *almost surely, and that* $\hat{h}_\tau$ *takes values in* $[-(1-\alpha), \alpha]$*. In addition, assume that* Θ *is finite and that, almost surely, there exists* $\theta \in \Theta$ *such that* $\sum_{i \in \mathcal{I}_{cal}} U_{AIPCW,i}(\theta; \hat{G}, \hat{h}_\tau, \hat{\mathcal{A}}) \geq 0$*. Then for any* $\epsilon \in (0,1)$*, there exists a constant* $\kappa > 0$ *such that, with probability at least* $1-\epsilon$ *over* $\mathcal{D}$*,*

$$\begin{aligned}
&\mathbb{P}\{T_{n+1} \in \hat{\mathcal{C}}_{AIPCW}(\bar{Z}_{\tau,n+1}) \mid T_{n+1} > \tau, \mathcal{D}\} \\
&\quad \geq 1 - \alpha - s_\tau^{-1}\eta^{-1} \sup_{\theta \in \Theta} [\min\{r_1(\theta),\, r_2(\theta) + (1-\eta)\, r_3(\theta)\}] \\
&\quad\quad - s_\tau^{-1} c_\eta \left(\sqrt{\frac{1}{2}\log\frac{1}{\epsilon}} + \kappa \right) \frac{1}{\sqrt{|\mathcal{D}_{cal}|}},
\end{aligned}$$

where $c_\eta = 2\max(\alpha, 1-\alpha)(2\eta^{-1}-1)$*, and* r_1*,* r_2*,* r_3*, defined in* (21)–(23) *in Appendix E.2, are integrals of products between the estimation errors for* h_τ *and conditional censoring distribution.*

Theorem 4.1 shows that the coverage gap of HAPS-A is bounded by two terms: the first term involves product errors of the nuisance estimators (r_1, r_2, r_3), and the second term is of order $O(|\mathcal{D}_{\text{cal}}|^{-1/2})$, which goes to zero as the calibration sample size gets larger. The proof is provided in Appendix E.2.

5 Numerical experiments

We consider the data generating mechanisms (DGMs) described in Appendix F.1, which vary in their event-time and censoring mechanisms and have censoring rates between 32%–54%. We construct two-sided prediction intervals for survivors at $\tau \in \{0, 3, 6\}$ using HAPS and its two extensions (HAPS-DR and HAPS-A). As comparators, we consider a baseline IPCW method that uses only baseline covariates for both prediction and censoring adjustment (details in Appendix F.2), denoted by 'IPCW'. In addition, we also include two natural adaptations of this baseline method to the dynamic prediction setting: 'trunc', which takes the intersection of the IPCW interval with $[\tau, \infty)$; and 'LM', which applies the baseline IPCW method to the landmark subgroup at time τ, i.e., individuals with $X > \tau$. Both baseline adaptations produce prediction intervals contained in $[\tau, \infty)$, the support of event time among survivors at τ.

We consider fitting prediction models for HAPS and its two extensions (HAPS-DR and HAPS-A) using a neural-network dynamic survival model adapted from Dynamic-DeepHit (Lee et al., 2019),

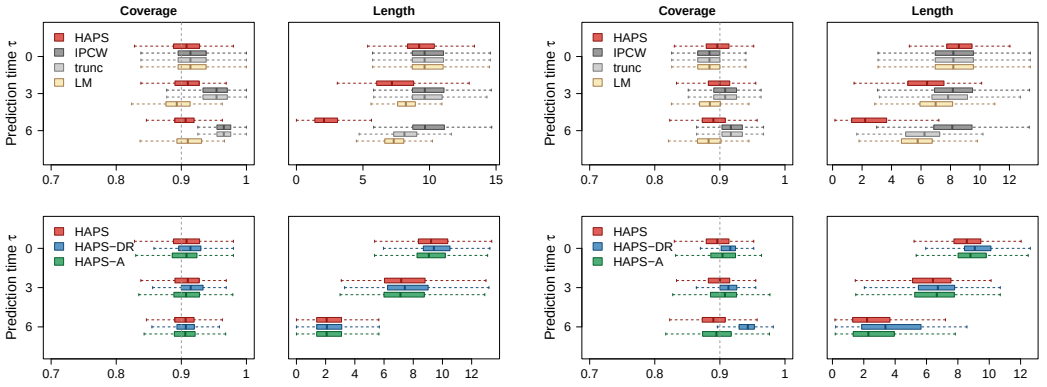
denoted by 'DDH', which leverages the covariate history up to the prediction time; landmarking-based Cox proportional hazards models ('Cox'); or random survival forests ('RSF'). For nuisance parameter estimation, we consider both semiparametric models and machine learning methods. Details of the prediction and nuisance models are provided in Appendix F.3.

We evaluate the performance of the prediction intervals over 200 Monte Carlo replications. In each replication, we generate an observed sample of size 1000 to construct the prediction intervals, and an independent evaluation sample of size 500 without right censoring, which allows direct assessment of coverage.

Results. We report the empirical coverage, interval length, and lower and upper endpoints of the two-sided prediction intervals for each method over the 200 Monte Carlo replications. Figure 2 summarizes the empirical coverage and interval length of the 90% prediction intervals under two simulation setups, with boxplots computed among survivors at each τ. Setup A generates data from DGM1, where the censoring distribution follows a proportional hazards model depending on time-varying covariates, and uses a Cox prediction model together with a correctly specified proportional hazards model for estimating the censoring distribution. Setup B generates data from DGM2, where the censoring distribution is a mixture of uniform and Weibull distributions depending on time-varying covariates, and uses a DDH prediction model together with machine learning methods for nuisance function estimation. Details of the simulation setups and implementation details, as well as additional summaries for the lower and upper endpoints of the two-sided intervals are provided in Appendix F.

We observe that HAPS and HAPS-A achieve close to 90% coverage in both setups for survivors at all three τ, while HAPS-DR tends to overcover. The prediction intervals for all three methods become tighter as τ increases, reflecting the efficiency gain from using more covariate information for prediction. We also observe that HAPS and HAPS-A have similar median interval length, while HAPS-DR is typically wider, reflecting its conservativeness due to the doubly robust post-processing step.

On the other hand, baseline methods ('IPCW', 'trunc', 'LM') can show large coverage gaps. In both simulation setups, 'IPCW' and 'trunc' overcover at $\tau = 3$ and $\tau = 6$, while 'LM' undercovers in Setup B at all three prediction times. Even at $\tau = 0$, where the coverage target is the full population, the baseline methods exhibit coverage gaps in both settings because adjustment for right censoring using only baseline covariates is insufficient. Table 1 summarizes the ratios of median prediction interval length for each version of HAPS relative to the baseline methods. Relative to the baseline methods, HAPS and its extensions reduce median interval length by **5%–23%** at $\tau = 3$ and **34–75%** at $\tau = 6$. To disentangle the two roles of covariate history in the empirical improvements, Appendix F.7 considers additional baselines. The results suggest that using covariate history for handling censoring in calibration ensures close to nominal coverage, and incorporating covariate history in the prediction model leads to shorter intervals. Appendix F.8 reports additional simulation results with sample sizes that mimic those in the two applications.



(a) Setup A: DGM1 with Cox prediction model.

(b) Setup B: DGM2 with DDH prediction model.

Figure 2: Summary of 90% prediction intervals across 200 Monte Carlo replications comparing HAPS with the baseline methods (top) and extensions of HAPS (bottom).

Table 1: Ratio of median interval length for HAPS and its extensions relative to baseline methods.

		$\tau = 3$			$\tau = 6$		
Setup	Method	IPCW	trunc	LM	IPCW	trunc	LM
A	HAPS	0.79	0.79	0.91	0.25	0.29	0.32
DGM1:Cox	HAPS-DR	0.81	0.81	0.95	0.25	0.29	0.33
	HAPS-A	**0.77**	**0.77**	**0.90**	0.25	0.29	0.32
B	HAPS	**0.78**	**0.81**	**0.91**	**0.32**	**0.43**	**0.46**
DGM2:DDH	HAPS-DR	0.82	0.84	0.95	0.46	0.61	0.66
	HAPS-A	0.80	0.83	0.93	0.33	0.44	0.48

Compared with HAPS, the advantage of HAPS-A and HAPS-DR is robustness to censoring estimation errors. Appendix F.5 presents additional simulations to assess this robustness, including settings with misspecified censoring models and settings with flexible machine learning methods to estimate the censoring distribution. Appendix F.6 further studies the impact of prediction model accuracy on prediction intervals. We observe that HAPS can undercover when the censoring model is misspecified. HAPS-DR is more robust to censoring model misspecification, although at the cost of more conservative prediction intervals, especially in settings with small sample sizes or poorly trained prediction models. When machine learning methods are used for censoring estimation, HAPS shows robust coverage under different censoring mechanisms, especially with larger sample sizes. In practice, HAPS is often preferred for its simplicity, whereas HAPS-DR is preferred when robustness is the primary concern. Although HAPS-A is also doubly robust, it requires careful estimation of additional nuisance parameters for the augmentation term, which can be computationally demanding (Appendix C.3).

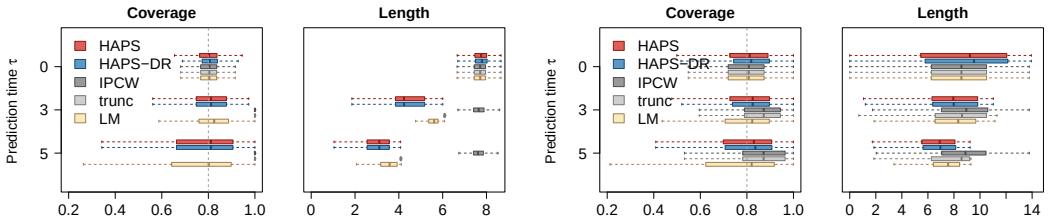
6 Applications

We apply the proposed methods to two publicly available benchmark datasets from the `survival` R package: the colon cancer dataset '`colon`' and the primary biliary cholangitis (PBC) dataset '`pbcseq`'. For both datasets, we focus on predicting overall survival time. The colon cancer dataset contains 929 patients and has a censoring rate of approximately 51.3%. The PBC dataset contains 312 patients with primary biliary cholangitis and has a censoring rate of approximately 55.1%. Details of the datasets and the covariates included in the analysis are summarized in Appendix G.1.

Results. We construct two-sided prediction intervals using the DDH prediction model and a proportional hazards model for censoring. For evaluation, we repeat 200 random splits of each dataset into an 80% sample for interval construction and a 20% test sample. Unlike in the simulations, empirical survivor-conditional coverage cannot be computed directly on the test sample because event times are not fully observed under right censoring. We therefore report IPCW estimates of survivor-conditional coverage, with details provided in Appendix G.2. These estimates are not the true coverage or the empirical coverage computed with the true event times; they approximate the true coverage only when the censoring weights are close to the truth. Appendix G.5 analyzes the sensitivity of the coverage estimates to the censoring models used for evaluation.

Figure 3 summarizes the 80% prediction intervals among survivors at $\tau = 0, 3, 5$ years for both data sets, with additional results for other prediction times provided in Appendix G. Across the two data applications, the prediction intervals from HAPS and HAPS-DR are similar and attain IPCW-estimated coverage close to 80%. Their median interval lengths decrease as τ increases, reflecting the additional information provided by covariate histories observed up to later prediction times. Table 7 in the appendix summarizes the ratios of median prediction interval lengths of HAPS and HAPS-DR relative to the baseline methods ('IPCW', 'trunc', 'LM'). Compared to the baseline methods, HAPS and HAPS-DR reduce median interval length by **9%–22%** in the PBC data and **13%–60%** in the colon cancer data for predictions at year 5.

Computational resources used for the simulation and data-application experiments are summarized in Appendix H.1.



(a) Colon cancer data

(b) PBC data

Figure 3: Summary of 80% prediction intervals across 200 random splits for data applications; coverage is IPCW-estimated (Appendix G.2).

7 Discussion

We develop History-Aware Prediction Sets (HAPS), a conformal framework for constructing prediction intervals for individual event times using covariate histories observed up to a prediction time τ. Across simulations and data applications, HAPS produces substantially shorter prediction intervals than baseline methods while maintaining close-to-nominal coverage, showing its potential to aid decision-making in settings where updated covariate information is collected over time. The code for implementing the method and reproducing the simulation results is available at: `https://github.com/wangyuyao98/HAPS_neurips2026`.

An interesting future direction is to extend HAPS to competing risks and semi-competing risks settings, leveraging advances in modern dynamic prediction algorithms such as Dynamic-DeepHit (Lee et al., 2019). Another promising direction is to extend HAPS to left-truncated and right-censored data, which arise in studies with delayed entry, by leveraging inverse probability of truncation weighting or doubly robust methods to account for covariate-dependent left truncation (Vakulenko-Lagun et al., 2022; Wang et al., 2024a,b; Han, 2024). A third extension is to multi-source and federated settings. Building on multi-source conformal inference under distribution shift (Liu et al., 2024), adaptive borrowing across sites could improve calibration efficiency while targeting coverage among survivors in a specified target population. Related advances in federated causal survival analysis (Liu et al., 2026) motivate implementations that avoid sharing individual-level records.

Limitations. A limitation of HAPS and its extensions is that their validity relies on conditionally independent censoring given the observed time-varying covariate history, as well as positivity. Violations of these assumptions may lead to prediction intervals with invalid coverage. With extreme weights, and especially for small α, the algorithm may return trivial intervals (Appendix G.6). While these conditions are weaker than those imposed by existing conformal methods for right-censored time-to-event outcomes, practitioners should still conduct diagnostics in practice, especially for high-stakes applications. Sensitivity analyses can help to evaluate the impact of potential violations.

An additional limitation of HAPS is its sensitivity to errors in estimating the censoring distribution; misspecification of the censoring model can lead to coverage gaps. To address this limitation, we propose two doubly robust extensions, HAPS-DR and HAPS-A, which improve robustness but introduce additional trade-offs. HAPS-DR can be conservative due to its post-processing step, whereas HAPS-A requires estimating additional nuisance parameters and may be computationally challenging or unstable in small samples or under positivity violations. In practice, HAPS-DR may be preferred when robustness is prioritized, while HAPS may be preferable when simplicity and efficiency are primary concerns.

Broader impacts. HAPS provides a model-agnostic framework for uncertainty quantification that can be wrapped around existing prediction algorithms, taking a step toward more reliable use of modern dynamic survival algorithms in high-stakes settings such as clinical care. As shown in our simulations and applications, HAPS and its extensions produce substantially shorter prediction intervals than baseline methods, improving the practical utility of conformal prediction. Potential negative impacts may arise if practitioners over-rely on the resulting intervals without assessing the plausibility of the required assumptions or the sensitivity of conclusions to assumption violations.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers and the area chair for their constructive comments. This research was supported in part by a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-2025C1-44648-IC), the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (K01AI193070; L.H.), and the U.S. National Science Foundation (SES-2242776; S.Y.). This research also uses services provided by the Open Science Grid (OSG) Consortium (Pordes et al., 2007; Sfiligoi et al., 2009; OSG, 2006, 2015), which is supported by National Science Foundation awards #2030508 and #1836650. The content, including any opinions, findings, conclusions, or recommendations, is solely the responsibility of the authors and does not necessarily represent the views of PCORI, its Board of Governors or Methodology Committee, or the official views of the National Institutes of Health or the U.S. National Science Foundation. The authors declare no competing interests.

References

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. arXiv:2411.11824, 2024.

Christopher RS Banerji, Tapabrata Chakraborti, Chris Harbron, and Ben D MacArthur. Clinical AI tools must convey predictive uncertainty for each individual patient. *Nature Medicine*, 29(12): 2996–2998, 2023.

Devesh Batra, Salvatore Mercuri, and Raad Khraishi. Conformal predictions for longitudinal data. arXiv:2310.02863, 2023.

Emmanuel Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):24–45, 2023.

Hen Davidov, Shai Feldman, Gil Shamai, Ron Kimmel, and Yaniv Romano. Conformalized survival analysis for general right-censored data. In *The Thirteenth International Conference on Learning Representations*, 2025.

Rebecca Farina, Eric Tchetgen Tchetgen, and Arun Kumar Kuchibhotla. Doubly robust and efficient calibration of prediction sets for right-censored time-to-event outcomes. *Biometrika*, 113 (3):asag036, 09 2026.

Chenyin Gao, Peter B Gilbert, and Larry Han. On the role of surrogates in conformal inference of individual causal effects. arXiv:2412.12365, 2024.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Isaac Gibbs and Emmanuel J Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Yu Gui, Rohan Hore, Zhimei Ren, and Rina Foygel Barber. Conformalized survival analysis with adaptive cut-offs. *Biometrika*, 111(2):459–477, 2024.

Larry Han. Truncated, not forgottenhandling left truncation in time-to-event studies, 2024.

Robin Henderson, Margaret Jones, and Janez Stare. Accuracy of point predictions in survival analysis. *Statistics in Medicine*, 20(20):3083–3096, 2001.

Chris Holmes and Ariane Marandon. Two-sided conformalized survival analysis. arXiv:2410.24136, 2024.

Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond. *Journal of Machine Learning Research*, 25 (16):1–59, 2024.

John A Laurie, Charles G Moertel, Thomas R Fleming, Harry S Wieand, John E Leigh, Jebal Rubin, Greg W McCormack, James B Gerstner, James E Krook, and James Malliard. Surgical adjuvant therapy of large-bowel carcinoma: an evaluation of levamisole and the combination of levamisole and fluorouracil. the north central cancer treatment group and the mayo clinic. *Journal of clinical oncology*, 7(10):1447–1456, 1989.

Changhee Lee, Jinsung Yoon, and Mihaela Van Der Schaar. Dynamic-DeepHit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 67(1):122–133, 2019.

Jeffrey Lin and Sheng Luo. Deep learning for the dynamic prediction of multivariate longitudinal and survival data. *Statistics in Medicine*, 41(15):2894–2907, 2022.

Yi Liu, Alexander Levis, Sharon-Lise Normand, and Larry Han. Multi-source conformal inference under distribution shift. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 31344–31382. PMLR, 2024. URL `https://proceedings.mlr.press/v235/liu24ag.html`.

Yi Liu, Alexander Levis, Ke Zhu, Shu Yang, Peter Gilbert, and Larry Han. Privacy-protected causal survival analysis under distribution shift. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://proceedings.iclr.cc/paper_files/paper/2026/hash/98b4a12cd33fe955953d275aaa2a0785-Abstract-Conference.html`.

Judith J Lok, Shu Yang, Brian Sharkey, and Michael D Hughes. Estimation of the cumulative incidence function under multiple dependent and independent censoring mechanisms. *Lifetime data analysis*, 24(2):201–223, 2018.

Jiyu Luo, Denise Rava, Jelena Bradic, and Ronghui Xu. Doubly robust estimation under a possibly misspecified marginal structural cox model. *Biometrika*, 112(1):asae065, 2025.

Charles G Moertel, Thomas R Fleming, John S Macdonald, Daniel G Haller, John A Laurie, Phyllis J Goodman, James S Ungerleider, William A Emerson, Douglas C Tormey, John H Glick, et al. Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *New England Journal of Medicine*, 322(6):352–358, 1990.

Paul A Murtaugh, E Rolland Dickson, Gooitzen M Van Dam, Michael Malinchoc, Patricia M Grambsch, Alice L Langworthy, and Chris H Gips. Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits. *Hepatology*, 20(1):126–134, 1994.

OSG. Ospool, 2006. OSG. https://doi.org/10.21231/906P-4D78.

OSG. Open science data federation, 2015. OSG. https://doi.org/10.21231/0KVZ-VE57.

Kaci L Pickett, Krithika Suresh, Kristen R Campbell, Scott Davis, and Elizabeth Juarez-Colunga. Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker. *BMC medical research methodology*, 21(1):216, 2021.

Ruth Pordes, Don Petravick, Bill Kramer, Doug Olson, Miron Livny, Alain Roy, Paul Avery, Kent Blackburn, Torre Wenaus, Frank Würthwein, Ian Foster, Rob Gardner, Mike Wilde, Alan Blatecky, John McGee, and Rob Quick. The open science grid. In *J. Phys. Conf. Ser.*, volume 78, page 012057, 2007. doi: 10.1088/1742-6596/78/1/012057.

Jing Qin, Jin Piao, Jing Ning, and Yu Shen. Conformal predictive intervals in survival analysis: a resampling approach. *Biometrics*, 81(2):ujaf063, 2025.

Hongxiang Qiu, Edgar Dobriban, and Eric Tchetgen Tchetgen. Prediction sets adaptive to unknown covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5): 1680–1705, 2023.

Hongxiang Qiu, Marco Carone, Alex Luedtke, and Peter B Gilbert. Multiply robust estimation of conditional survival probability with time-varying covariates. arXiv:2510.10372, 2026.

Grace Rhodes, Marie Davidian, and Wenbin Lu. Dynamic prediction of residual life with longitudinal covariates using long short-term memory networks. *Annals of Applied Statistics*, 17(3): 2039–2058, 2023.

Dimitris Rizopoulos and Jeremy MG Taylor. Optimizing dynamic predictions from joint models using super learning. *Statistics in medicine*, 43(7):1315–1328, 2024.

Dimitris Rizopoulos, Geert Molenberghs, and Emmanuel MEH Lesaffre. Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal*, 59(6):1261–1276, 2017.

James M Robins and Andrea Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS epidemiology*, pages 297–331. Springer, New York, NY, 1992.

A Rotnitzky and JM Robins. Inverse probability weighting in survival analysis. In *Encyclopedia of Biostatistics*, volume 4, pages 2619–2625. Wiley, New York, NY, 2005.

Andrea Rotnitzky, Ezequiel Smucler, and James M Robins. Characterization of parameters with a mixed bias property. *Biometrika*, 108(1):231–238, 2021.

Matteo Sesia and Vladimir Svetnik. Doubly robust conformalized survival analysis with right-censored data. In *The International Conference on Machine Learning*, 2025.

Igor Sfiligoi, Daniel C Bradley, Burt Holzman, Parag Mhashilkar, Sanjay Padhi, and Frank Wurthwein. The pilot way to grid resources using glideinwms. In *2009 WRI World Congress on Computer Science and Information Engineering*, volume 2, pages 428–432, 2009. doi: 10.1109/CSIE.2009.950.

Wenwen Si and Hongxiang Qiu. Training-set conditionally valid prediction sets with right-censored data. arXiv:2505.02213, 2025.

Kamaryn T Tanner, Linda D Sharples, Rhian M Daniel, and Ruth H Keogh. Dynamic survival prediction combining landmarking with a machine learning ensemble: Methodology and empirical comparison. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(1):3–30, 2021.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Bella Vakulenko-Lagun, Jing Qian, Sy Han Chiou, Nancy Wang, and Rebecca A Betensky. Nonparametric estimation of the survival distribution under covariate-induced dependent truncation. *Biometrics*, 78(4):1390–1401, 2022.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Yuyao Wang, Andrew Ying, and Ronghui Xu. Doubly robust estimation under covariate-induced dependent left truncation. *Biometrika*, 111(3):789–808, 2024a.

Yuyao Wang, Andrew Ying, and Ronghui Xu. A liberating framework from truncation and censoring, with application to learning treatment effects. arXiv:2411.18879, 2024b.

Ted Westling, Alex Luedtke, Peter B Gilbert, and Marco Carone. Inference for treatment-specific survival curves using machine learning. *Journal of the American Statistical Association*, 119 (546):1541–1553, 2024.

Wenyi Xie, Donglin Zeng, and Yuanjia Wang. Support vector machine for dynamic survival prediction with time-dependent covariates. *Annals of Applied Statistics*, 18(3):2166–2186, 2024.

Chen Xu and Yao Xie. Conformal prediction for time series. *IEEE transactions on pattern analysis and machine intelligence*, 45(10):11575–11587, 2023a.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *International Conference on Machine Learning*, pages 38707–38727. PMLR, 2023b.

Shu Yang, Karen Pieper, and Frank Cools. Semiparametric estimation of structural failure time models in continuous-time processes. *Biometrika*, 107(1):123–136, 2020.

Yachong Yang, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(4):943–965, 2024.

Menghan Yi, Ze Xiao, Huixia Judy Wang, and Yanlin Tang. Survival conformal prediction under random censoring. *Stat*, 14(2):e70052, 2025.

Andrew Ying. Asymptotic theory for doubly robust estimators with continuous-time nuisance parameters. arXiv preprint arXiv:2302.06739, 2023.

A Summary of the conformal prediction literature for right censored time-to-event outcomes

Early developments on conformal prediction for time-to-event outcomes considered *type-I censoring*, where each individuals potential censoring time is always observed regardless of whether censoring occurs. Under this setting, Candès et al. (2023) developed calibrated lower prediction bounds by discarding individuals with early censoring times and correcting for the induced covariate shift using weighted conformal prediction (Tibshirani et al., 2019). Gui et al. (2024) extended this framework by allowing the censoring-time threshold to be selected in a covariate-dependent and data-adaptive manner, yielding tighter lower bounds while preserving validity.

Subsequent work has addressed the more common right censoring mechanism, where only the minimum of the event time and censoring time is observed. Under this censoring setting, several approaches have been developed for constructing the lower prediction bounds including methods that reduce the problem to the type-I censoring setting through imputation of missing censoring times (Sesia and Svetnik, 2025), approaches that incorporate weights to account for right censoring (Davidov et al., 2025), as well as methods that leverage semiparametric efficiency theory to handle censoring in doubly robust and efficient manners (Farina et al., 2026; Si and Qiu, 2025).

Beyond lower prediction bounds, several recent works developed two-sided prediction intervals for time-to-event outcomes under right censoring. Holmes and Marandon (2024) developed mixed-type prediction intervals using a classification-based approach: individuals are first classified into non-censored and censored groups, after which two-sided prediction intervals are constructed for individuals that are classified as non-censored with enough confidence; lower prediction bounds are constructed otherwise. Qin et al. (2025) developed a bootstrap-based approach under working models for the conditional event time distribution given covariates. Yi et al. (2025) developed a weighted conformal method that accounts for right censoring using inverse probability of censoring weighting (IPCW), with weights estimated from a localized Kaplan Meier estimator of the conditional censoring distribution given covariates.

B General construction of the candidate prediction sets based on conformity scores

To form the candidate class $\{\mathcal{C}_{\tau,\theta}\}_{\theta\in\Theta}$, one general strategy is to use a conformity score $R_\tau(\bar{z}_\tau, t; \mathcal{A})$ which measures how well a candidate event time t conforms to the prediction based on covariate history $\bar{z}_\tau$ (smaller values indicating better conformity); see Angelopoulos et al. (2024) for a comprehensive review. Specifically, the candidate class can be formed as $\mathcal{C}_{\tau,\theta}(\bar{z}_\tau; \mathcal{A}) = \{t : R_\tau(\bar{z}_\tau, t; \mathcal{A}) \leq \theta\}$, for $\theta \in \Theta$, where Θ contains all possible values of the conformity score.

In practice, to solve for $\hat{\theta}$ in Algorithm 1, one may perform a grid search on a discretized version of Θ defined by a grid over the empirical range of the conformity score evaluated on the observed data. In particular, including the maximum observed conformity score in the grid guarantees that the empirical version of (2) is satisfied for at least one $\theta \in \Theta$, ensuring the existence of a solution.

There is a trade-off between grid resolution and computational cost: finer grids may yield more informative prediction sets but incur higher search cost. In practice, domain knowledge and multi-stage search can be used to restrict the range of Θ and improve computational efficiency.

C Nuisance parameter estimation

C.1 Estimation for G with general time-varying covariate histories

The nuisance parameter G can be estimated using existing regression methods for right censored time-to-event outcomes that accommodate dependent right censoring with time-varying covariates. For example, one can fit a proportional hazards model for censoring, using the `coxph` function (with the counting process formulation) in the '`survival`' R package and treating $1 - \Delta$ as the event indicator. When stronger assumptions such as independent censoring or conditional independent censoring given baseline covariates hold, simpler estimators such as the Kaplan–Meier estimator or survival regression estimators that incorporate baseline covariates may be used instead.

C.2 Estimation for G and h_τ with covariates measured at discrete visit times

As noted in the main paper, estimating the function h_τ with general time-varying covariate histories is statistically challenging, because h_τ is defined through conditional expectations involving both the event time and the covariate history. When the covariate history is observed in continuous time, nonparametric estimation of h_τ is generally infeasible without imposing strong modeling assumptions on the joint distribution of the event time and the covariate process.

Here, we consider the setting where time-varying covariates are measured at prespecified times $0 = t_1 < \cdots < t_K$. This setting is highly relevant in real-world applications where covariates are collected at planned visit times. For notational convenience, let $t_{K+1} = \infty$. For each $k = 1, \ldots, K$, we define the interval-specific contributions of the event time and censoring time over $(t_k, t_{k+1}]$ as

$$T_k = (T \wedge t_{k+1} - t_k)\mathbb{1}(T > t_k), \qquad C_k = (C \wedge t_{k+1} - t_k)\mathbb{1}(C > t_k).$$

Let $X_k = \min(T_k, C_k)$ and $\Delta_k = \mathbb{1}(0 < T_k < C_k)$. With this notation, the original random variables can be written as

$$T = \sum_{k=1}^{K} T_k, \qquad C = \sum_{k=1}^{K} C_k, \qquad X = \sum_{k=1}^{K} X_k, \qquad \Delta = \sum_{k=1}^{K} \Delta_k.$$

Figure 4 illustrates the data structure and the definitions of T_k and C_k.

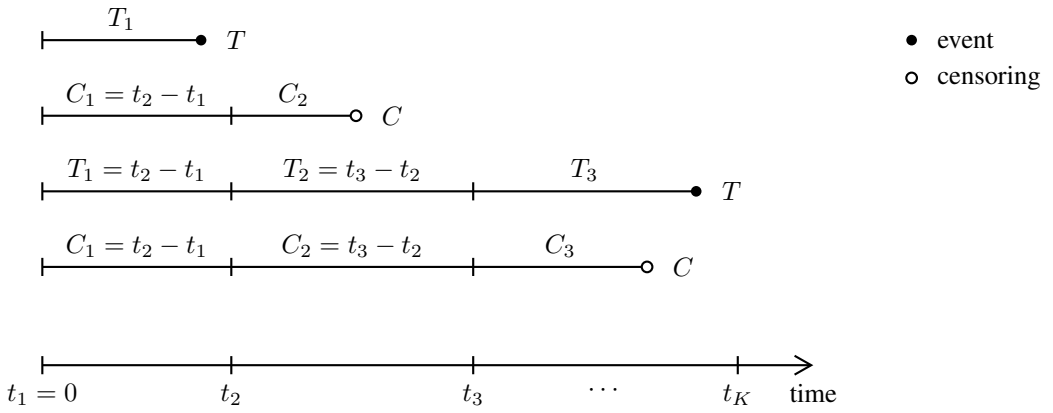


Figure 4: Illustration for the data structure and the definitions of T_k and C_k.

Since the time-varying covariate process Z_t changes values only at $t_1, \ldots, t_K$, we use the simplified notation $L_k = Z_{t_k}$ and $\bar{L}_k = (L_1, \ldots, L_k)$ to denote the covariate history observed up to time t_k. We focus on prediction at $\tau = t_\nu$ for a fixed $\nu \in \{1, \ldots, K\}$. Since for any $u \in [t_k, t_{k+1})$, $\bar{Z}_u$ contains the same information as $\bar{L}_k$, we use these two notations interchangeably in this section and write $\mathcal{C}_{\tau,\theta}(\bar{L}_\nu; \mathcal{A})$ for the candidate prediction sets.

Under this setting, Assumption 2.2 can be equivalently expressed as follows.

Assumption C.1 (Sequential conditional independent censoring)**.** For each $k = 1, \ldots, K$,

$$C_k \perp\!\!\!\perp (T_k, \ldots, T_K, L_{k+1}, \ldots, L_K) \mid (T > t_k, \bar{L}_k).$$

We assume that T and C are absolutely continuous random variables and introduce the following notation for their interval-specific conditional survival functions. For $k = 1, \ldots, K$ and $t \in (t_k, \infty)$, denote

$$S_k(t|\bar{L}_k) = \mathbb{P}(T > t \mid \bar{L}_k, X > t_k),$$
$$G_k(t|\bar{L}_k) = \mathbb{P}(C > t \mid \bar{L}_k, X > t_k).$$

C.2.1 Piecewise estimation for G.

In addition to the estimation approach described in Appendix C.1, which fits a global time-to-event model with time-varying covariates, an alternative approach is to exploit the following factorization

of G in terms of the interval-specific censoring survival functions G_k: for $t \in (t_k, t_{k+1}]$,

$$G(t|\bar{Z}_t) = \left\{ \prod_{j=1}^{k-1} G_j(t_{j+1}|\bar{L}_j) \right\} \cdot G_k(t|\bar{L}_k), \tag{10}$$

where the empty product is defined as one. One can then fit separate models for each $G_k(t|\bar{L}_k)$ on the interval $t \in (t_k, t_{k+1}]$, and estimate G by plugging the resulting estimates of G_k into (10).

The conditional censoring survival curve $G_k(t|\bar{L}_k)$, for $t \in (t_k, t_{k+1}]$, can be estimated by fitting a survival regression model for the censoring time C conditional on $\bar{L}_k$, using subjects who remain at risk at t_k and imposing administrative censoring at t_{k+1}. Specifically, the model is fitted using data

$$\left\{ \left(X_i \wedge t_{k+1}, \tilde{\Delta}_i, \bar{L}_{k,i} \right) : X_i > t_k \right\},$$

where

$$\tilde{\Delta}_i = (1 - \Delta_i) \mathbb{1}\{X_i \in (t_k, t_{k+1}]\}$$

is the event indicator for censoring. Existing software for right-censored time-to-event outcomes can be used to estimate G_k, including proportional hazards models implemented in the '`survival`' R package, random survival forests implemented in the '`randomForestSRC`' R package, and the boosting trees implemented in the '`xgboost`' R package, among others.

C.2.2 Estimation of h_τ

For ease of exposition, we focus on the quantile-based candidate prediction sets used in the numerical experiments of this paper:

$$\mathcal{C}_{\tau,\theta}(\bar{L}_\nu; \mathcal{A}) = \left(q_{1-\theta}(\bar{L}_\nu; \mathcal{A}), q_\theta(\bar{L}_\nu; \mathcal{A}) \right),$$

where $\theta \in [0.5, 1]$ and $q_\theta(\bar{L}_\nu; \mathcal{A})$ denotes the conditional θ-quantile of $T \mid (T > t_\nu, \bar{L}_\nu)$ computed from the prediction model $\mathcal{A}$. The same estimation approach can also be extended to other candidate prediction sets.

For any real numbers a and b, we use $a \vee b$ and $a \wedge b$ to denote the maximum and minimum of a and b, respectively. For any $u \in (t_k, t_{k+1}]$, we can show that (with the proof provided at the end of this subsection)

$$\begin{aligned}
&h_\tau(u, \bar{Z}_u; \theta, \mathcal{A}) \\
&= \mathbb{E}\left(\mathbb{1}(T > \tau) \left[\mathbb{1}\{T \in \mathcal{C}_{\tau,\theta}(\bar{L}_\nu; \mathcal{A})\} - (1 - \alpha) \right] \middle| \bar{L}_k, T \geq u \right) \\
&= \begin{cases} \dfrac{\xi_k(\bar{L}_k; \theta, \mathcal{A}, \alpha)}{S_k(u|\bar{L}_k)}, & \text{if } k < \nu, \text{ i.e., } u < \tau, \\ \dfrac{S_k\left(u \vee q_{1-\theta}(\bar{L}_\nu; \mathcal{A}) \mid \bar{L}_k\right) - S_k\left(u \vee q_\theta(\bar{L}_\nu; \mathcal{A}) \mid \bar{L}_k\right)}{S_k(u|\bar{L}_k)} - (1 - \alpha), & \text{if } k \geq \nu, \text{ i.e., } u \geq \tau, \end{cases}
\end{aligned} \tag{11}$$

where

$$\xi_k(\bar{L}_k; \theta, \mathcal{A}, \alpha) = \mathbb{E}\left(\mathbb{1}(T > \tau) \left[\mathbb{1}\{T \in \mathcal{C}_{\tau,\theta}(\bar{L}_\nu; \mathcal{A})\} - (1 - \alpha) \right] \middle| \bar{L}_k, X \geq t_k \right).$$

Therefore, it remains to estimate $\xi_k(\bar{L}_k)$ and $S_k(t|\bar{L}_k)$ for $t \in [t_k, \infty)$ and $k = 1, \ldots, K$. The function h_τ can then be estimated by plugging these estimates into (11).

Estimation of S_k. We first consider estimating $S_k(t|\bar{L}_k)$ for $t \in [t_k, \infty)$. A natural approach is to fit a survival regression model for T conditional on $\bar{L}_k$ using individuals in the landmark risk set at t_k, i.e., those with $X > t_k$. However, this approach is generally biased for $k < K$, because censoring after t_k may depend on future covariates that are associated with the event time. Specifically, under Assumption C.1, C may remain dependent on T conditional on $\{\bar{L}_k, X > t_k\}$ for $k < K$.

To account for this residual dependence between the censoring time and the event time, we use inverse probability of censoring weighting. Denote $\bar{t}_{j,t} = t_{j+1} \wedge t$. We use the convention that

$$G_j(s|\bar{L}_j) = 1 \quad \text{for } s \leq t_j.$$

Then, for $t > t_k$, define

$$H_k(t) = \prod_{j=k}^{K} G_j(\bar{t}_{j,t}|\bar{L}_j).$$

Under the sequential conditional independent censoring assumption (Assumption C.1), $H_k(t)$ can be interpreted as the conditional probability of remaining uncensored through time t in the landmark population at t_k, given the covariate history up to time t:

$$H_k(t) = \mathbb{P}(C > t \mid \bar{Z}_t, X > t_k).$$

Then, for each $k = 1, \ldots, K-1$, the survival curve $S_k(t|\bar{L}_k)$ can be estimated by fitting a weighted regression model for the event time T conditional on $\bar{L}_k$ among individuals with $X > t_k$, using the uncensored observations with IPCW weights $\Delta/H_k(X)$. Equivalently, one may fit the model using observations with $X > t_k$ and $\Delta = 1$, weighted by $1/H_k(X)$.

For the last interval, no future covariates remain. Thus, under Assumption C.1, $S_K(t|\bar{L}_K)$ for $t > t_K$ can be estimated by fitting a standard right-censored survival regression model for T conditional on $\bar{L}_K$ among individuals with $X > t_K$.

In practice, this weighted estimation step can be implemented using existing software for survival outcomes, as long as the chosen method accommodates observation-level weights. For example, weighted Cox proportional hazards models can be fitted using the `survival` R package, and random survival forests and boosted survival models can be implemented using packages such as `randomForestSRC` and `xgboost`, respectively. The choice of model may be guided by the effective sample size in each landmark risk set.

Estimation of ξ_k. If there were no right censoring, ξ_k could be estimated by regressing

$$\mathbb{1}(T > \tau)\left[\mathbb{1}\{T \in \mathcal{C}_{\tau,\theta}(\bar{L}_\nu; \mathcal{A})\} - (1-\alpha)\right] \tag{12}$$

on $\bar{L}_k$ among individuals at risk at t_k. With right censoring, (12) is not always observed, because T and the covariate history $\bar{L}_\nu$ may be censored. We therefore again use inverse probability of censoring weights to handle censoring.

Specifically, for each θ, $\xi_k(\bar{L}_k; \theta, \mathcal{A}, \alpha)$ can be estimated by fitting a weighted regression of the pseudo outcome

$$\mathbb{1}(X > \tau)\left[\mathbb{1}\{X \in \mathcal{C}_{\tau,\theta}(\bar{L}_\nu; \mathcal{A})\} - (1-\alpha)\right] \tag{13}$$

on $\bar{L}_k$, using uncensored individuals who are at risk at t_k, i.e., those with $X > t_k$ and $\Delta = 1$, with IPCW weights $\Delta/H_k(X)$. For individuals with observed event times before τ, the pseudo outcome (13) is zero.

We observe that the pseudo-outcome in (13) takes only three possible values: 0, α, and $-(1-\alpha)$. Therefore, ξ_k can be estimated either by fitting a weighted regression model directly to this pseudo-outcome, or by fitting a weighted multiclass classification model for the three possible outcome categories and then computing the conditional mean using the estimated class probabilities. Existing regression or classification software that accommodates observation-level weights can be used for estimating ξ_k. For example, flexible models can be implemented using boosting trees in the `xgboost` R package or random forests in the `randomForestSRC` R package. More generally, any regression or machine learning method that supports case weights may be used to estimate the conditional mean function $\xi_k(\bar{L}_k; \theta, \mathcal{A}, \alpha)$, and the choice of model may be guided by the effective sample size in each landmark risk set.

Remark. As an alternative to the approach described above, $S_k(t|\cdot)$ at a fixed $t \in (t_k, \infty)$ and ξ_k can also be nonparametrically identified and estimated using an iterative procedure similar to that of Qiu et al. (2026). However, in our setting, we require the entire conditional survival curve $S_k(t|\cdot)$ over $t \in (t_k, \infty)$. Applying the iterative procedure to this task would be computationally prohibitive, as it would require fitting separate models for each time point t.

Proof of Equation (11)**.** Let

$$Y_\theta = \mathbb{1}(T > \tau)\left[\mathbb{1}\{T \in \mathcal{C}_{\tau,\theta}(\bar{L}_\nu; \mathcal{A})\} - (1-\alpha)\right].$$

For $u \in (t_k, t_{k+1}]$, we have

$$\begin{aligned} h_\tau(u, \bar{Z}_u; \theta, \mathcal{A}) &= \mathbb{E}\left(Y_\theta \mid \bar{L}_k, T \geq u\right) \\ &= \frac{\mathbb{E}\left\{\mathbb{1}(T \geq u) Y_\theta \mid \bar{L}_k, T > t_k\right\}}{\mathbb{P}(T \geq u \mid \bar{L}_k, T > t_k)}. \end{aligned}$$

Under the independent censoring assumption and the absolute continuity of T, conditioning on $T > t_k$ can equivalently be replaced by conditioning on the observed landmark risk set $X > t_k$. Thus,

$$\mathbb{P}(T \geq u \mid \bar{L}_k, T > t_k) = \mathbb{P}(T \geq u \mid \bar{L}_k, X > t_k) = S_k(u|\bar{L}_k),$$

where the distinction between $>$ and $\geq$ is immaterial by absolute continuity.

(1) First consider $k < \nu$, so that $u < \tau$. Recall that $\tau = t_\nu$. Since $T > \tau$ implies $T \geq u$, we have

$$\begin{aligned} h_\tau(u, \bar{Z}_u; \theta, \mathcal{A}) &= \frac{\mathbb{E}\left[\mathbb{1}(T > \tau)\left\{\mathbb{1}\{T \in \mathcal{C}_{\tau,\theta}(\bar{L}_\nu; \mathcal{A})\} - (1-\alpha)\right\} \mid \bar{L}_k, X > t_k\right]}{S_k(u|\bar{L}_k)} \\ &= \frac{\xi_k(\bar{L}_k; \theta, \mathcal{A}, \alpha)}{S_k(u|\bar{L}_k)}. \end{aligned}$$

Next consider $k \geq \nu$, so that $u \geq \tau$. In this case, $\bar{L}_\nu$ is contained in $\bar{L}_k$, and hence $q_{1-\theta}(\bar{L}_\nu; \mathcal{A})$ and $q_\theta(\bar{L}_\nu; \mathcal{A})$ are known conditional on $\bar{L}_k$ and given $\mathcal{A}$. For the quantile-based candidate set

$$\mathcal{C}_{\tau,\theta}(\bar{L}_\nu; \mathcal{A}) = \left(q_{1-\theta}(\bar{L}_\nu; \mathcal{A}), q_\theta(\bar{L}_\nu; \mathcal{A})\right), \quad \theta \in [0.5, 1],$$

we have

$$\begin{aligned} h_\tau(u, \bar{Z}_u; \theta, \mathcal{A}) &= \frac{\mathbb{E}\left[\mathbb{1}(T \geq u)\left\{\mathbb{1}\left(q_{1-\theta}(\bar{L}_\nu; \mathcal{A}) < T < q_\theta(\bar{L}_\nu; \mathcal{A})\right) - (1-\alpha)\right\} \mid \bar{L}_k, X > t_k\right]}{S_k(u|\bar{L}_k)} \\ &= \frac{\mathbb{P}\left(T > u \vee q_{1-\theta}(\bar{L}_\nu; \mathcal{A}),\, T < q_\theta(\bar{L}_\nu; \mathcal{A}) \mid \bar{L}_k, X > t_k\right)}{S_k(u|\bar{L}_k)} - (1-\alpha) \\ &= \frac{S_k\left(u \vee q_{1-\theta}(\bar{L}_\nu; \mathcal{A}) \mid \bar{L}_k\right) - S_k\left(u \vee q_\theta(\bar{L}_\nu; \mathcal{A}) \mid \bar{L}_k\right)}{S_k(u|\bar{L}_k)} - (1-\alpha). \end{aligned}$$

Combining the two cases, Equation (11) follows.

C.3 Computational cost of HAPS-A

The most computationally expensive step of HAPS-A is fitting the nuisance models for the augmentation term. Table 2 summarizes the nuisance models needed for HAPS and HAPS-A. We see that the number of nuisance models needed for HAPS-A is of order $O(K|\Theta|)$, where $|\Theta|$ denotes the size of the grid Θ. Therefore, the computation time can be long when the number of visit times K is large or when the grid is fine (so $|\Theta|$ is large).

To reduce the computation cost, one may apply the localization approach of Kallus et al. (2024), which has also been used in conformal prediction (Gao et al., 2024). Specifically, the localized version computes the augmentation term only at the initial HAPS estimator $\hat{\theta}$ and solves the estimating inequality with it. Then only $K - 1$ models for ξ_k need to be fitted at $\hat{\theta}$.

Table 3 summarizes the computation time for HAPS, HAPS-A, and localized HAPS-A under different numbers of visit times K. We observe that localized HAPS-A is much faster than HAPS-A. For HAPS-A, the computation time is approximately linear in K.

Table 2: Nuisance models needed for HAPS and HAPS-A at a single prediction time τ.

Method	Nuisance models needed at a single τ
HAPS	1 prediction model + 1 censoring model (or K censoring models using the piecewise approach in Appendix C.2.1)
HAPS-A	Models in HAPS + K models for S_k + $(K-1)\lvert\Theta\rvert$ models in total for ξ_k at each $\theta \in \Theta$
Localized HAPS-A	Models in HAPS + K models for S_k + $(K-1)$ models in total for ξ_k at $\hat{\theta}$

Table 3: Computation time in seconds on a MacBook Air (M3 chip) for 3 replicates, under simulation Setup A, at $\tau = 3$, and with Θ being a uniform grid on $[0.5, 1]$ with increment 0.0025 ($|\Theta| = 201$).

n	Method	$K=3$	$K=5$	$K=9$
300	HAPS	0.03	0.04	0.05
	HAPS-A	2.9	4.2	7.2
	Localized HAPS-A	0.21	0.13	0.19
1000	HAPS	0.14	0.12	0.14
	HAPS-A	18.4	28.0	45.5
	Localized HAPS-A	0.49	0.53	0.75

D Proof for identification of θ^*

Proof of Lemma 3.1. Recall $G(t|\bar{Z}_t) = \exp\Big\{-\int_0^t \lambda_C(u \mid \bar{Z}_u)\, du\Big\}$. Under assumption 2.2, we have

$$G(t|\bar{Z}_t) = \exp\Big\{-\int_0^t \lambda_C(u \mid T, \bar{Z}_T)\, du\Big\} = \mathbb{P}(C > t \mid T, \bar{Z}_T), \quad t > 0. \tag{14}$$

Recall $\Delta = \mathbb{1}(T < C)$ and $X = \min(T, C)$. When $\Delta = 1$, we have $X = T$ and $\bar{Z}_X = \bar{Z}_T$. Therefore,

$$\begin{aligned} U(\theta; G, \mathcal{A}) &= \frac{\Delta \mathbb{1}(X > \tau)}{G(X|\bar{Z}_X)} \left[\mathbb{1}\{X \in \mathcal{C}_{\tau,\theta}(\bar{Z}_\tau; \mathcal{A})\} - (1-\alpha)\right] \\ &= \frac{\Delta \mathbb{1}(T > \tau)}{G(T|\bar{Z}_T)} \left[\mathbb{1}\{T \in \mathcal{C}_{\tau,\theta}(\bar{Z}_\tau; \mathcal{A})\} - (1-\alpha)\right] \\ &= \frac{\mathbb{1}(T < C)\mathbb{1}(T > \tau)}{G(T|\bar{Z}_T)} \left[\mathbb{1}\{T \in \mathcal{C}_{\tau,\theta}(\bar{Z}_\tau; \mathcal{A})\} - (1-\alpha)\right] \end{aligned}$$

By tower property of expectations,

$$\begin{aligned} \mathbb{E}\left\{U(\theta; G, \mathcal{A})\right\} &= \mathbb{E}\left(\frac{\mathbb{1}(T < C)\mathbb{1}(T > \tau)}{G(T|\bar{Z}_T)} \left[\mathbb{1}\{T \in \mathcal{C}_{\tau,\theta}(\bar{Z}_\tau; \mathcal{A})\} - (1-\alpha)\right]\right) \\ &= \mathbb{E}\left\{\mathbb{E}\left(\frac{\mathbb{1}(T < C)\mathbb{1}(T > \tau)}{G(T|\bar{Z}_T)} \left[\mathbb{1}\{T \in \mathcal{C}_{\tau,\theta}(\bar{Z}_\tau; \mathcal{A})\} - (1-\alpha)\right] \middle| T, \bar{Z}_T\right)\right\} \\ &= \mathbb{E}\left(\frac{\mathbb{E}\left\{\mathbb{1}(T < C) \mid T, \bar{Z}_T\right\}}{G(T|\bar{Z}_T)} \cdot \mathbb{1}(T > \tau) \left[\mathbb{1}\{T \in \mathcal{C}_{\tau,\theta}(\bar{Z}_\tau; \mathcal{A})\} - (1-\alpha)\right]\right) \\ &= \mathbb{E}\left(\frac{\mathbb{P}\left(C > T \mid T, \bar{Z}_T\right)}{G(T|\bar{Z}_T)} \cdot \mathbb{1}(T > \tau) \left[\mathbb{1}\{T \in \mathcal{C}_{\tau,\theta}(\bar{Z}_\tau; \mathcal{A})\} - (1-\alpha)\right]\right) \\ &= \mathbb{E}\left(\mathbb{1}(T > \tau) \left[\mathbb{1}\{T \in \mathcal{C}_{\tau,\theta}(\bar{Z}_\tau; \mathcal{A})\} - (1-\alpha)\right]\right) \qquad (15) \\ &= \mathbb{E}\{U^*(\theta; \mathcal{A})\}, \end{aligned}$$

where (15) holds by (14).

Therefore,

$$\theta^* = \inf\{\theta \in \Theta : \mathbb{E}\{U^*(\theta; \mathcal{A})\} \geq 0\} = \inf\{\theta \in \Theta : \mathbb{E}\{U(\theta; G, \mathcal{A})\} \geq 0\}.$$

□

E Additional detials and proofs for the coverage guarantees

E.1 Proof of Theorem 3.2

We first prove the following two lemmas, which will be used to prove the coverage guarantee in Theorem 3.2.

Recall that $\mathcal{D}$ denotes the observed sample, which is split into two disjoint sets $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{cal}}$. The set $\mathcal{D}_{\text{tr}}$ is used to fit the prediction model $\mathcal{A}$ and estimate the nuisance parameter G; the set $\mathcal{D}_{\text{cal}}$ is used to obtain $\hat{\theta}$.

Lemma E.1. *Suppose that Assumptions 2.2 and 2.3 hold, and that $\hat{G}(X \mid \bar{Z}_X) > \eta$ almost surely. Then for any $\epsilon \in (0, 1)$, there exists a constant $\kappa > 0$ such that with probability at least $1 - \epsilon$,*

$$\sup_{\theta} \left| \frac{1}{|\mathcal{D}_{cal}|} \sum_{i \in \mathcal{I}_{cal}} U_i(\theta; \hat{G}, \hat{\mathcal{A}}) - \mathbb{E}\left\{ U(\theta; \hat{G}, \hat{\mathcal{A}}) \middle| \mathcal{D} \right\} \right| \leq \eta^{-1} \left(\sqrt{\frac{1}{2} \log \frac{1}{\epsilon}} + \kappa \right) \frac{1}{\sqrt{|\mathcal{D}_{cal}|}}.$$

Proof of Lemma E.1. The conclusion follows by applying McDiarmid's inequality for empirical processes and Lemma 4 of Yang et al. (2024), using similar arguments as the proof of Lemma S.2.1 in the supplement of Farina et al. (2026). □

Lemma E.2. *Suppose that Assumptions 2.2 and 2.3 hold, and that $\hat{G}(X \mid \bar{Z}_X) > \eta$ almost surely. Then we have*

$$\sup_{\theta} \left| \mathbb{E}\left\{ U(\theta; \hat{G}, \hat{\mathcal{A}}) \middle| \mathcal{D} \right\} - \mathbb{E}\left\{ U(\theta; G, \hat{\mathcal{A}}) \middle| \mathcal{D} \right\} \right| \leq \eta^{-2} \|\hat{G} - G\|_2.$$

Proof of Lemma E.2. The conclusion follows by applying Jensen's inequality and Cauchy-Schwarz inequality, using similar arguments as the proof of Lemma S.2.2 in the supplement of Farina et al. (2026). □

Proof of Theorem 3.2. Recall from Section 2 that $(T_{n+1}, \bar{Z}_{\tau,n+1})$ are from the new test individual which is independent from $\mathcal{D}$. Recall from (5) the definition of U^*, and from Algorithm 1 that $\hat{\mathcal{C}}(\bar{Z}_{\tau,n+1}) = \mathcal{C}_{\tau,\hat{\theta}}(\bar{Z}_{\tau,n+1}; \hat{\mathcal{A}})$. Since $\hat{\mathcal{A}}$ and $\hat{\theta}$ are estimated from $\mathcal{D}$, they are fixed after conditional on $\mathcal{D}$. Therefore, by Lemma 3.1,

$$\begin{aligned}
&\mathbb{E}\left\{ U_{n+1}(\hat{\theta}; G, \hat{A}) \middle| \mathcal{D} \right\} \qquad (16)\\
&= \mathbb{E}\left\{ U^*_{n+1}(\hat{\theta}; \hat{A}) \middle| \mathcal{D} \right\}\\
&= \mathbb{E}\left(\mathbb{1}(T_{n+1} > \tau) \cdot \left[\mathbb{1}\{T_{n+1} \in \hat{\mathcal{C}}(\bar{Z}_{\tau,n+1})\} - (1-\alpha) \right] \middle| \mathcal{D} \right)\\
&= \mathbb{P}\left\{ T_{n+1} > \tau,\ T_{n+1} \in \hat{\mathcal{C}}(\bar{Z}_{\tau,n+1}) \middle| \mathcal{D} \right\} - (1-\alpha) \cdot \mathbb{P}(T_{n+1} > \tau)\\
&= \mathbb{P}\left\{ T_{n+1} \in \hat{\mathcal{C}}(\bar{Z}_{\tau,n+1}) \middle| T_{n+1} > \tau, \mathcal{D} \right\} \cdot \mathbb{P}(T_{n+1} > \tau) - (1-\alpha) \cdot \mathbb{P}(T_{n+1} > \tau)\\
&= \left[\mathbb{P}\left\{ T_{n+1} \in \hat{\mathcal{C}}(\bar{Z}_{\tau,n+1}) \middle| T_{n+1} > \tau, \mathcal{D} \right\} - (1-\alpha) \right] \cdot s_\tau. \qquad (17)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\mathbb{E}\left\{ U_{n+1}(\hat{\theta}; G, \hat{A}) \middle| \mathcal{D} \right\} \\
&= \frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{i \in \mathcal{I}_{\text{cal}}} U_i(\hat{\theta}; \hat{G}, \hat{\mathcal{A}}) - \left[\frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{i \in \mathcal{I}_{\text{cal}}} U_i(\hat{\theta}; \hat{G}, \hat{\mathcal{A}}) - \mathbb{E}\left\{ U_{n+1}(\hat{\theta}; G, \hat{A}) \middle| \mathcal{D} \right\} \right] \\
&\geq \frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{i \in \mathcal{I}_{\text{cal}}} U_i(\hat{\theta}; \hat{G}, \hat{\mathcal{A}}) - \left| \frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{i \in \mathcal{I}_{\text{cal}}} U_i(\hat{\theta}; \hat{G}, \hat{\mathcal{A}}) - \mathbb{E}\left\{ U_{n+1}(\hat{\theta}; G, \hat{A}) \middle| \mathcal{D} \right\} \right| \\
&\geq -\sup_{\theta} \left| \frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{i \in \mathcal{I}_{\text{cal}}} U_i(\theta; \hat{G}, \hat{\mathcal{A}}) - \mathbb{E}\left\{ U_{n+1}(\theta; G, \hat{A}) \middle| \mathcal{D} \right\} \right|, \qquad (18)
\end{aligned}$$

where (18) holds because $\sum_{i \in \mathcal{I}_{\text{cal}}} U_i(\hat{\theta}; \hat{G}, \hat{\mathcal{A}}) \geq 0$ (by the definition of $\hat{\theta}$ in Algorithm 1).

Combining (17) and (18), we have

$$\begin{aligned}
&\mathbb{P}\left\{ T_{n+1} \in \hat{\mathcal{C}}(\bar{Z}_{\tau,n+1}) \middle| T_{n+1} > \tau \right\} \\
&\geq 1 - \alpha - s_\tau^{-1} \cdot \sup_{\theta} \left| \frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{i \in \mathcal{I}_{\text{cal}}} U_i(\hat{\theta}; \hat{G}, \hat{\mathcal{A}}) - \mathbb{E}\left\{ U_{n+1}(\hat{\theta}; G, \hat{A}) \middle| \mathcal{D} \right\} \right|. \qquad (19)
\end{aligned}$$

We now bound the above supremum term. By Lemma E.1 and Lemma E.2, we have that for any $\epsilon \in (0,1)$, there exists a constant $\kappa > 0$ such that with probability at least $1 - \epsilon$,

$$\begin{aligned}
&\sup_{\theta} \left| \frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{i \in \mathcal{I}_{\text{cal}}} U_i(\theta; \hat{G}, \hat{\mathcal{A}}) - \mathbb{E}\{U(\theta; G, \hat{\mathcal{A}}) \mid \mathcal{D}\} \right| \\
&\leq \sup_{\theta} \left| \frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{i \in \mathcal{I}_{\text{cal}}} U_i(\theta; \hat{G}, \hat{\mathcal{A}}) - \mathbb{E}\{U(\theta; \hat{G}, \hat{\mathcal{A}}) \mid \mathcal{D}\} \right| \\
&\quad + \sup_{\theta} \left| \mathbb{E}\{U(\theta; \hat{G}, \hat{\mathcal{A}}) \mid \mathcal{D}\} - \mathbb{E}\{U(\theta; G, \hat{\mathcal{A}}) \mid \mathcal{D}\} \right| \\
&\leq \eta^{-1} \left(\sqrt{\frac{1}{2} \log \frac{1}{\epsilon}} + \kappa \right) \frac{1}{\sqrt{|\mathcal{D}_{\text{cal}}|}} + \eta^{-2} \|\hat{G} - G\|_2. \qquad (20)
\end{aligned}$$

Combining (19) and (20) concludes the proof. □

E.2 Additional details and proof for the coverage guarantee of HAPS-A

As in the main paper, we assume that $\hat{G}$ admits a nonnegative censoring hazard function $\hat{\lambda}_C$, with corresponding cumulative hazard

$$\hat{\Lambda}_C(t \mid \bar{Z}_t) = \int_0^t \hat{\lambda}_C(u \mid \bar{Z}_u)\, du = -\log \hat{G}(t \mid \bar{Z}_t).$$

Let $R(u) = \mathbb{1}(X \geq u)$ and let $\hat{M}_C(t) = N_C(t) - \int_0^t R(u) \hat{\lambda}_C(u \mid \bar{Z}_u)\, du$. For a function $f(u, \bar{z}_u)$, possibly depending on $\mathcal{D}_{\text{tr}}$, define the following norms using a new observation independent of $\mathcal{D}$:

$$\begin{aligned}
\|f(u, \bar{Z}_u)\|_2 &= \mathbb{E}\{R(u) |f(u, \bar{Z}_u)|^2 \mid \mathcal{D}\}^{1/2}, \\
\|f(X, \bar{Z}_X)\|_2 &= \mathbb{E}\{|f(X, \bar{Z}_X)|^2 \mid \mathcal{D}\}^{1/2}, \\
\|f\|_{\sup,2} &= \mathbb{E}\Big\{ \sup_{0 \leq u \leq X} |f(u, \bar{Z}_u)|^2 \,\Big|\, \mathcal{D} \Big\}^{1/2}.
\end{aligned}$$

To simplify notation, we use the following shorthand whenever there is no ambiguity:

$$\begin{aligned}
Y_\theta &= U^*(\theta; \hat{\mathcal{A}}) = \mathbb{1}(T > \tau)\big[\mathbb{1}\{T \in \mathcal{C}_{\tau,\theta}(\bar{Z}_\tau; \hat{\mathcal{A}})\} - (1 - \alpha)\big], \\
h_\tau(u; \theta) &= h_\tau(u, \bar{Z}_u; \theta, \hat{\mathcal{A}}) = \mathbb{E}(Y_\theta \mid \bar{Z}_u, T \geq u, \mathcal{D}), \\
\hat{h}_\tau(u; \theta) &= \hat{h}_\tau(u, \bar{Z}_u; \theta, \hat{\mathcal{A}}).
\end{aligned}$$

From their definitions, $Y_\theta \in \{0, \alpha, -(1-\alpha)\}$ and $h_\tau(u;\theta) \in [-(1-\alpha), \alpha]$. Using this shorthand and the same norm notation for these processes, we consider the following product errors:

$$r_1(\theta) = \int_0^\infty \|\hat{h}_\tau(u;\theta) - h_\tau(u;\theta)\|_2 \, \|\hat{\lambda}_C(u \mid \bar{Z}_u) - \lambda_C(u \mid \bar{Z}_u)\|_2 \, du, \tag{21}$$

$$r_2(\theta) = \|\hat{h}_\tau(X;\theta) - h_\tau(X;\theta)\|_2 \, \|\hat{\Lambda}_C(X \mid \bar{Z}_X) - \Lambda_C(X \mid \bar{Z}_X)\|_2 + \int_0^\infty \|\Lambda_C(u \mid \bar{Z}_u) - \hat{\Lambda}_C(u \mid \bar{Z}_u)\|_2 \left\| \frac{d}{du}\{\hat{h}_\tau(u;\theta) - h_\tau(u;\theta)\} \right\|_2 du, \tag{22}$$

$$r_3(\theta) = \|\hat{\Lambda}_C - \Lambda_C\|_{\sup,2} \, \|\hat{h}_\tau(\cdot;\theta) - h_\tau(\cdot;\theta)\|_{\sup,2}. \tag{23}$$

The derivative in (22) is taken along the realized covariate history. We set $r_2(\theta) = \infty$ if the function $u \mapsto \hat{h}_\tau(u;\theta) - h_\tau(u;\theta)$ fails to be absolutely continuous on $[0, X]$ with positive conditional probability given $\mathcal{D}$.

Lemma E.3. *Suppose that Assumptions 2.2 and 2.3 hold, that $\hat{G}(X \mid \bar{Z}_X) > \eta$ almost surely, and that $\hat{h}_\tau$ takes values in $[-(1-\alpha), \alpha]$. Then for any $\epsilon \in (0,1)$, with probability at least $1-\epsilon$ over $\mathcal{D}$,*

$$\sup_{\theta \in \Theta} \left| \frac{1}{|\mathcal{D}_{cal}|} \sum_{i \in \mathcal{I}_{cal}} U_{AIPCW,i}(\theta; \hat{G}, \hat{h}_\tau, \hat{\mathcal{A}}) - \mathbb{E}\{U_{AIPCW}(\theta; \hat{G}, \hat{h}_\tau, \hat{\mathcal{A}}) \mid \mathcal{D}\} \right| \leq \frac{c_\eta}{\sqrt{|\mathcal{D}_{cal}|}} \left(\sqrt{\frac{1}{2} \log \frac{1}{\epsilon}} + \kappa \right),$$

where $c_\eta = 2\max(\alpha, 1-\alpha)(2\eta^{-1} - 1)$ and $\kappa = \sqrt{\frac{1}{2}\log(2|\Theta|)}$.

Proof of Lemma E.3. Conditional on $\mathcal{D}_{\text{tr}}$, the fitted prediction model and nuisance estimates are fixed, and the calibration observations are i.i.d. For each fixed θ, the conditional mean of $U_{\text{AIPCW},i}(\theta; \hat{G}, \hat{h}_\tau, \hat{\mathcal{A}})$ given $\mathcal{D}_{\text{tr}}$ equals the population expectation $\mathbb{E}\{U_{\text{AIPCW}}(\theta; \hat{G}, \hat{h}_\tau, \hat{\mathcal{A}}) \mid \mathcal{D}\}$ in the lemma. Indeed, the latter expectation is taken over a new observation independent of $\mathcal{D}$, and all fitted functions depend only on $\mathcal{D}_{\text{tr}}$.

Expanding (9) gives

$$U_{\text{AIPCW}}(\theta; \hat{G}, \hat{h}_\tau, \hat{\mathcal{A}}) = \frac{\Delta Y_\theta + (1-\Delta)\hat{h}_\tau(X;\theta)}{\hat{G}(X \mid \bar{Z}_X)} - \int_0^X \hat{h}_\tau(u;\theta) \frac{\hat{\lambda}_C(u \mid \bar{Z}_u)}{\hat{G}(u \mid \bar{Z}_u)} \, du.$$

The numerator of the first term is bounded in absolute value by $\max(\alpha, 1-\alpha)$. For the integral term, the representation of $\hat{G}$ gives

$$\int_0^X \frac{\hat{\lambda}_C(u \mid \bar{Z}_u)}{\hat{G}(u \mid \bar{Z}_u)} \, du = \frac{1}{\hat{G}(X \mid \bar{Z}_X)} - 1.$$

By boundedness of $\hat{h}_\tau$ and positivity of $\hat{G}$,

$$\left| U_{\text{AIPCW}}(\theta; \hat{G}, \hat{h}_\tau, \hat{\mathcal{A}}) \right| \leq \max(\alpha, 1-\alpha) \left\{ \frac{2}{\hat{G}(X \mid \bar{Z}_X)} - 1 \right\} \leq \frac{c_\eta}{2},$$

uniformly over $\theta \in \Theta$.

Let $m = |\mathcal{D}_{\text{cal}}|$ and let W denote the supremum in the lemma. For each fixed θ, the estimating-function values lie in $[-c_\eta/2, c_\eta/2]$, an interval of length c_η. Applying Hoeffding's inequality conditional on $\mathcal{D}_{\text{tr}}$ and taking a union bound over Θ gives

$$\mathbb{P}(W > t \mid \mathcal{D}_{\text{tr}}) \leq 2|\Theta| \exp\left(-\frac{2mt^2}{c_\eta^2} \right), \qquad t > 0.$$

Taking $t = c_\eta\sqrt{\log(2|\Theta|/\epsilon)/(2m)}$ and using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ gives the stated bound with conditional probability at least $1-\epsilon$ given $\mathcal{D}_{\text{tr}}$. By the law of total probability, the bound holds with probability at least $1-\epsilon$ over $\mathcal{D}$. □

Lemma E.4. *Suppose that Assumptions 2.2 and 2.3 hold, that $\hat{G}(X \mid \bar{Z}_X) > \eta$ almost surely, and that $\hat{h}_\tau$ is bounded. Then, almost surely, for each $\theta \in \Theta$,*

$$\mathbb{E}\{U_{\textit{AIPCW}}(\theta; G, h_\tau, \hat{\mathcal{A}}) \mid \mathcal{D}\} = \mathbb{E}(Y_\theta \mid \mathcal{D}),$$

and

$$\left|\mathbb{E}\{U_{\textit{AIPCW}}(\theta; \hat{G}, \hat{h}_\tau, \hat{\mathcal{A}}) \mid \mathcal{D}\} - \mathbb{E}\{U_{\textit{AIPCW}}(\theta; G, h_\tau, \hat{\mathcal{A}}) \mid \mathcal{D}\}\right| \tag{24}$$

$$\le \min\{\eta^{-1} r_1(\theta),\, \eta^{-1} r_2(\theta) + (\eta^{-1} - 1) r_3(\theta)\}. \tag{25}$$

Proof of Lemma E.4. (1) We first show the first claim. Fix θ and condition on $\mathcal{D}$, so the fitted functions are fixed and the expectations are over a new observation. With some algebra, (9) can be written as

$$U_{\text{AIPCW}}(\theta; \hat{G}, \hat{h}_\tau, \hat{\mathcal{A}}) = Y_\theta - \int_0^X \frac{Y_\theta - \hat{h}_\tau(u;\theta)}{\hat{G}(u \mid \bar{Z}_u)}\, d\hat{M}_C(u).$$

Consider the filtration $\mathcal{F}_t^* = \sigma(\mathcal{D}, T, \bar{Z}_T, \{N_C(s) : 0 \le s \le t\})$. By Assumption 2.2, the intensity of $N_C(u)$ in this filtration is $R(u)\lambda_C(u \mid \bar{Z}_u)$, so M_C is a martingale. The fitted functions, Y_θ, and the covariate history through T are all known at time zero. Thus, extending the integrand by zero for $u > X$ gives a predictable process, since $R(u) = \mathbb{1}(X \ge u)$ is predictable. The integrand is bounded on $[0, X]$ because Y_θ and $\hat{h}_\tau$ are bounded and $\hat{G}(u \mid \bar{Z}_u) \ge \hat{G}(X \mid \bar{Z}_X) > \eta$ for $u \le X$. Moreover, the compensator has finite conditional expectation because N_C has at most one jump. It follows that

$$\mathbb{E}\left\{\int_0^X \frac{Y_\theta - \hat{h}_\tau(u;\theta)}{\hat{G}(u \mid \bar{Z}_u)}\, dM_C(u) \,\middle|\, \mathcal{D}\right\} = 0.$$

By Assumption 2.2 and the definition of h_τ, we have

$$\mathbb{E}(Y_\theta \mid \bar{Z}_u, X \ge u, \mathcal{D}) = \mathbb{E}(Y_\theta \mid \bar{Z}_u, T \ge u, \mathcal{D}) = h_\tau(u;\theta).$$

In addition, we have

$$d\hat{M}_C(u) = dM_C(u) + R(u)\{\lambda_C(u \mid \bar{Z}_u) - \hat{\lambda}_C(u \mid \bar{Z}_u)\}\, du.$$

Substituting this decomposition and applying iterated expectations gives

$$\begin{aligned}
&\mathbb{E}\{U_{\text{AIPCW}}(\theta; \hat{G}, \hat{h}_\tau, \hat{\mathcal{A}}) \mid \mathcal{D}\} - \mathbb{E}(Y_\theta \mid \mathcal{D}) \\
&\quad = \mathbb{E}\left\{\int_0^X \frac{\hat{h}_\tau(u;\theta) - h_\tau(u;\theta)}{\hat{G}(u \mid \bar{Z}_u)}\, d\{\Lambda_C(u \mid \bar{Z}_u) - \hat{\Lambda}_C(u \mid \bar{Z}_u)\} \,\middle|\, \mathcal{D}\right\}.
\end{aligned} \tag{26}$$

Setting $(\hat{G}, \hat{h}_\tau) = (G, h_\tau)$ makes the right-hand side of (26) zero, proving the first claim.

(2) We now show the second claim, and we will show that the left hand side (L.H.S.) of (25) can be upper bounded by both $\eta^{-1} r_1(\theta)$ and $\eta^{-1} r_2(\theta) + (\eta^{-1} - 1) r_3(\theta)$, which concludes the proof.

To obtain the first bound, write the cumulative-hazard differential as the hazard difference times du. Since $\hat{G}(u \mid \bar{Z}_u) > \eta$ for $u \le X$, the absolute value of the right-hand side of (26) is bounded by

$$\begin{aligned}
&\eta^{-1} \int_0^\infty \mathbb{E}\big[R(u)|\hat{h}_\tau(u;\theta) - h_\tau(u;\theta)|\, |\lambda_C(u \mid \bar{Z}_u) - \hat{\lambda}_C(u \mid \bar{Z}_u)| \mid \mathcal{D}\big]\, du \\
&\quad \le \eta^{-1} r_1(\theta),
\end{aligned}$$

where the last inequality follows from the Cauchy–Schwarz inequality conditional on $\mathcal{D}$ with the at-risk weight $R(u)$.

For the second bound, suppose that $r_2(\theta) < \infty$. Suppressing the covariate history in the arguments of $\hat{G}$, Λ_C, and $\hat{\Lambda}_C$, integration by parts and the product rule give

$$\begin{aligned}
&\int_0^X \frac{\hat{h}_\tau(u;\theta) - h_\tau(u;\theta)}{\hat{G}(u)}\, d\{\Lambda_C(u) - \hat{\Lambda}_C(u)\} \\
&\quad = \frac{\{\hat{h}_\tau(X;\theta) - h_\tau(X;\theta)\}\{\Lambda_C(X) - \hat{\Lambda}_C(X)\}}{\hat{G}(X)} \\
&\qquad - \int_0^X \frac{\Lambda_C(u) - \hat{\Lambda}_C(u)}{\hat{G}(u)} \frac{d}{du}\{\hat{h}_\tau(u;\theta) - h_\tau(u;\theta)\}\, du \\
&\qquad - \int_0^X \{\Lambda_C(u) - \hat{\Lambda}_C(u)\}\{\hat{h}_\tau(u;\theta) - h_\tau(u;\theta)\}\, d\{1/\hat{G}(u)\}. \qquad (27)
\end{aligned}$$

By positivity and the Cauchy–Schwarz inequality conditional on $\mathcal{D}$, the sum of the conditional expected absolute values of the first two terms in (27) is at most $\eta^{-1} r_2(\theta)$.

Since $1/\hat{G}(u)$ is nondecreasing and satisfies

$$\int_0^X d\{1/\hat{G}(u)\} = \frac{1}{\hat{G}(X)} - 1 \le \eta^{-1} - 1.$$

Therefore, the absolute value of the last integral in (27) is bounded by

$$\begin{aligned}
&\left| \int_0^X \{\Lambda_C(u) - \hat{\Lambda}_C(u)\}\{\hat{h}_\tau(u;\theta) - h_\tau(u;\theta)\}\, d\{1/\hat{G}(u)\} \right| \\
&\quad \le \sup_{0 \le u \le X} |\Lambda_C(u) - \hat{\Lambda}_C(u)| \sup_{0 \le u \le X} |\hat{h}_\tau(u;\theta) - h_\tau(u;\theta)| \int_0^X d\{1/\hat{G}(u)\} \\
&\quad \le (\eta^{-1} - 1) \sup_{0 \le u \le X} |\Lambda_C(u) - \hat{\Lambda}_C(u)| \sup_{0 \le u \le X} |\hat{h}_\tau(u;\theta) - h_\tau(u;\theta)|
\end{aligned}$$

Taking expectations conditional on $\mathcal{D}$ and applying the Cauchy–Schwarz inequality to the product of the two suprema bounds the conditional expected absolute value of last integral in (27) by

$$(\eta^{-1} - 1)\|\Lambda_C - \hat{\Lambda}_C\|_{\sup,2} \|\hat{h}_\tau(\cdot;\theta) - h_\tau(\cdot;\theta)\|_{\sup,2} = (\eta^{-1} - 1) r_3(\theta).$$

Combining the above gives the second bound. Taking the minimum of the two bounds proves the result.

When $r_2(\theta) = \infty$, the first bound alone gives the stated minimum. □

Proof for Theorem 4.1. For fixed $\theta \in \Theta$, write

$$\bar{U}_{\text{AIPCW}}(\theta) = \frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{i \in \mathcal{I}_{\text{cal}}} U_{\text{AIPCW},i}(\theta; \hat{G}, \hat{h}_\tau, \hat{\mathcal{A}}), \qquad \mu_0(\theta) = \mathbb{E}(Y_\theta \mid \mathcal{D}).$$

By feasibility of the augmented calibration problem and finiteness of Θ, the selected parameter $\hat{\theta}_{\text{AIPCW}}$ satisfies $\bar{U}_{\text{AIPCW}}(\hat{\theta}_{\text{AIPCW}}) \ge 0$. Hence

$$\begin{aligned}
\mu_0(\hat{\theta}_{\text{AIPCW}}) &\ge \bar{U}_{\text{AIPCW}}(\hat{\theta}_{\text{AIPCW}}) - \left| \bar{U}_{\text{AIPCW}}(\hat{\theta}_{\text{AIPCW}}) - \mu_0(\hat{\theta}_{\text{AIPCW}}) \right| \\
&\ge - \left| \bar{U}_{\text{AIPCW}}(\hat{\theta}_{\text{AIPCW}}) - \mu_0(\hat{\theta}_{\text{AIPCW}}) \right| \\
&\ge - \sup_{\theta \in \Theta} \left| \bar{U}_{\text{AIPCW}}(\theta) - \mu_0(\theta) \right|.
\end{aligned}$$

The triangle inequality bounds this supremum by the sum of the empirical deviation and the population bias. Lemma E.3 bounds the empirical deviation. Lemma E.4 bounds the population bias.

Thus, with probability at least $1-\epsilon$ over $\mathcal{D}$,

$$\mu_0(\hat{\theta}_{\text{AIPCW}}) \geq -\frac{c_\eta}{\sqrt{|\mathcal{D}_{\text{cal}}|}}\left(\sqrt{\frac{1}{2}\log\frac{1}{\epsilon}}+\kappa\right) - \eta^{-1}\sup_{\theta\in\Theta}\min\{r_1(\theta),\, r_2(\theta)+(1-\eta)r_3(\theta)\}.$$

Conditional on $\mathcal{D}$, $\hat{\theta}_{\text{AIPCW}}$ is fixed. Since the test individual is independent of $\mathcal{D}$, evaluating μ_0 at $\hat{\theta}_{\text{AIPCW}}$ and using the definition of Y_θ gives

$$\mu_0(\hat{\theta}_{\text{AIPCW}}) = s_\tau\left[\mathbb{P}\{T_{n+1}\in\hat{\mathcal{C}}_{\text{AIPCW}}(\bar{Z}_{\tau,n+1}) \mid T_{n+1}>\tau,\mathcal{D}\}-(1-\alpha)\right].$$

Substituting the lower bound on $\mu_0(\hat{\theta}_{\text{AIPCW}})$ and dividing by $s_\tau>0$ proves the theorem. □

F Additional details and results for simulation

F.1 Data generating mechanisms

F.1.1 DGM1

We generate a covariate process that is piecewise constant over prespecified time intervals defined by $0 = t_1 < \cdots < t_K < t_{K+1} = \infty$, where $K \geq 2$. Specifically, $Z_t = Z_{t_k}$ for all $t \in [t_k, t_{k+1})$, $k = 1, \ldots, K$. The covariate values $\{Z_{t_k}\}_{k=1}^K$ are generated from an autoregressive model of order 1:

$$Z_{t_{k+1}} = \rho Z_{t_k} + \sqrt{1-\rho^2}\,\epsilon_k, \qquad k = 1, \ldots, K,$$

where $Z_0 \sim N(0,1)$, $\epsilon_1, \ldots, \epsilon_K \overset{\text{i.i.d.}}{\sim} N(0,1)$ are independent of Z_0, and $\rho \in [-1, 1]$ controls the temporal dependence of the covariate process. In particular, $\rho = \mathrm{Corr}(Z_{t_k}, Z_{t_{k-1}})$. The setting with $\rho = 0$ corresponds to i.i.d. covariates $Z_{t_k} \sim N(0,1)$, whereas $\rho = 1$ corresponds to time-invariant covariates with $Z_{t_k} = Z_0$ for all $k \in \{1, \ldots, K\}$.

Conditional on the covariate history $\bar{Z}_{t_k}$, the event time increment T_k and the censoring time increment C_k for the interval $[t_k, t_{k+1})$ are generated independently from Weibull distributions:

$$T_k \sim \mathrm{Weibull}\big(a_{T,k}, \sigma_{T,k}(\bar{Z}_{t_k})\big)\,,$$
$$C_k \sim \mathrm{Weibull}\big(a_{C,k}, \sigma_{C,k}(\bar{Z}_{t_k})\big)\,,$$

where $a_{T,k}, a_{C,k} > 0$ are shape parameters, and the scale parameters are given by

$$\sigma_{T,k}(\bar{Z}_{t_k}) = \exp\left\{-\frac{\beta_{T,0}^{(k)} + \beta_T^{(k)} Z_{t_k}}{a_{T,k}}\right\}, \tag{28}$$

$$\sigma_{C,k}(\bar{Z}_{t_k}) = \exp\left\{-\frac{\beta_{C,0}^{(k)} + \beta_C^{(k)} Z_{t_k}}{a_{C,k}}\right\}. \tag{29}$$

Let $\zeta_1 = 1$ and $\zeta_k = \zeta_{k-1} \cdot \mathbb{1}(T_{k-1} > t_k - t_{k-1})$ for $k = 2, ..., K$ denote the indicators for whether the individual survived to t_k. Let $\zeta_1^c = 1$ and $\zeta_k^c = \zeta_{k-1}^c \cdot \mathbb{1}(C_{k-1} > t_k - t_{k-1})$ for $k = 2, ..., K$ denote the indicators that the subject hasn't experienced censoring yet till t_k. The latent event time T and censoring time C are obtained by:

$$T = \sum_{k=1}^{K} \zeta_k \cdot \min(T_k,\ t_{k+1} - t_k), \quad C = \sum_{k=1}^{K} \zeta_k^c \cdot \min(C_k,\ t_{k+1} - t_k).$$

Note that (28) and (29) are used only for illustration, under which T_k and C_k depend only on the most recent covariates Z_{t_k}, resulting in a Markov data-generating mechanism. Our framework does not rely on the Markov assumption and applies more generally to settings where T and C may depend on the entire covariate history.

Denote $\alpha_* = (\alpha_*^{(1)}, \ldots, \alpha_*^{(K)})$ and $\beta_* = (\beta_*^{(1)}, \ldots, \beta_*^{(K)})$, where $*$ is a place holder. We take $K = 3$, $(t_2, t_3) = (3, 6)$, $\rho = 0.3$, and

$$\alpha_T = (4,\ 4,\ 5), \quad \beta_{T,0} = (-8,\ -8,\ -5), \quad \beta_T = (1,\ 2,\ 3),$$
$$\alpha_C = (3,\ 3,\ 3), \quad \beta_{C,0} = (-6,\ -6,\ -5), \quad \beta_C = (2,\ 2,\ 2).$$

The resulting censoring rate is approximately 36.2%. The survival probabilities are approximately $\mathbb{P}(T > 3) = 95.8\%$ and $\mathbb{P}(T > 6) = 86.3\%$, and the corresponding at-risk probabilities are approximately $\mathbb{P}(X > 3) = 79.6\%$ and $\mathbb{P}(X > 6) = 63.9\%$.

F.1.2 DGM2

DGM2 uses the same piecewise-constant time-varying covariate process as DGM1. $[t_k, t_{k+1})$. In addition, we generate a baseline binary covariate $B \sim \mathrm{Bernoulli}(0.5)$, independent of the time-varying covariate process.

Conditional on $(B, \bar{Z}_{t_k})$, the event time is generated interval-by-interval from Weibull distributions

$$T_k \sim \text{Weibull}\{a_{T,k}, \sigma_{T,k}(B, Z_{t_k})\},$$

where

$$\sigma_{T,k}(B, Z_{t_k}) = \exp\left\{-\frac{\beta_{T,0}^{(k)} + \beta_{TL}^{(k)} Z_{t_k} + \beta_{TB}^{(k)} B}{a_{T,k}}\right\}.$$

The censoring mechanism is a subgroup mixture. For subjects with $B = 0$, the latent censoring time is generated as

$$C \sim \text{Unif}(0, 15).$$

For subjects with $B = 1$, censoring is generated interval-by-interval from Weibull waiting times

$$C_k \sim \text{Weibull}\{a_{C,k}, \sigma_{C,k}(Z_{t_k})\},$$

where

$$\sigma_{C,k}(Z_{t_k}) = \exp\left\{-\frac{\beta_{C,0}^{(k)} + \beta_{CL}^{(k)} Z_{t_k} + \beta_{CL2}^{(k)} Z_{t_k}^2}{a_{C,k}}\right\}.$$

We take

$$a_T = (4, 4, 5), \quad \beta_{T,0} = (-8, -8, -5), \quad \beta_{TL} = (1, 2, 3), \quad \beta_{TB} = (0.7, 0.9, 1.1),$$

$$a_C = (3, 3, 3), \quad \beta_{C,0} = (-8.0, -7.5, -7.0), \quad \beta_{CL} = (0.5, 0.6, 0.7), \quad \beta_{CL2} = (0.4, 0.5, 0.6).$$

The resulting censoring rate is approximately 32.1%. The survival probabilities are approximately $\mathbb{P}(T > 3) = 93.9\%$ and $\mathbb{P}(T > 6) = 81.1\%$, and the corresponding at-risk probabilities are approximately $\mathbb{P}(X > 3) = 83.5\%$ and $\mathbb{P}(X > 6) = 62.3\%$.

Under this data generating mechanism, the Cox model is misspecified for the censoring mechanism: for $B = 0$ the censoring time follows a uniform distribution, whereas for $B = 1$ the censoring hazard depends nonlinearly on Z_{t_k} through both Z_{t_k} and $Z_{t_k}^2$.

F.1.3 DGM3

DGM3 uses the same structure as DGM2 except with different parameters for generating the baseline covariate and the censoring time. The baseline covariate $B \sim \text{Bernoulli}(0.4)$, and we take coefficients

$$a_C = (1, 1, 1), \quad \beta_{C,0} = (-2.8, -2.6, -2.4), \quad \beta_{CL} = (-0.8, -1, -1.2), \quad \beta_{CL2} = (0, 0, 0).$$

The resulting censoring rate is approximately 53.7%. The survival probabilities are approximately $\mathbb{P}(T > 3) = 94.2\%$ and $\mathbb{P}(T > 6) = 82.1\%$, and the corresponding at-risk probabilities are approximately $\mathbb{P}(X > 3) = 74.9\%$ and $\mathbb{P}(X > 6) = 47.8\%$.

F.2 Details of the baseline IPCW method

The baseline IPCW method is an extension of the IPCW approach in Farina et al. (2026) to constructing two-sided prediction intervals and can also be viewed as a generalization of the weighting approach in Yi et al. (2025) by allowing arbitrary conformity scores and flexible estimation of the censoring weights. Algorithm 2 provides the details of this baseline.

Algorithm 2: Baseline IPCW method

Input: observed data $\mathcal{D} = \{O_i\}_{i=1}^n$; coverage level $1-\alpha$; prediction model for the conditional distribution of $T|Z_0$; censoring model for the conditional distribution of $C|Z_0$; baseline covariates $Z_{0,n+1}$ from a new test individual.

1. Randomly split $\mathcal{D}$ into a training set $\mathcal{D}_{\text{tr}}$ and a calibration set $\mathcal{D}_{\text{cal}}$ with almost equal size.
2. Fit the prediction model on $\mathcal{D}_{\text{tr}}$, and let $\tilde{q}_\theta(Z_{0,n+1})$ denote the estimated conditional θ-quantile function of T_{n+1} given $Z_{0,n+1}$. Consider the following class of candidate intervals:
$$\left\{\left(\tilde{q}_{1-\theta}(Z_{0,n+1}),\ \tilde{q}_\theta(Z_{0,n+1})\right) :\ \theta \in \Theta\right\},$$
where Θ is taken as a grid on $[0.5, 1]$ with increments 0.01.
3. Fit the censoring model on $\mathcal{D}_{\text{tr}}$, and let $\hat{S}_c(t|z_0)$ denote the estimator for $\mathbb{P}(C > t \mid Z_0 = z_0)$.
4. Using a grid search over Θ, compute
$$\tilde{\theta} = \inf\left\{\theta \in \Theta : \frac{1}{|\mathcal{D}_{\text{cal}}|}\sum_{i\in\mathcal{D}_{\text{cal}}} \frac{\Delta_i \mathbb{1}\left\{\tilde{q}_{1-\theta}(Z_{0,i}) \le T_i \le \tilde{q}_\theta(Z_{0,i})\right\}}{\hat{S}_c(X_i|Z_{0,i})} \ge 1-\alpha\right\}. \quad (30)$$

Output: prediction interval $\left(\tilde{q}_{1-\tilde{\theta}}(Z_{0,n+1}),\ \tilde{q}_{\tilde{\theta}}(Z_{0,n+1})\right)$.

F.3 Implementation details for simulation

F.3.1 Prediction models

Landmarking prediction models. At each prediction time τ, prediction models for $T \mid \bar{Z}_\tau, T > \tau$ are fitted using the landmarking approach, in which we consider the following models:

- 'Cox': proportional hazards model fitted using the `coxph` function in the `survival` R package.
- 'RSF': random survival forests fitted using the `rfsrc` function in the `randomForestSRC` R package, with `ntree = 1000` and `nodesize = 15`.

For a given landmarking time τ, the above models are fitted with individuals that are at risk at time τ, i.e., those with $X > \tau$. The covariates included in the proportional hazards model and random survival forests are different for different conformal methods: for HAPS and its two extensions, Z_τ is included; whereas for baseline methods, only baseline covariates Z_0 are included.

Neural-network prediction model. As an additional prediction model, we implemented a single-event neural-network dynamic prediction model inspired by Dynamic-DeepHit (Lee et al., 2019), denoted by 'DDH'. At each landmark time τ, the model takes as input the observed covariate history up to τ for subjects still at risk. A gated recurrent unit (GRU) is used to encode the history into a hidden representation, and is then used to output the estimated discrete hazards.

The residual time from τ is discretized using quantiles of the observed residual follow-up times in the training set, with at most 40 time bins. The GRU representation is then used to predict discrete hazards in each bin, which is combined to output the survival curve estimate. The network is trained by minimizing the discrete-time right-censored survival negative log-likelihood. Unlike the original Dynamic-DeepHit implementation, we do not include the ranking loss or competing-risk architecture, since our setting involves a single event type and the model is used to estimate conditional quantiles for conformal prediction.

For the neural network structure and tuning parameters, we used a one-layer GRU with hidden size 64, dropout 0.1, learning rate 10^{-3}, weight decay 10^{-5}, batch size 64, 40 residual-time bins, and 50 training epochs. Training was performed on CPU. In degenerate landmark splits with fewer than 8 at-risk training subjects or no observed events, the implementation falls back to a Kaplan-Meier estimate.

F.3.2 Nuisance models

For the censoring survival function G, we consider the following working models:

- Cox: a proportional hazards model implemented using the `coxph` function in the `survival` R package.
- RSF: a random survival forest implemented using the `rfsrc` function in the `randomForestSRC` R package, with `ntree = 1000`, `nodesize = 30`, and `nsplit = 10`.

For the event-survival nuisance functions S_k, we consider the following working models:

- Cox: a proportional hazards model implemented using the `coxph` function in the `survival` R package.
- RSF: a random survival forest implemented using `rfsrc` in the `randomForestSRC` R package, with `ntree = 1000`, `nodesize = 30`, and `nsplit = 10`.
- XGB: a gradient-boosted Cox model implemented using the `xgboost` R package with objective `survival:cox`, `nrounds = 200`, `eta = 0.05`, `max_depth = 4`, `min_child_weight = 1`, `subsample = 0.8`, `colsample_bytree = 0.8`, `lambda = 1`, and `gamma = 0`.

For the augmentation nuisance functions ξ_k, we use the following working model:

- XGB: a gradient-boosted multiclass classifier implemented using the `xgboost` R package with objective `multi:softprob`. The three classes correspond to whether the pseudo-outcome is zero, negative, or positive. We use `num_class = 3`, `eval_metric = mlogloss`, `nrounds = 200`, `eta = 0.05`, `max_depth = 4`, `min_child_weight = 1`, `subsample = 0.8`, `colsample_bytree = 0.8`, `lambda = 1`, and `gamma = 0`. Then the conditional expectation ξ_k is computed with the estimated probabilities for the three classes.

For HAPS and HAPS-DR, the censoring survival function G is estimated using the global approach described in Section C.1 when the model for G is Cox. When the model for G is RSF, we use the piecewise estimation approach described in Section C.2.1. For HAPS-A, we use the piecewise estimation approach in Section C.2.1 for estimating G under both Cox and RSF models.

F.4 Additional details and simulation results for the two simulation setups in the main paper

Simulation setups used in the main paper. The simulation figures in the main paper report two representative settings, summarized in Table 4.

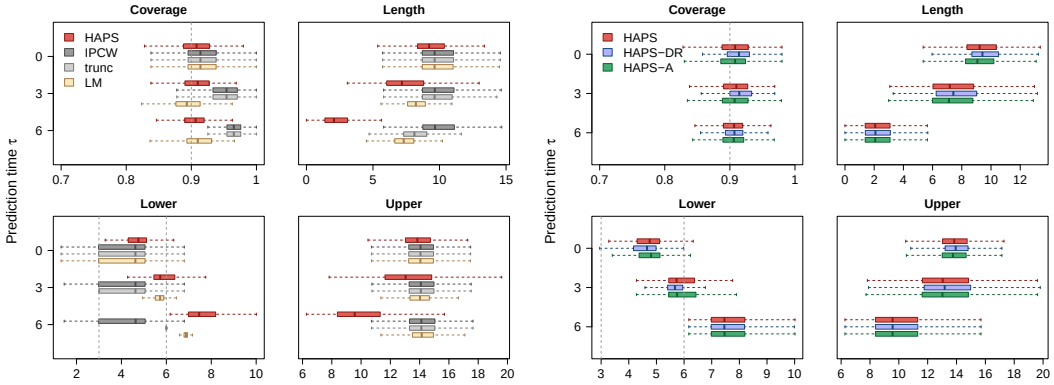
In Setup A, data are generated from DGM1, and a correctly specified Cox model is used to estimate G. This represents a setting in which a semiparametric model is used to estimate the censoring distribution and the model is correctly specified.

In Setup B, data are generated from DGM2, where the conditional censoring distribution is more complex, and flexible machine learning methods are used to estimate the censoring distribution.

Table 4: DGMs and estimation models used in the two main simulation settings.

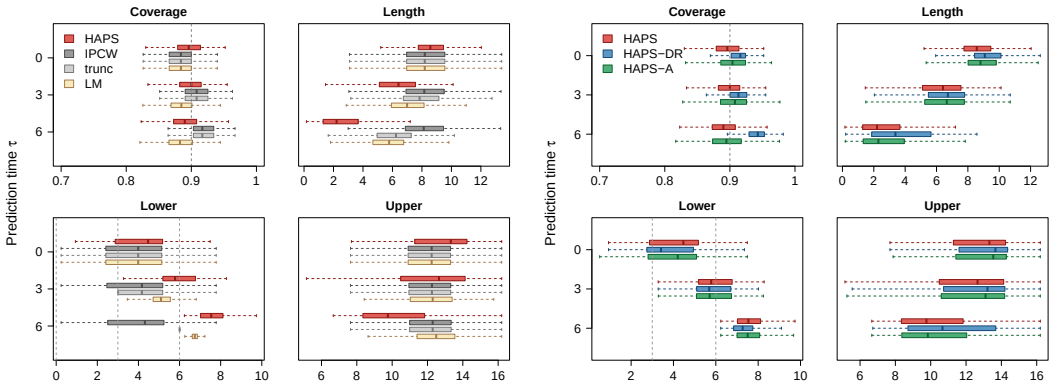
Setups	DGM	Prediction model	G	S_k	ξ_k	Baseline methods
Setup A	DGM1	Cox	Cox	Cox	XGB	Cox prediction, Cox model for G
Setup B	DGM2	DDH	RSF	XGB	XGB	RSF prediction, RSF model for G

Additional results on upper and lower endpoints. Figure 5 and Figure 6 summarize the 90% prediction intervals in simulation Setups A and B, including the lower and upper endpoints of the intervals. We observe that, for HAPS and its extensions, the lower endpoints tend to increase and the upper endpoints tend to decrease at later prediction times. This suggests that the prediction intervals become more informative over time as additional covariate history is incorporated.



(a) HAPS versus baseline methods. (b) HAPS and its two extensions.

Figure 5: Summaries for the 90% prediction intervals in Setup A of the main paper.



(a) HAPS versus baseline methods. (b) HAPS and its two extensions.

Figure 6: Summaries for the 90% prediction intervals in Setup B of the main paper.

F.5 Assessing robustness to censoring estimation errors

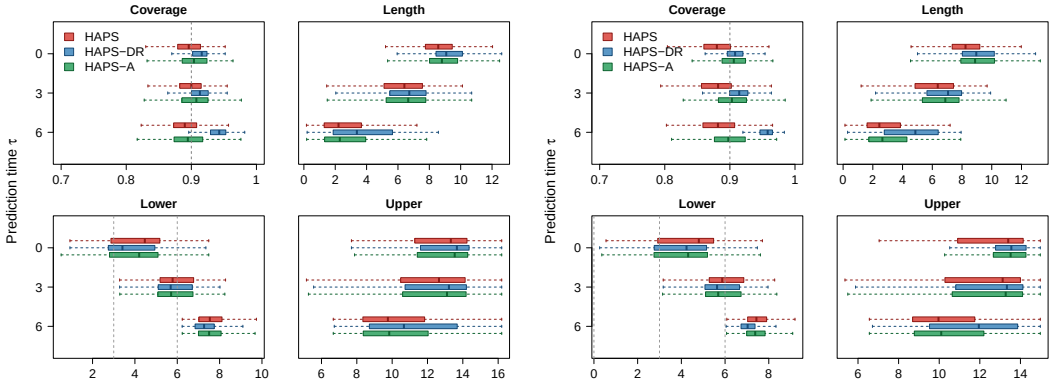
Additional simulation studies are performed to investigate the robustness of HAPS and its extensions (HAPS-DR, HAPS-A) to censoring estimation errors.

Appendix F.5.1 studies the performance of HAPS and its extensions when misspecified semiparametric models are used to estimate G, and Appendix F.5.2 studies the performance of HAPS and its extensions when flexible machine learning methods are used to estimate the censoring distribution.

F.5.1 Under censoring model misspecification

Figure 7 summarizes the prediction intervals from HAPS and its two extensions (HAPS-DR, HAPS-A) under DGM2 and DGM3 when Cox model is used to estimate the censoring distribution. In both DGMs, censoring follows a mixture distribution, so a Cox model is misspecified for the conditional censoring distribution.

The results show that HAPS noticeably undercovers in DGM3, suggesting that it can be sensitive to censoring model misspecification when a parametric or semiparametric censoring model is used. In contrast, the two extensions, HAPS-A and HAPS-DR, improve coverage, indicating robustness to censoring model misspecification. We also observe that HAPS-DR tends to overcover and produces wider intervals than HAPS and HAPS-A, which is expected from its post-processing construction.



(a) DGM2 (b) DGM3

Figure 7: Summary of 90% prediction intervals with DDH prediction model and a misspecified Cox model for censoring. HAPS uses DDH for the prediction model, Cox for the censoring model; HAPS-A uses XGB to estimate additional nuisance estimators S_k and ξ_k in the augmentation term.

F.5.2 Robustness when using machine learning methods to estimate nuisance parameters

Figure 8 summarizes the 80% prediction intervals from HAPS and its two extensions (HAPS-DR, HAPS-A) that use RSF for the prediction model, RSF for the censoring model, and XGB for the additional augmentation nuisance parameters S_k and ξ_k for HAPS-A, under DGM1-DGM3 and various sample sizes.

We observe that all three methods achieve close to nominal coverage in large sample sizes, reflecting the robustness of HAPS and its extensions when using flexible machine learning methods to estimate the nuisance parameters.

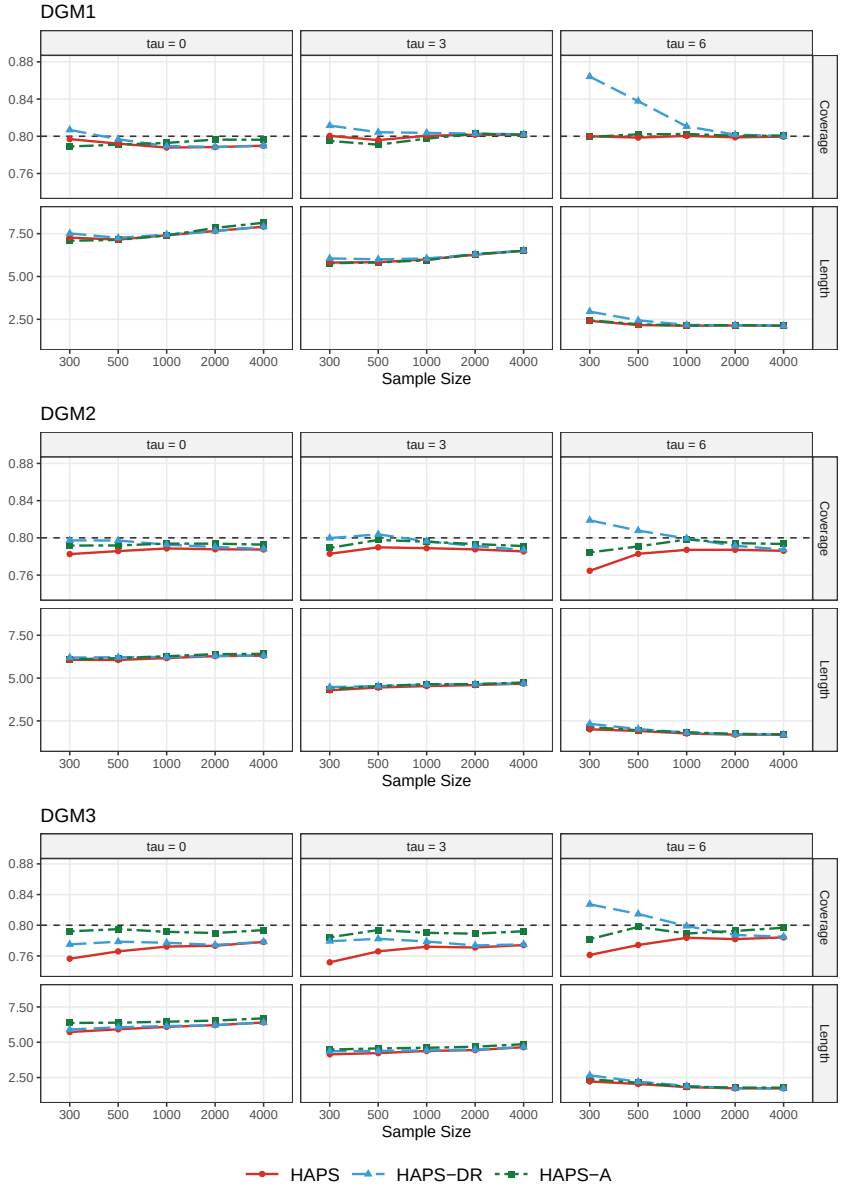


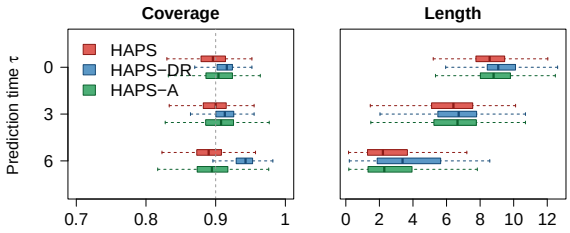
Figure 8: Mean coverage and median interval length of 80% prediction intervals from HAPS, HAPS-DR, and HAPS-A across sample sizes for DGM1–DGM3, across 200 Monte Carlo replications. The dashed horizontal line marks the nominal 80% coverage level.

F.6 Impact of prediction model accuracy on prediction intervals

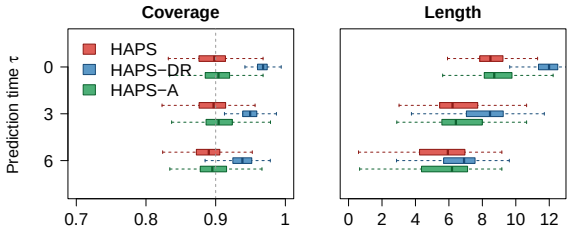
In this section, we study the impact of prediction model accuracy on prediction intervals from HAPS and its two extensions, HAPS-DR and HAPS-A. Specifically, we compare the prediction intervals obtained using the DDH prediction model described in Appendix F.3.1 with those obtained using a smaller DDH model with hidden size 32 and 20 training epochs.

Figure 9 summarizes the 90% prediction intervals under DGM2, where an RSF model is used to estimate the censoring distribution and XGB models are used to estimate the additional augmentation nuisance parameters S_k and ξ_k.

We observe that in both settings, the prediction intervals achieve close to nominal or above nominal coverage. However, when the smaller DDH model is used, the prediction intervals tend to be wider, especially at $\tau = 3$ and $\tau = 6$, reflecting the poorer prediction performance of this model. We also observe that, compared with HAPS and HAPS-A, the overcoverage of HAPS-DR becomes more pronounced when the smaller DDH model is used; and in this case, the prediction intervals from HAPS-DR are substantially wider than those from HAPS and HAPS-A.



(a) DDH model described in Section F.3.1. with hidden size 64 and 50 training epochs.



(b) Smaller DDH model with hidden size 32 and 20 training epochs.

Figure 9: Summary of 90% prediction intervals for HAPS and its two extensions in DGM2 under two DDH prediction model tuning choices. All nuisance models are held fixed: RSF is used to estimate the censoring distribution, and XGB models are used to estimate additional augmentation nuisance parameters S_k and ξ_k.

F.7 Additional baselines

To disentangle the two roles of covariate history in the empirical improvements of HAPS, we consider three additional baselines:

- LM-Z_τ: fits the same prediction model as HAPS but applies static IPCW calibration on the landmark sample at time τ, using only baseline covariates and time-varying covariates measured at τ to handle censoring.
- HAPS-Z_0: uses the covariate history for handling censoring in calibration (same as HAPS) and uses only Z_0 to fit the prediction model.
- Quantile: the interval formed by the $\alpha/2$ and $(1-\alpha/2)$ quantiles from the fitted prediction model (uncalibrated).

The LM-Z_τ method uses the same prediction model as HAPS, so any difference between the performance of HAPS and LM-Z_τ is due to using time-varying covariates for censoring adjustments.

The HAPS-Z_0 method uses only baseline covariates for fitting the prediction model and uses the same censoring adjustments as HAPS, so any difference between HAPS and HAPS-Z_0 reflects how including time-varying covariates in the prediction model affects performance.

Table 5 shows the coverage (Cov) and median interval length (Len) under the two simulation setups in Section 5, where the nominal coverage probability is 0.9. We observe that HAPS and HAPS-Z_0 achieve close to the nominal 90% coverage for all τ and in both Setup A and Setup B, while LM-Z_τ over-covers at $\tau = 3$ in Setup A, indicating that using covariate history for handling censoring in calibration ensures close to nominal coverage. In addition, the median interval length from HAPS-Z_0 is greater than that from HAPS, indicating that incorporating covariate history in the prediction model leads to shorter intervals. LM-Z_τ shows close to 90% coverage at $\tau = 6$ because, under the data generating mechanisms in both setups, Z_τ can properly address censoring in the last time interval. Finally, we observe that the uncalibrated quantile intervals can have large coverage gaps, which shows the importance of conformal calibration.

For Setup B, we also observe that HAPS-Z_0 at $\tau = 3$ and $\tau = 6$ and HAPS at $\tau = 6$ appear to be slightly undercovered, possibly due to the slow convergence rate of the RSF model for censoring. The results for HAPS in Setup B also agree with Theorem 3.2, in that the coverage gap can be larger at later τ, where s_τ tends to be smaller.

Table 5: Simulation results for Setups A and B with additional baselines at different τ. Cov: mean coverage (Monte Carlo SE); Len: median interval length. The nominal coverage probability is 0.9.

	$\tau = 3$		$\tau = 6$	
Method	Cov (SE)	Len	Cov (SE)	Len
Setup A				
HAPS	.899 (.0020)	7.31	.899 (.0017)	2.35
LM-Z_τ	**.917** (.0024)	7.81	.903 (.0018)	2.37
HAPS-Z_0	.895 (.0022)	**8.32**	.900 (.0019)	**7.04**
Quantile (Cox)	.894 (.0018)	7.24	**.883** (.0016)	2.25
Setup B				
HAPS	.891 (.0018)	6.17	**.885** (.0020)	2.57
LM-Z_τ	**.887** (.0018)	6.10	**.886** (.0020)	2.58
HAPS-Z_0	**.883** (.0021)	**7.00**	**.885** (.0021)	**5.61**
Quantile (DDH)	.905 (.0014)	6.50	**.940** (.0013)	**3.67**

F.8 Simulations under smaller sample sizes

In both data applications, 20% of individuals have been used as test data. The training and calibration sample size is 250 (125 training / 125 calibration) for the PBC data example and 744 (372 training / 372 calibration) for the colon cancer data example.

Table 6 reports additional simulation results under simulation Setup A with sample sizes $n = 250$ and $n = 750$ to mimic the sample sizes in the two applications. We report the mean coverage (Cov) and its SE across 200 Monte Carlo replications and the median length of the prediction intervals (Len). We also report the rate (Triv) at which the estimating inequality does not admit a solution, which can happen especially when the sample size is small. In such cases, the algorithm returns a trivial interval $(\tau, 50)$, where 50 is a large enough upper bound for T under this paper's data-generating mechanisms.

We observe that under the smaller sample sizes $n = 250$ and $n = 750$, HAPS still achieves close to nominal coverage, and its median interval length is much smaller than that of baselines using only baseline covariates. The conclusions match those of the simulation in Section 5 (sample size 1000; 500 training + 500 calibration).

Table 6: Simulation results under Setup A with smaller sample sizes: $n = 250$ (125 training / 125 calibration, same size as the PBC application) and $n = 750$ (375 training / 375 calibration, similar to the size of the colon cancer application). Cov: mean coverage (SE); Len: median interval length; Triv: rate of trivial intervals; –: not applicable.

	$\tau = 3$			$\tau = 6$		
Method	Cov (SE)	Len	Triv	Cov (SE)	Len	Triv
$n = 250$						
HAPS	.899 (.0038)	7.69	.060	.901 (.0036)	2.55	.095
HAPS-A	.890 (.0041)	7.40	.065	.886 (.0055)	2.50	.065
HAPS-DR	**.919** (.0028)	8.08	.060	.909 (.0028)	2.60	.095-
HAPS-Z_0	.892 (.0039)	8.54	.055	.903 (.0034)	**7.25**	.050
LM-Z_τ	.909 (.0036)	8.02	.080	.900 (.0038)	2.54	.120
LM	**.884** (.0033)	8.23	.035	.901 (.0031)	**7.22**	.050
IPCW	**.938** (.0029)	**9.86**	.046	**.959** (.0017)	**9.91**	.046
trunc	**.938** (.0029)	**9.79**	.046	**.959** (.0017)	**8.14**	.046
Quantile (Cox)	**.881** (.0028)	7.10	–	**.856** (.0029)	2.24	–
$n = 750$						
HAPS	.899 (.0022)	7.41	.000	.898 (.0021)	2.37	.005
HAPS-A	.898 (.0022)	7.39	.000	.898 (.0021)	2.37	.000
HAPS-DR	**.915** (.0019)	7.90	.000	.909 (.0018)	2.44	.005
HAPS-Z_0	.895 (.0025)	**8.48**	.000	.900 (.0022)	**7.15**	.000
LM-Z_τ	**.914** (.0026)	7.83	.015	.899 (.0017)	2.38	.000
LM	**.884** (.0022)	8.03	.000	.904 (.0020)	**7.13**	.000
IPCW	**.949** (.0020)	**9.74**	.000	**.963** (.0012)	**9.76**	.000
trunc	**.949** (.0020)	**9.71**	.000	**.963** (.0012)	**8.17**	.000
Quantile (Cox)	.893 (.0018)	7.25	–	**.879** (.0018)	2.25	–

G Additional details and results for application

G.1 Details of the data sets

Colon cancer data. We analyze data from one of the first successful trials of adjuvant chemotherapy for colon cancer (Laurie et al., 1989; Moertel et al., 1990), available as the `colon` data set in the `survival` R package and consisting of 929 colon cancer patients. We focus on predicting overall survival, i.e., time to death. The censoring rate is around 51.3%. We include the following baseline covariates: age (years), sex (male/female), treatment group (observation, Levamisole, or Levamisole+5-FU), and number of positive lymph nodes (more than 4 vs. not); as well as a time-varying covariate for recurrence, recorded as a 01 process indicating whether recurrence has occurred by each time.

PBC data. We analyze data from the Mayo Clinic trial in primary biliary cholangitis (PBC) conducted between 1974 and 1984, available as the `pbcseq` data set in `survival` R package and consisting of 312 patients with longitudinal laboratory measurements in addition to baseline covariates. Again we focus on predicting overall survival. The censoring rate is approximately 55.1%. Following Murtaugh et al. (1994), we consider the following (transformed) covariates: age, $\log$(bilirubin), $\log$(prothrombin time), $\log$(albumin), and edema score (0/0.5/1). Among these, bilirubin, prothrombin time, albumin, and edema score are time-varying.

G.2 IPCW estimator for the survivor-conditional coverage

As mentioned before, the event times in real data may be censored. As a result, the empirical survivor-conditional coverage cannot be computed directly on test data. We therefore consider its IPCW estimator in (31) below. Such an estimator for coverage probability has also been reported in Farina et al. (2026) in static conformal settings. It is a consistent estimator of the coverage probability when the censoring model is correctly specified. The details of the estimator are provided below.

For a given prediction set $\mathcal{C}(\bar{Z}_\tau)$, Let $\beta = \mathbb{P}\{T \in \mathcal{C}(\bar{Z}_\tau) \mid T > \tau\}$ denote the coverage probability among survivors at time τ. Recall from Section 3.2 the definition of G. With an estimator of G, denoted by $\hat{G}$, and a test sample indexed by $\mathcal{I}_{\text{test}}$, the IPCW estimator for β is

$$\hat{\beta} = \frac{\sum_{i \in \mathcal{I}_{\text{test}}, X_i > \tau} \mathbb{1}\{X_i \in \hat{\mathcal{C}}(\bar{Z}_\tau)\} \cdot \Delta_i / \hat{G}(X_i | \bar{Z}_{X_i,i})}{\sum_{i \in \mathcal{I}_{\text{test}}, X_i > \tau} \Delta_i / \hat{G}(X_i | \bar{Z}_{X_i,i})}. \tag{31}$$

The justification of the estimator is provided below.

Justification. Let

$$D^*(\beta) = \mathbb{1}(T > \tau) \left[\mathbb{1}\{T \in \mathcal{C}(\bar{Z}_\tau)\} - \beta \right]$$

By definition, we have that $D^*(\beta)$ is an estimating function for β in the sense that $\mathbb{E}\{D^*(\beta)\} = 0$. Under right censoring, T and $\bar{Z}_\tau$ are not always completely observed. So we apply IPCW (Robins and Rotnitzky, 1992; Rotnitzky and Robins, 2005), resulting in the following estimating function for β in the observed data:

$$D(\beta; G) = \frac{\Delta}{G(X | \bar{Z}_X)} \cdot \mathbb{1}(X > \tau) \left[\mathbb{1}\{X \in \mathcal{C}(\bar{Z}_\tau)\} - \beta \right]. \tag{32}$$

Therefore, with a random test sample indexed by $\mathcal{I}_{\text{test}}$, if we have an estimator of G, denoted by $\hat{G}$, then β can be estimated by solving the following estimating equation:

$$\sum_{i \in \mathcal{I}_{\text{test}}} D_i(\beta; \hat{G}) = 0. \tag{33}$$

By plugging in the expression (32) into (33) and solve for β, we get a closed-form solution

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i \in \mathcal{I}_{\text{test}}} \mathbb{1}(X_i > \tau) \mathbb{1}\{X_i \in \mathcal{C}(\bar{Z}_{\tau,i})\} \cdot \Delta_i / \hat{G}(X_i | \bar{Z}_{X_i,i})}{\sum_{i \in \mathcal{I}_{\text{test}}} \mathbb{1}(X_i > \tau) \cdot \Delta_i / \hat{G}(X_i | \bar{Z}_{X_i,i})} \\ &= \frac{\sum_{i \in \mathcal{I}_{\text{test}}, X_i > \tau} \mathbb{1}\{X_i \in \mathcal{C}(\bar{Z}_{\tau,i})\} \cdot \Delta_i / \hat{G}(X_i | \bar{Z}_{X_i,i})}{\sum_{i \in \mathcal{I}_{\text{test}}, X_i > \tau} \Delta_i / \hat{G}(X_i | \bar{Z}_{X_i,i})}. \end{aligned}$$

Remark. Evaluating the coverage of prediction intervals with right censored data is challenging. The evaluation strategy we use here is based on IPCW and is valid only when the censoring model is correctly specified.

In addition to IPCW, Farina et al. (2026) also considered a regression-based estimator, which requires correct modeling of the conditional event-time distribution given covariates, and an augmented IPCW (AIPCW) estimator, which is consistent if at least one of the censoring model or the event-time model is correctly specified. Extending the regression-based or AIPCW-based evaluation strategies to survivor-conditional coverage in our setting with time-varying covariates is challenging because they require modeling not only the conditional event time distribution but also the covariate process.

Alternatively, Sesia and Svetnik (2025) reported lower and upper bounds on the true miscoverage rate and used their average as a point estimate. However, this midpoint estimator is a heuristic, without a theoretical guarantee for convergence to the true miscoverage rate.

G.3 Additional results for colon cancer data application

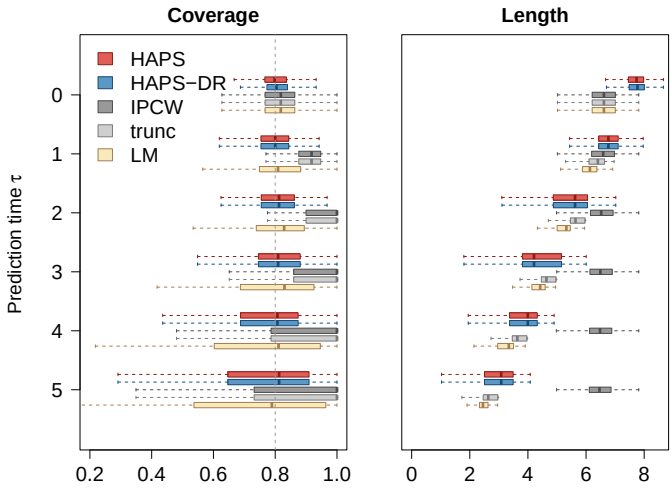


Figure 10: Summary of 80% prediction intervals across 200 random splits for the colon cancer data; coverage is IPCW-estimated.

G.4 Additional results for PBC data application

G.4.1 Ratio of prediction interval lengths from HAPS and HAPS-DR relative to baseline methods

Table 7: Ratio of median interval length for HAPS and HAPS-DR relative to baseline methods in the two data applications.

		$\tau = 3$			$\tau = 5$		
Data	Method	IPCW	trunc	LM	IPCW	trunc	LM
PBC	HAPS	0.90	0.94	0.98	0.78	0.87	0.91
	HAPS-DR	0.90	0.94	0.98	0.78	0.87	0.91
Colon	HAPS	0.56	0.71	0.78	0.40	0.74	0.87
	HAPS-DR	0.56	0.71	0.78	0.40	0.74	0.87

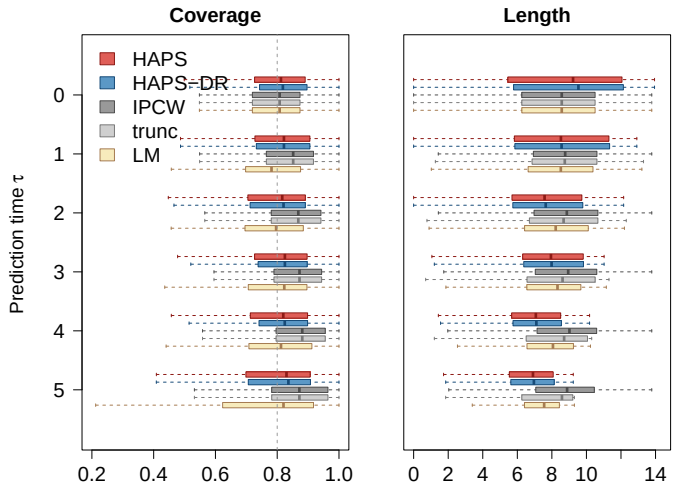


Figure 11: Summary of 80% prediction intervals across 200 random splits for the PBC data; coverage is IPCW-estimated.

G.5 Sensitivity of the coverage estimates to the censoring model

The coverages reported in the data applications are IPCW estimates, and they approximate the true coverage only when the censoring weights are close to the truth. Here we analyze the sensitivity of the coverage estimates to the censoring models used.

Coverage estimates under different censoring models. To evaluate how sensitive the coverage estimate is to the censoring model used in the IPCW evaluation weights, we compute additional IPCW coverage estimates under two censoring models:

- KM: the Kaplan–Meier estimator for the censoring survival curve, which does not incorporate covariates.
- RSF: a random survival forest with baseline covariates (since existing RSF software does not handle time-varying covariates).

Table 8 summarizes the changes in the IPCW coverage estimates for the two data applications under these evaluation censoring models, relative to the estimates reported in Section 6. Almost all of the differences are no larger than .02 in absolute value, indicating that the coverage estimates are robust to the censoring model used for evaluation. This sensitivity analysis cannot rule out potential misspecification of the censoring model used for evaluation; however, it demonstrates the stability of the IPCW coverage estimate under different censoring models.

Additional comparison to empirical coverage in the uncensored subgroup. We also report, in Table 9, the empirical coverage among the uncensored subgroup, for which the true event times are observed. However, we note that this empirical coverage is for the subgroup $\{T > \tau, T < C\}$, which is different from the survivor subgroup $\{T > \tau\}$ targeted by the proposed method. Table 9 summarizes the coverage estimates for HAPS under different estimators, as well as the number of uncensored subjects that the estimates are based on. We observe that for both applications, at $\tau = 5$, only about 10 uncensored subjects are used in the coverage estimate, reflecting that the coverage estimates may be unreliable.

Table 8: Changes in the IPCW estimate for coverage in the data applications, relative to the estimate with the Cox model in Section 6.

Data	Method	Censoring model for Eval.	$\tau = 0$	$\tau = 3$	$\tau = 5$
PBC	HAPS	KM	$-.003$	$-.002$	$-.001$
PBC	HAPS	RSF	$-.019$	$-.010$	$-.018$
PBC	IPCW	KM	$+.007$	$+.005$	$+.005$
PBC	IPCW	RSF	$-.011$	$-.015$	$-.020$
PBC	LM	KM	$+.007$	$+.020$	$+.033$
PBC	LM	RSF	$-.011$	$+.002$	$+.006$
colon	HAPS	KM	$+.001$	$+.001$	$+.007$
colon	HAPS	RSF	$+.002$	$+.002$	$+.002$
colon	IPCW	KM	$-.002$	$-.000$	$+.000$
colon	IPCW	RSF	$-.001$	$+.000$	$+.000$
colon	LM	KM	$-.002$	$-.004$	$-.011$
colon	LM	RSF	$-.001$	$-.002$	$-.005$

Table 9: The mean coverage estimate (Monte Carlo SE) of HAPS over 200 splits, $\alpha = 0.2$. "% uncens." and "# uncens." refer to the uncensored percentage and the number of uncensored subjects within the at-risk subgroup $X > \tau$ in a single test split (recall that $X = \min(T, C)$), averaged over 200 splits.

Dataset	τ	IPCW-Cox	IPCW-KM	IPCW-RSF	Complete-case	% uncens.	# uncens.
colon	0	.800 (.004)	.801 (.004)	.802 (.004)	.780 (.004)	48.8	90.2
	3	.803 (.007)	.804 (.007)	.805 (.007)	.761 (.008)	23.9	29.8
	5	.767 (.013)	.774 (.013)	.770 (.013)	.684 (.013)	9.6	9.7
PBC	0	.799 (.008)	.796 (.008)	.780 (.008)	.761 (.008)	44.8	27.7
	3	.794 (.010)	.792 (.010)	.784 (.010)	.721 (.011)	33.0	16.1
	5	.781 (.012)	.780 (.012)	.763 (.013)	.698 (.013)	26.1	10.6

G.6 Effective sample size, weight truncation, and positivity

From the coverage bound in Theorem 3.2, the coverage gap for HAPS can be large when η is small (i.e., under positivity violations for censoring) or when s_τ is small (which happens when τ is very large, i.e., prediction at the tail).

Miscoverage under positivity violations can also be viewed from the effective sample size perspective. Recall that the estimating function in (7) involves the weights $\Delta \mathbb{1}(X > \tau)/G(X \mid \bar{Z}_X)$, so only uncensored subjects with $X > \tau$ contribute to the calibration inequality. Therefore, when τ is large or when $G(X \mid \bar{Z}_X)$ is close to zero for some individuals, the effective sample size used for calibration is small. Truncating the weights can improve the effective sample size, but it can introduce bias when estimating θ, which can lead to bias in coverage. (See also Table 9 and the discussion in Appendix G.5 on the unreliability of the coverage estimates for the two applications at $\tau = 5$ due to the small number of uncensored subjects.)

With extreme weights, and especially for small α, the algorithm may return very wide or even trivial intervals. Any subject whose standardized weight exceeds α must be covered for the weighted calibration inequality to hold. Since $G(X \mid \bar{Z}_X)$ is small for large X, such subjects tend to be those with large X, and covering them forces the candidate set in (3) toward wide intervals. In small samples, the inequality may even be infeasible, in which case we return the largest candidate interval, i.e., $\theta = 1$. The IPCW coverage estimate is likewise unreliable here, as it is dominated by the few subjects with large weights.

Table 10 shows the fraction of the 200 replications that selected $\hat{\theta} = 1$, i.e., the largest interval in the candidate set. We observe that this fraction is larger with smaller α, indicating that the prediction intervals are not very informative in such settings. In Table 11, we also report the IPCW-estimated coverage and median interval length of HAPS for the two data applications under different α.

Table 10: Fraction of the 200 replications with $\hat{\theta} = 1$ in HAPS.

Data	τ	$\alpha = 0.2$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
PBC	0	.065	.540	.910	1.000
PBC	3	.095	.400	.725	.960
PBC	5	.070	.280	.605	.915
colon	0	.000	.030	.470	.995
colon	3	.065	.540	.860	.995
colon	5	.090	.390	.730	.950

Table 11: Mean IPCW-estimated coverage and median interval length (in parentheses) of HAPS among the 200 replications.

Data	τ	$\alpha = 0.2$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
PBC	0	.799 (8.36)	.904 (13.33)	.942 (13.68)	.950 (13.72)
PBC	3	.794 (7.67)	.886 (9.14)	.926 (10.54)	.936 (10.70)
PBC	5	.781 (6.62)	.897 (7.82)	.939 (8.82)	.955 (8.98)
colon	0	.800 (7.75)	.902 (8.29)	.942 (8.54)	.946 (8.59)
colon	3	.803 (4.33)	.887 (5.66)	.908 (5.88)	.915 (5.91)
colon	5	.767 (3.03)	.878 (3.57)	.915 (3.96)	.926 (3.99)

H Additional information

H.1 Computational resources for experiments

All experiments were run on CPUs; no GPU acceleration was used. Smaller-scale experiments and real-data analyses were conducted on a MacBook Air, 13-inch, 2024, with an Apple M3 chip and 16 GB memory. For the main laptop-based simulation experiments with $R = 200$, $n = 1000$, and $n_{\text{test}} = 500$, the Cox-based DGM1 experiments took a few minutes per method. The DDH experiment under DGM2 with RSF censoring estimation and XGB nuisance-parameter estimation was the most computationally intensive local run, requiring approximately 4.7 hours in total, or about 84 seconds per Monte Carlo replication, while constructing HAPS, HAPS-DR, HAPS-A, and the raw DDH intervals in the same run. Analogous DDH runs with a Cox censoring model required approximately 1 hour. The two real-data DDH analyses with $R = 200$ required approximately 6–20 minutes, depending on the dataset and censoring model.

The larger sample-size experiments were run on CPU through Open Science Grid (OSG) (Pordes et al., 2007; Sfiligoi et al., 2009; OSG, 2006, 2015) as parallel array jobs. For the appendix sample-size experiments at nominal 80% coverage, the median runtime per prediction time was approximately 32–44 seconds for $n = 300$, 55–58 seconds for $n = 500$, 117–125 seconds for $n = 1000$, 389–444 seconds for $n = 2000$, and 1243–1548 seconds for $n = 4000$. Equivalently, the median runtime for one Monte Carlo replication across the three prediction times was approximately 2.2 minutes for $n = 300$, 3 minutes for $n = 500$, 6 minutes for $n = 1000$, 20–22 minutes for $n = 2000$, and 63–79 minutes for $n = 4000$. These OSG jobs were run in parallel, so the serial runtimes should not be interpreted as the wall-clock time required for the full experiment.

H.2 Existing assets and licenses

The real-data applications use the publicly available `colon` and `pbcseq` datasets distributed with the `survival` R package. We cite the original studies associated with these datasets and use the datasets in accordance with the `survival` package license.

Our implementation builds on existing open-source software. The R code uses `survival`, `randomForestSRC`, `xgboost`, `dplyr`, `tidyr`, `ggplot2`, `scales`, and `reticulate`. The neural-network prediction model is implemented in Python through `reticulate` and uses `torch`, `numpy`, and `pandas`. These packages are used under their corresponding open-source licenses. The versions of all R and Python packages used in our experiments are listed in the `README.md` file included with the supplementary code.

The implementation of the DDH prediction model was adapted from the publicly available Dynamic-DeepHit implementation by Lee, Yoon, and van der Schaar (`https://github.com/chl8856/Dynamic-DeepHit`), associated with Lee et al. (2019). The repository did not specify an explicit software license at the time of writing.

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: The abstract and introduction state the paper's main contributions and scope: HAPS constructs prediction sets for censored time-to-event outcomes using covariate histories observed up to a decision time, targets survivor-conditional coverage, handles right censoring through IPCW, and includes two doubly robust extensions with a post-processing step or AIPCW. Theoretical results show PAAC survivor-conditional coverage under consistent censoring estimation. The empirical results from simulations and two public benchmark datasets show that HAPS and its two extensions achieve substantially narrower prediction intervals while maintaining close to nominal coverage.

   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: The paper discusses the limitations of HAPS and its two extensions in the Discussion section, including their reliance on conditional independent censoring given the observed covariate history and the positivity assumption. The paper also discusses the sensitivity of HAPS to errors in estimating the censoring distribution, which motivates the two doubly robust extensions, HAPS-DR and HAPS-A. The paper empirically evaluates the performance of HAPS and its extensions across simulations with different data-generating mechanisms, nuisance-estimation strategies, and model-specification settings.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: The paper states the assumptions required for the theoretical results, including conditional independent censoring given covariate history and positivity. Each theoretical result explicitly references the relevant assumptions, and the proofs are provided in the appendix and cross-referenced from the main text.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The paper provides the information needed to reproduce the main experimental results. The HAPS algorithm and its two extensions are described in detail, and additional implementation details for nuisance-parameter estimation, prediction-model fitting, simulation setups, and data analyses are provided in the appendix and referenced in the main text. The datasets used in the data applications are publicly available, and code for reproducing the simulation and data analysis results is included in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via

detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: The paper provides open access to anonymized code in the supplementary material, together with instructions for reproducing the main simulation and data-analysis results. The simulation data are generated from the data-generating mechanisms described in the appendix, and the data applications use publicly available benchmark datasets from the `survival` R package. The supplementary material includes scripts for implementing HAPS, HAPS-DR, HAPS-A, and the baseline methods, as well as scripts for reproducing the reported figures and tables.

   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: The paper specifies the experimental settings needed to understand and reproduce the main results. The main text describes the simulation design, prediction times, sample sizes, evaluation samples, competing methods, and the number of Monte Carlo replications. Additional details on the baseline methods, prediction models, nuisance-parameter models, model hyperparameters, and data-analysis details are provided in the appendix. The real-data analyses also specify the public datasets, covariates, random split strategy, and IPCW-based coverage-evaluation procedure.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: The paper reports variability in the experimental results using boxplots and tables for the main performance metrics, including empirical coverage, interval length, and the lower and upper endpoints of the two-sided prediction intervals. In the simulation studies, the boxplots summarize variability across 200 Monte Carlo replications; in the real-data applications, they summarize variability across 200 random train-test splits. These summaries show not only the central tendency of each metric but also its variation across repeated experimental runs, which is appropriate for assessing the stability of the empirical claims.

The boxplots use standard conventions, with the center line indicating the median, boxes indicating the interquartile range, and whiskers summarizing the remaining spread.

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: The paper reports the computational resources used to reproduce the experiments. All experiments were performed on CPUs, with smaller-scale simulations and real-data analyses conducted on a MacBook Air with an Apple M3 chip and 16 GB memory, and larger sample-size experiments conducted on CPU through Open Science Grid (OSG) (Pordes et al., 2007; Sfiligoi et al., 2009; OSG, 2006, 2015) as parallel array jobs. The appendix reports approximate runtimes for the main laptop-based simulations, real-data analyses, and larger OSG sample-size experiments, including per-replication and total-runtime summaries where applicable.

Guidelines:

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

9. **Code of ethics**

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

Answer: [Yes]

Justification: The research conforms to the NeurIPS Code of Ethics. The paper proposes a methodological framework for conformal prediction with censored time-to-event outcomes and evaluates it using simulated data and publicly available benchmark datasets. The work does not involve collecting new human-subject data, crowdsourcing, or deploying a system with direct interaction with human participants. The paper discusses relevant assumptions and limitations, including conditional independent censoring, positivity, and sensitivity to censoring-model estimation, and provides reproducibility information through detailed experimental descriptions and supplementary code. The supplementary materials are anonymized to preserve the double-blind review process.

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: The paper discusses potential positive impacts, including improved uncertainty quantification for dynamic survival prediction and more reliable decision support in high-stakes settings such as clinical care. It also discusses potential negative impacts, including over-reliance on the method without assessing assumptions, and invalid conclusions under assumption violations.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release data or models with a high risk for misuse. The work develops a statistical methodology and provides code for reproducing simulations and data analyses. The experiments use simulated data generated from the data-generating mechanisms described in the paper and publicly available benchmark datasets from the `survival` R package. The paper does not release pre-trained generative models, scraped datasets, private or sensitive datasets, or deployed predictive systems. Therefore, safeguards for responsible release of high-risk data or models are not applicable.

    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: The paper and supplementary material credit the original owners of the existing assets used in the experiments. The real-data applications use publicly available benchmark datasets from the `survival` R package, and the original studies associated with these datasets are cited. The implementation of the DDH prediction model was adapted from the publicly available Dynamic-DeepHit implementation by Lee, Yoon, and van der Schaar, which is credited in the paper. The relevant software packages and their licenses are reported in Appendix H.2.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: The paper introduces new code assets for implementing HAPS, HAPS-DR, HAPS-A, and the experimental baselines. The supplementary material provides anonymized code together with documentation for reproducing the simulation and data-analysis results, including scripts for data generation, model fitting, interval construction, and figure/table generation. No new datasets or pretrained models are released; simulated data are generated from the data-generating mechanisms described in the paper, and the real-data analyses use publicly available benchmark datasets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or new research 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: The paper does not involve crowdsourcing or new research with 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: The core method development in this paper does not involve LLMs as an important, original, or non-standard component.

    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.