PRIVACY-PROTECTED CAUSAL SURVIVAL ANALYSIS UNDER DISTRIBUTION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal inference across multiple data sources can improve the generalizability and reproducibility of scientific findings. However, for time-to-event outcomes, data integration methods remain underdeveloped, especially when populations are heterogeneous and privacy constraints prevent direct data pooling. We propose a federated learning method for estimating target site-specific causal effects in multi-source survival settings. Our approach dynamically re-weights source contributions to correct for distributional shifts, while preserving privacy. Leveraging semiparametric efficiency theory, data-adaptive weighting and flexible machine learning, the method achieves both double robustness and efficiency improvement. Through simulations and two real data applications: (i) multi-site randomized trials of monoclonal antibodies for HIV-1 prevention among cisgender men and transgender persons in the United States, Brazil, Peru, and Switzerland, as well as women in sub-Saharan Africa, and (ii) an analysis of sex disparities across biomarker groups for all-cause mortality using the "flchain" dataset, we demonstrate the validity, efficiency gains, and practical utility of the approach. Our findings highlight the promise of federated methods for efficient, privacy-preserving causal survival analysis under distribution shift.

1 INTRODUCTION

Data fusion, or data integration, can substantially enhance the generalizability, transportability, and replicability of scientific findings. By combining heterogeneous studies, researchers can gain larger and more diverse samples, extend insights beyond single settings, and strengthen causal conclusions. Yet, integration is challenging: distributional shifts in covariates, outcomes, or censoring can invalidate naive pooling, while privacy regulations such as the General Data Protection Regulation (GDPR) in Europe and the Health Insurance Portability and Accountability Act (HIPAA) in the United States often preclude sharing individual-level data.

Limitations of existing work. Most existing causal data fusion methods focus on binary or continuous outcomes (Han et al., 2025; 2024; Yang & Ding, 2019; Liu et al., 2024; Han et al., 2023; Li et al., 2023; Makhija et al., 2024; Almodóvar et al., 2024), neglecting the timing of events that is often crucial in biomedical and policy settings. Distinguishing between preventing versus delaying hospitalization, for example, has markedly different implications.

Existing extensions to survival data rely on restrictive assumptions. For example, the Cox proportional hazards (PH) model imposes a log-linear hazard structure (Hernán, 2010; Han, 2023; Nagpal et al., 2023), or the assumption of a common conditional outcome distribution (CCOD) across sites (Lee et al., 2022; Cao et al., 2024; Wen et al., 2025) may fail under heterogeneous data distributions. Violations of these assumptions yield biased estimates and inference. In addition, privacy-preserving methods avoiding sharing raw data across sites for survival outcomes remain scarce (Jia et al., 2021).

Time-to-event outcomes are typically analyzed within single-site studies using nonparametric survival curves such as the Kaplan–Meier estimator (Kaplan & Meier, 1958). With covariate-rich data, semiparametric extensions such as the Cox model (Cox, 1972; Xie & Liu, 2005; Bull & Spiegelhalter, 1997) and doubly robust estimators (Bai et al., 2013) are standard. More recently, Westling et al. (2024) integrated double machine learning (Chernozhukov et al., 2018) to flexibly estimate nuisance functions in survival analysis (Wolock et al., 2024; Cui et al., 2023; van der Laan et al., 2007).

However, these methods remain focused on single-study contexts and do not address how to combine survival data across multiple sources.

Contributions. Recognizing that pooling is often infeasible and that CCOD may not hold, we develop a federated estimator with adaptive site weighting that accommodates both continuous- and discrete-time outcomes. Our approach leverages influence function theory to construct site-specific estimators based only on local summary statistics, combined through a constrained convex optimization that upweights informative sites and downweights or excludes biased ones. We establish consistency, asymptotic normality, and conditions under which our method improves efficiency over target-only analysis. By integrating cross-fitting (Chernozhukov et al., 2018) and ensemble learning (Díaz et al., 2019; Díaz, 2020; Westling et al., 2024; van der Laan et al., 2007), our estimator avoids restrictive assumptions while retaining fast convergence rates.

We validate the method through extensive Monte Carlo simulation studies and two real applications: (i) multi-site randomized trials of monoclonal antibodies for HIV-1 prevention among cisgender men and transgender persons in the United States, Brazil, Peru, and Switzerland, as well as women in sub-Saharan Africa, and (ii) an analysis of sex disparities in all-cause mortality using the `flchain` dataset in the `survival` R package, stratified into biomarker-defined groups. Together, these examples highlight the potential of federated methods to enable efficient, privacy-preserving causal inference for time-to-event outcomes in realistic multi-source settings.

2 METHODOLOGY

2.1 PROBLEM SETUP AND TARGET ESTIMAND

Observed data. Consider K studies, each of which may be randomized or observational. For each participant, we observe baseline covariates $\mathbf{X}$, a binary treatment $A \in \{0,1\}$, and right-censored outcomes. Let $T^{(a)}$ and $C^{(a)}$ denote the potential event and censoring times under treatment $a \in \{0,1\}$. By the stable unit treatment value assumption (SUTVA) (Rosenbaum & Rubin, 1983), the observed event and censoring times are $T = AT^{(1)} + (1-A)T^{(0)}, \quad C = AC^{(1)} + (1-A)C^{(0)}$. With right censoring, however, we only observe $Y = \min(T, C)$ and $\Delta = \mathbb{I}(T \leq C)$.

Denote a copy of the independent and identically distributed (i.i.d.) data by $\mathcal{O}$. The observed data across all sites are then given by

$$\{\mathcal{O}_i = (\mathbf{X}_i, A_i, Y_i, \Delta_i, R_i) : i = 1, \ldots, n\},$$

where $R \in \{0, 1, \ldots, K-1\}$ denotes the site, with $R = 0$ indicating the target site and $R = 1, \ldots, K-1$ the external sources.

Target estimand. Our goal is to estimate the treatment-specific survival function in the target population over a finite horizon $\tau < \infty$:

$$\theta^0(t, a) = \mathbb{P}(T^{(a)} > t \mid R = 0), \quad a \in \{0,1\},\ t \in [0, \tau].$$

This function gives the probability that a target-site individual on treatment a ($a = 1$ for treated, $a = 0$ for control) survives beyond time t.

Conditional survival functions. For each site k, define the conditional survival function $S^k(t \mid a, \mathbf{X}) = \mathbb{P}(T > t \mid A = a, \mathbf{X}, R = k)$. To simultaneously accommodate continuous- and discrete-time outcomes, we use the product integral representation (Gill & Johansen, 1990):

$$S^k(t \mid a, \mathbf{X}) = \prod_{(0,t]} \{1 - \Lambda^k(du \mid a, \mathbf{X})\},$$

where $\Lambda^k(t \mid a, \mathbf{X})$ is the conditional cumulative hazard function. This notation unifies both discrete and continuous-time survival models, because in discrete time the product integral becomes the standard discrete product $\prod$, and in continuous time it becomes $\exp\{-\Lambda^k(t \mid a, \mathbf{X})\}$.

We impose three standard assumptions for causal survival analysis:

Assumption 2.1 (Unconfoundedness). $A \perp\!\!\!\perp T^{(a)} \mid \mathbf{X}, R$ and $A \perp\!\!\!\perp C^{(a)} \mid \mathbf{X}, R$.

Assumption 2.2 (Treatment-specific non-informative censoring). $C^{(a)} \perp\!\!\!\perp T^{(a)} \mid A = a, \mathbf{X}, R$.

Assumption 2.3 (Positivity)**.** There exists $\eta > 0$ such that $\mathbb{P}(R = k) \geq 1/\eta$, and for almost all $\mathbf{X}$,

$$\min_{k=0,\ldots,K-1} \{\pi^k(a \mid \mathbf{X}),\ G^k(t \mid a, \mathbf{X})\} \geq 1/\eta, \quad \min_k S^k(t \mid a, \mathbf{X}) > 0.$$

Here $\pi^k(a \mid \mathbf{X}) = \mathbb{P}(A = a \mid \mathbf{X}, R = k)$ is the site-specific propensity score for treatment $A = a$, and $G^k(t \mid a, \mathbf{X}) = \mathbb{P}(C > t \mid A = a, \mathbf{X}, R = k)$ the conditional survival function of censoring. Each treatment and censoring mechanism has non-vanishing probability, and each site contributes a non-negligible fraction of participants.

2.2 Single-site estimation

Auxiliary process. For later use, define

$$\mathcal{H}_{t,a}(\mathcal{O}; S^k, G^k) = \frac{\mathbb{I}(Y \leq t, \Delta = 1)}{S^k(Y \mid a, \mathbf{X}) G^k(Y \mid a, \mathbf{X})} - \int_0^{t \wedge Y} \frac{\Lambda^k(du \mid a, \mathbf{X})}{S^k(u \mid a, \mathbf{X}) G^k(u \mid a, \mathbf{X})}, \tag{1}$$

where $t \wedge Y = \min(t, Y)$. This functional plays a role as the inverse probability-weighted mean-zero residual (part of an augmentation term) in doubly robust estimators for right-censored data.

Efficient influence function (EIF). When using only target-site data ($R = 0$), the nonparametric EIF of $\theta^0(t, a)$ given $t \in [0, \tau]$ and $a \in \{0, 1\}$ is given by (Westling et al., 2024):

$$\varphi^{*0}_{t,a}(\mathcal{O}; \mathbb{P}) = \frac{\mathbb{I}(R = 0)}{\mathbb{P}(R = 0)} \left[\left\{ 1 - \frac{\mathbb{I}(A = a)}{\pi^0(a \mid \mathbf{X})} \mathcal{H}_{t,a}(\mathcal{O}; S^0, G^0) \right\} S^0(t \mid a, \mathbf{X}) - \theta^0(t, a) \right].$$

This representation highlights two components: (i) an *anchor term* that $S^0(t \mid a, \mathbf{X}) - \theta^0(t, a)$, which anchors estimation through the conditional survival function under an outcome model by using target data; and (ii) an *augmentation term*—the weighted part involving $\mathcal{H}_{t,a}(\mathcal{O}; S^0, G^0)$ and $\pi^0(a \mid \mathbf{X})$, which adjusts for censoring and treatment assignment.

Here, $\mathbb{P}$ in $\varphi^{*0}_{t,a}(\mathcal{O}; \mathbb{P})$ indicates that the EIF depends on nuisance functions under the true data distribution. In other words, $\varphi^{*0}_{t,a}(\mathcal{O}; \mathbb{P}) \equiv \varphi^{*0}_{t,a}(\mathcal{O}; S^0, G^0, \pi^0)$. The same convention applies to other EIFs specified later. Throughout, $\mathbb{P}_n[f(\mathcal{O})] = n^{-1} \sum_{i=1}^n f(\mathcal{O}_i)$ denotes the empirical average.

Target-only estimator. Motivated by the EIF, we define $\widehat{\theta}^0_n(t, a)$ as the solution to the estimating equation

$$0 = \mathbb{P}_n[\widehat{\varphi}^{*0}_{t,a}(\mathcal{O}; \widehat{\mathbb{P}})],$$

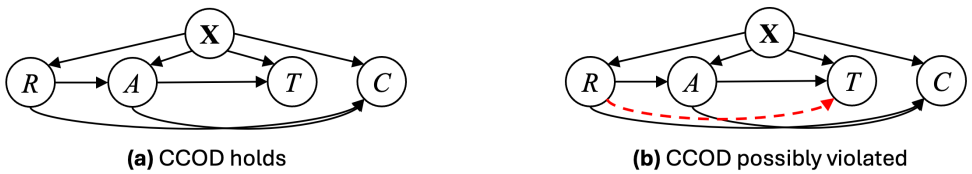
where $\widehat{\mathbb{P}}$ denotes that nuisance functions are replaced by their sample estimates. Under regularity conditions, $\widehat{\theta}^0_n(t, a)$ is regular and asymptotically linear (RAL) and achieves the semiparametric efficiency bound uniformly over $t \in [0, \tau]$ when only target-site data are available.

2.3 The CCOD Assumption

When multiple data sources are available, precision can be improved by data fusion. A common simplifying assumption is that conditional survival functions are identical across sites given covariates.

Assumption 2.4 (Common conditional outcome distribution)**.** $T^{(a)} \perp\!\!\!\perp R \mid \mathbf{X}$ for $a \in \{0, 1\}$.

Assumption 2.4 implies that $S^k(t \mid a, \mathbf{X}) = \bar{S}(t \mid a, \mathbf{X}) \equiv \mathbb{P}(T > t \mid A = a, \mathbf{X})$ for all k, while still allowing shifts in the covariate distribution $\mathbf{X}$ across sites, i.e., adjusted for covariates, the event-time distribution no longer depends on the site.



(a) CCOD holds **(b)** CCOD possibly violated

Figure 1: Data structures under and without CCOD. **(a)** Under CCOD, site R and event time T are conditionally independent given treatment A and covariates $\mathbf{X}$. **(b)** When CCOD is possibly violated, indicated by the red dashed arrow, R and T may not be conditionally independent.

Figure 1 illustrates the data structure through a directed acyclic graph (DAG), depicting the relationships among covariates $\mathbf{X}$, treatment A, site indicator R, event time T, and censoring time C, and compares scenarios with and without the CCOD assumption.

2.4 Federated estimation under distribution shifts and privacy

Motivation. In many settings, pooling individual-level data across sites is infeasible due to privacy constraints. At the same time, CCOD may fail, so naïve pooling is invalid. Still, some sites may provide information that improves estimation for the target population. We propose a federated method that adaptively re-weights source sites using only summary-level information.

2.4.1 Local site-level estimation

For each source site k, we temporarily posit a working partial CCOD assumption, $S^k(t \mid a, \mathbf{X}) = S^0(t \mid a, \mathbf{X})$ almost surely, in order to derive an EIF. This assumption is used only for formulating site-level estimators; violations will later be corrected by adaptive weighting in Section 2.4.2.

Theorem 2.5. *For $k \in \{0, 1, \ldots, K-1\}$, $\theta^0(t,a)$ is a pathwise differentiable parameter given $t \in [0, \tau]$ and $a \in \{0, 1\}$. Under the working partial CCOD assumption, the semiparametric EIF is given by $\varphi^{*k,0}_{t,a}(\mathcal{O}; \mathbb{P}) =$*

$$\frac{\mathbb{I}(R=0)}{\mathbb{P}(R=0)}\{S^0(t \mid a, \mathbf{X}) - \theta^0(t,a)\} - \frac{\mathbb{I}(R=k)}{\mathbb{P}(R=k)}\omega^{k,0}(\mathbf{X})S^k(t \mid a, \mathbf{X})\frac{\mathbb{I}(A=a)}{\pi^k(a \mid \mathbf{X})}\mathcal{H}_{t,a}(\mathcal{O}; S^k, G^k),$$

where $\omega^{k,0}(\mathbf{X}) = \mathbb{P}(\mathbf{X} \mid R=0)/\mathbb{P}(\mathbf{X} \mid R=k)$ is a density ratio comparing covariate distributions between the target site and source site k.

Local estimator. Each site computes $\widehat{\theta}^{k,0}_n(t,a)$ by solving $0 = \mathbb{P}_n[\widehat{\varphi}^{*k,0}_{t,a}(\mathcal{O}; \widehat{\mathbb{P}})]$. The proof of Theorem 2.5, along with regularity conditions and asymptotic properties of $\widehat{\theta}^{k,0}_n(t,a)$ are presented in Appendix E.1.

Interpretation. (i) The first term ("anchor") in the EIF uses target-site data ($R=0$), while the second term ("augmentation") leverages site-k data ($R=k$), adjusted by the density ratio to re-weight towards the target covariate distribution; (ii) Because $\omega^{k,0}(\mathbf{X})$ can be estimated using coarse statistics under flexible models (Han et al., 2025), individual-level covariates need not be shared; and (iii) For $S^k(t \mid a, \mathbf{X})$ in the augmentation term, we train a model on the target site and apply its predictions to site k, since $S^0(t \mid a, \mathbf{X})$ and $S^k(t \mid a, \mathbf{X})$ are exchangeable under partial CCOD. If partial CCOD is violated, we can detect site heterogeneity by the difference between $\widehat{\theta}^{k,0}_n(t,a)$ and $\widehat{\theta}^0_n(t,a)$.

2.4.2 Aggregation across sites

Data-adaptive weighting. We define the site-specific discrepancy measure $\widehat{\chi}^{k,0}_{n,t,a} = \widehat{\theta}^{k,0}(t,a) - \widehat{\theta}^0(t,a)$ and the weight vector $\boldsymbol{\eta}_{t,a} = (\eta^0_{t,a}, \eta^1_{t,a}, \ldots, \eta^{K-1}_{t,a})$. To aggregate information, we solve an ℓ_1-penalized convex optimization problem: we minimize $Q(\boldsymbol{\eta}_{t,a})$, where

$$Q(\boldsymbol{\eta}_{t,a}) = \mathbb{P}_n\left[\left\{\widehat{\varphi}^{*0}_{t,a}(\mathcal{O}; \widehat{\mathbb{P}}) - \sum_{k=1}^{K-1}\eta^k_{t,a}\widehat{\varphi}^{*k,0}_{t,a}(\mathcal{O}; \widehat{\mathbb{P}})\right\}^2\right] + \frac{1}{n}\lambda\sum_{k=1}^{K-1}|\eta^k_{t,a}|\big(\widehat{\chi}^{k,0}_{n,t,a}\big)^2, \qquad (2)$$

subject to $\eta^k_{t,a} \geq 0$ and $\sum_{k=0}^{K-1}\eta^k_{t,a} = 1$; λ is a tuning parameter that controls the bias-variance trade-off and is chosen by cross-validation.

Interpretation. The objective function balances two goals: aligning site-level EIFs with the target distribution and excluding sites that would induce bias. The quadratic term ensures that sites well-aligned with the target survival distribution contribute more to the estimation, while the ℓ_1 penalty induces sparsity by driving the weights of misaligned sites exactly to zero. This contrasts with an ℓ_2 penalty, which merely shrinks weights without fully removing them. As a result, the procedure asymptotically includes only the informative sources.

Federated estimator. The final estimator is obtained as a weighted average of the estimated local survival curves:

$$\widehat{\theta}_n^{\text{fed}}(t,a) = \sum_{k=0}^{K-1} \widehat{\eta}_{t,a}^{k}\, \widehat{\theta}_n^{k,0}(t,a).$$

The variance of $\widehat{\theta}_n^{\text{fed}}(t,a)$ can be estimated from its influence function, with the explicit formula given in Appendix E.2. Importantly, all steps require only summary-level transmission, never raw participant data.

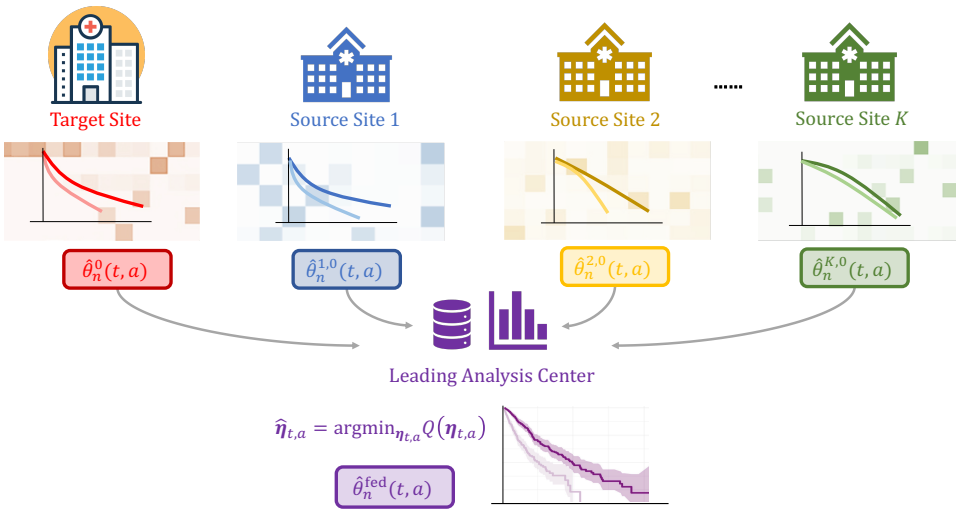


Figure 2: Flow of the Federated Estimation Algorithm. Each site first estimates its underlying survival functions locally. The related summary-level information (EIFs) is then transmitted to a leading analysis center, where it is aggregated and used to compute federated weights by minimizing the $Q(\cdot)$ function. Finally, the federated estimate is obtained through weighted averaging.

Remark 2.6. We summarize the procedure of the federated method in Algorithm 1 and illustrate its flow in Figure 2. Implementation details can be found in Appendix D, including for the cross-fitting procedure for nuisance fitting. Figure 2 emphasizes that our approach follows a *federated learning* paradigm, where raw data remain local and only summarized EIF-based quantities are transmitted to a leading analysis center(McMahan et al., 2017), thereby preserving privacy. This contrasts with *fully decentralized learning* (Lian et al., 2017), where there is no central aggregator and sites interact directly to reach consensus. Our method also differs from *meta-analysis* (Borenstein et al., 2021), which relies only on coarse population-level summaries; such information is insufficient in our setting.

2.4.3 Theoretical properties

We now summarize the main asymptotic results and efficiency gain of the federated estimator; detailed proofs are in Appendices E.1 and E.2.

Theorem 2.7 (Asymptotic distribution)**.** *If regularity conditions for local estimates (Conditions E.1–E.3 in Appendix E.1) and the adaptive weights $\widehat{\boldsymbol{\eta}}_{t,a}$ recover the oracle set of unbiased sources (Appendix E.2), then $\widehat{\theta}_n^{fed}(t,a)$, at each $(t,a) \in [0,\tau] \times \{0,1\}$, has asymptotic distribution*

$$\sqrt{n/\widehat{\mathcal{V}}_{t,a}^{fed}}\left\{\widehat{\theta}_n^{fed}(t,a) - \theta^0(t,a)\right\} \to_d \mathcal{N}(0,1).$$

where $\widehat{\mathcal{V}}_{t,a}^{fed}$ is an influence-function-based consistent estimator for the underlying asymptotic variance of $\widehat{\theta}_n^{fed}(t,a)$ (see Appendix E.2).

Corollary 2.8 (Asymptotic efficiency)**.** *The asymptotic variance $\mathcal{V}_{t,a}^{fed}$ is no greater than that of the target-only estimator $\widehat{\theta}_n^0(t,a)$. Further, if at least one source site provides a consistent estimate of $\theta^0(t,a)$, then $\widehat{\theta}_n^{fed}(t,a)$ is strictly more efficient (strictly smaller asymptotic variance).*

Algorithm 1 Federated Learning for Multi-Source Causal Survival Analysis.

1: **Input:** Multi-source right-censored data $\{\mathcal{O}_i = (\mathbf{X}_i, A_i, Y_i, \Delta_i, R_i), i = 1, \ldots, n\}$, a time horizon $\tau > 0$; a fine time grid $\{0, \epsilon, 2\epsilon, \ldots, \tau\}$ for $[0, \tau]$ with a small $\epsilon > 0$; and the number of disjoint folds into which the data are split, M.

2: **Output:** Estimated treatment-specific survival curves $\widehat{\theta}_n^{\text{fed}}(t, a)$ and its estimated variance $\widehat{\mathcal{V}}_{t,a}^{\text{fed}}$ for $a \in \{0, 1\}$ and $t \in \{0, \epsilon, 2\epsilon, \ldots, \tau\}$.

3: **for** $(t, a) \in \{0, \epsilon, 2\epsilon, \ldots, \tau\} \times \{0, 1\}$ **do**

4: Estimate the EIFs via an M-fold cross-fitting (see full detail in Algorithm 2).

5: Obtain local estimates $\widehat{\theta}_n^{k,0}(t, a)$ as solutions of $0 = \mathbb{P}_n \left[\widehat{\varphi}_{t,a}^{k,0}(\mathcal{O}; \widehat{\mathbb{P}})\right]$, for $k = 0, \ldots, K-1$.

6: Obtain the site-specific discrepancy measure (difference of the target and source estimators) as $\widehat{\chi}_{n,t,a}^{k,0} = \mathbb{P}_n \left[\widehat{\varphi}_{t,a}^{k,0}(\mathcal{O}; \widehat{\mathbb{P}}) - \widehat{\varphi}_{t,a}^{0}(\mathcal{O}; \widehat{\mathbb{P}})\right]$, for $k = 1, \cdots K-1$.

7: Solve for aggregation treatment- and time-specific weights $\widehat{\boldsymbol{\eta}}_{t,a} = (\widehat{\eta}_{t,a}^0, \widehat{\eta}_{t,a}^1, \ldots, \widehat{\eta}_{t,a}^{K-1})$ that minimize

$$Q(\boldsymbol{\eta}_{t,a}) = \mathbb{P}_n \left[\left\{ \widehat{\varphi}_{t,a}^{*0}(\mathcal{O}; \widehat{\mathbb{P}}) - \sum_{k=1}^{K-1} \eta_{t,a}^k \widehat{\varphi}_{t,a}^{*k,0}(\mathcal{O}; \widehat{\mathbb{P}}) \right\}^2 \right] + \frac{1}{n} \lambda \sum_{k=1}^{K-1} |\eta_{t,a}^k| (\widehat{\chi}_{n,t,a}^{k,0})^2,$$

subject to $0 \leq \eta_{t,a}^k \leq 1$, for all $k \in \{0, 1, \ldots, K-1\}$ and $\sum_{k=0}^{K-1} \eta_{t,a}^k = 1$, and λ is a tuning parameter chosen by cross-validation.

8: **end for**

9: **Return:**

$$\widehat{\theta}_n^{\text{fed}}(t, a) = \sum_{k=0}^{K-1} \widehat{\eta}_{t,a}^k \widehat{\theta}_n^{k,0}(t, a), \text{ and } \widehat{\mathcal{V}}_{t,a}^{\text{fed}\dagger} \text{ for } (t, a) \in \{0, \epsilon, 2\epsilon, \ldots, \tau\} \times \{0, 1\}.$$

†: $\widehat{\mathcal{V}}_{t,a}^{\text{fed}}$ is computed based on the influence function of $\widehat{\theta}_n^{\text{fed}}(t, a)$ (see Remark E.6 in Appendix E.2).

Remark 2.9 (Regularity conditions). The three regularity conditions in Appendix E.1 serve distinct roles. Condition E.1 requires local nuisance estimators to converge to general limiting functions. Condition E.2 imposes positivity by bounding nuisance functions away from 0, 1, or infinity. Condition E.3 controls three product-type errors. Together, these ensure pointwise convergence of each local estimator to a bounded and well-defined limit.

Remark 2.10 (Selection consistency). The asymptotic validity of $\widehat{\theta}_n^{\text{fed}}(t, a)$ relies on selection consistency with respect to the oracle set $\mathcal{S}_{t,a}^*$ (see below). This guarantees that post-selection inference by the influence-function-based variance estimator $\widehat{\mathcal{V}}_{t,a}^{\text{fed}}$ remains valid, even in the presence of heterogeneous or biased sources.

Remark 2.11 (Efficiency gains). To quantify the efficiency gain of $\widehat{\theta}_n^{\text{fed}}(t, a)$, let $\mathcal{S} = \{1, \ldots, K-1\}$ denote the set of source sites, and define the oracle selection space for $\boldsymbol{\eta}_{t,a}$ as

$$\mathcal{S}_{t,a}^* = \{k \in \mathcal{S} : \theta^k(t, a) = \theta^0(t, a)\},$$

and the corresponding weight space as

$$\mathbb{R}^{S_{t,a}^*} = \{\boldsymbol{\eta}_{t,a} \in \mathbb{R}^{K-1} : \eta_{t,a}^j = 0, \ \forall j \notin \mathcal{S}_{t,a}^*\}.$$

Appendix E.2 provides lemmata and conditions under which our federated estimator recovers the following oracle-optimal weights:

$$\bar{\boldsymbol{\eta}}_{t,a} = \underset{\eta_{t,a}^k = 0, \ \forall k \notin \mathcal{S}_{t,a}^*}{\arg\min} \ \mathcal{V}_{t,a}^{\text{fed}}(\boldsymbol{\eta}_{t,a}),$$

where $\mathcal{V}_{t,a}^{\text{fed}}(\boldsymbol{\eta}_{t,a})$ denotes the asymptotic variance of the federated estimator under weight vector $\boldsymbol{\eta}_{t,a}$. The target-only estimator corresponds to the special case $\boldsymbol{\eta}_{t,a} = (1, 0, \ldots, 0)$, so its variance is no larger than that of any federated estimator. If the bias term $\widehat{\chi}_{n,t,a}^{k,0}$ remains asymptotically non-zero, then $\eta_{t,a}^k \to 0$, ensuring exclusion of biased sites. Proofs are adapted from Han et al. (2023; 2025).

3 SIMULATION STUDY

We conducted simulation experiments to evaluate the performance of our federated estimator (FED) relative to three competing approaches: target-only estimation (TGT), pooling (POOL), and inverse variance weighting (IVW). The TGT method relies exclusively on target-site data ($R = 0$). POOL aggregates data from all sites without adjustment, and IVW computes a weighted average of site-specific estimators with weights proportional to the inverse of their estimated variances. This comparison allows us to assess both the efficiency gains and robustness properties of FED under varying degrees of site heterogeneity.

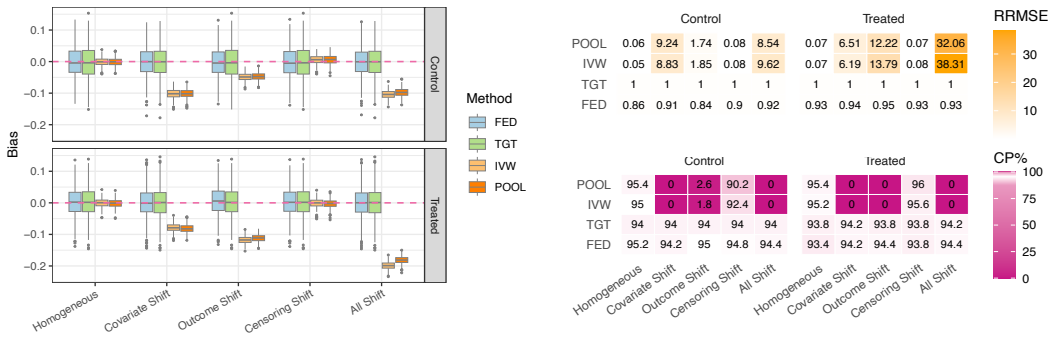
3.1 DATA GENERATING PROCESS

We conduct 500 independent Monte Carlo replications, with $n = \sum_{k=0}^{K-1} n_k$ observations distributed across $K = 5$ sites. The target site ($k = 0$) was fixed at $n_0 = 300$ observations, while source sample sizes were varied as $n_k \in \{300, 600, 1000\}$ for $k = 1, \ldots, 4$, representing small, moderate, and large external data. Covariates, treatments, and outcomes were generated according to the mechanisms described in Appendix B.1. The "truth" for each estimand was derived by averaging survival outcomes over a super-population of size $n_{\text{super}} = 10^8$ from the target distribution.

We modeled time-to-event outcomes over a one-year horizon (365 days), with administrative censoring at day 200. Performance was evaluated at days 30, 60, and 90. To investigate robustness under distribution shifts, we introduced five scenarios:

(i) Homogeneous: all sites follow identical processes;
(ii) Covariate Shift: covariate distributions vary across sites;
(iii) Outcome Shift: conditional outcome distributions differ;
(iv) Censoring Shift: censoring mechanisms vary; and
(v) All Shifts: simultaneous covariate, outcome, and censoring heterogeneity.

Figure 5 in Appendix B.1 depicts representative survival curves under outcome and covariate shifts, illustrating how site-specific heterogeneity can affect target estimation.



(A) Estimation Bias

(B) RRMSE (relative to TGT) and CP (nominal level: 95%)

Figure 3: Simulation results at day 90 with $n_0 = 300$ and $n_k = 600$ ($k = 1, 2, 3, 4$). Panel (A): bias across 500 replications. Panel (B): relative RMSE (RRMSE) compared to TGT.

3.2 PERFORMANCE METRICS AND RESULTS

We evaluated methods using three metrics: (i) **Bias**: assessed via boxplots of estimation error across 500 replications, (ii) **Relative root mean square error (RRMSE)**: defined as the RMSE of a method divided by that of TGT; values below 1 indicate efficiency gains, and (iii) **Coverage probability (CP%)**: the proportion of 95% Wald-type confidence intervals containing the truth. Values near 95% indicate valid inference. More detailed definitions of these metrics are provided in Appendix B.2. Full simulation results across all scenarios appear in Appendix B.3; here we summarize representative findings at day 90 with $n_k = 600$ for $k \geq 1$.

Main findings. As shown in Panel (A) of Figure 3, FED consistently yields negligible bias across all scenarios. In terms of efficiency, FED outperforms TGT in every heterogeneity setting: Panel (B) demonstrates 5–16% reductions in RMSE, with efficiency gains exceeding 20% in some cases (Appendix B.3). These results confirm that FED both preserves consistency and improves efficiency relative to target-only estimation.

Inferential validity. Both FED and TGT maintain CP% close to the nominal 95% across scenarios, validating our influence-function-based variance estimator. Further diagnostics, reported in Appendix Figures 6 and 7, show that federated weights $\widehat{\boldsymbol{\eta}}_{t,a}$ decrease systematically as site-specific bias measures $(\widehat{\chi}^{k,0}_{n,t,a})^2$ increase. Thus, FED adaptively upweights sites aligned with the target and downweights or excludes biased ones; the target site receives higher weights under covariate or outcome shifts, while contributions vary over time depending on alignment of survival functions.

Comparison with POOL and IVW. Although POOL and IVW exhibit lower variability (narrower boxplots), they perform poorly under Covariate, Outcome, or All shifts: bias is substantial such that RRMSE is elevated, and CP% drops far below 95%. The exception is under Censoring Shift, but this arises because censoring is treated as a nuisance function and estimated separately within each site, reducing sensitivity to between-site heterogeneity in censoring distributions.

4 Real Data Analysis

We illustrate our framework through two real-world applications. The first involves two coordinated randomized antibody-mediated prevention (AMP) trials, HVTN 704/HPTN 085 and HVTN 703/HPTN 081 (Corey et al., 2021; Ning et al., 2023), which enrolled 4,611 participants to evaluate whether a broadly neutralizing monoclonal antibody (bnAb) reduces HIV-1 acquisition. The second uses the "flchain" dataset from the `survival` R package, comprising 7,874 participants stratified into three groups defined by biomarker information, to examine sex disparities in all-cause mortality. For brevity, we focus here on the AMP trials and defer the flchain analysis to Appendix C.2.

4.1 AMP trial data

The AMP trials considered HIV diagnosis by week 80 as the primary survival endpoint, a rare event with only 3.77% incidence. Loss to follow-up was relatively low (less than 10% per treatment arm) (Corey et al., 2021). We divided participants into four regional subsets: (i) **SA:** South Africa, (ii) **OA:** other sub-Saharan African countries, (iii) **BP:** Brazil or Peru, and (iv) **US:** United States or Switzerland. Participants in (i) and (ii) were women, while those in (iii) and (iv) were cisgender men or transgender individuals, reflecting substantive population differences.

Because of event sparsity, we applied 2-fold cross-fitting. Conditional survival and censoring functions were estimated via an ensemble of Kaplan-Meier, Cox proportional hazards regression, and survival random forests, implemented in the `survSuperLearner` package (Westling et al., 2024). Propensity scores and density ratios were estimated using ensembles of logistic regression and LASSO via the Super Learner (van der Laan et al., 2007). Predictors included baseline age, a standardized machine-learning-derived HIV risk score, and body weight.

4.2 Results with South Africa as target site

We highlight results for the South Africa (SA) region as the target (Figure 4). Additional analyses treating OA, BP, or US as the target, as well as direct comparisons of regional survival curves and baseline covariates, appear in Appendix C. Table 1 shows that OA closely resembles SA, while BP and US differ markedly in baseline risk score, weight, and HIV prevalence, consistent with covariate and outcome shifts. This pattern is reflected in the federated weights: Panel (B) of Figure 4 shows SA receiving the highest weights on average, followed by OA, US, and BP.

Efficiency and coverage. Panel (A) shows that FED and TGT produce nearly identical survival curves, but FED offers narrower confidence intervals in some cases. In particular, TGT fails to yield valid intervals at certain early time points due to unstable variance estimates, while FED is able to recover narrower intervals by borrowing useful information from aligned sites. These efficiency

gains mirror our simulation findings (Section 3), highlighting the ability of FED to improve inference without introducing bias.

Comparison with POOL and IVW. Although POOL and IVW exhibit lower nominal variance (smaller relative efficiency values), both methods deviate from the trends of TGT and FED when targeting the SA population, suggesting bias under distributional shifts. Moreover, IVW fails to return estimates at early times due to extreme weights (arising from inverses of small site-specific variances), underscoring a practical limitation in survival applications.

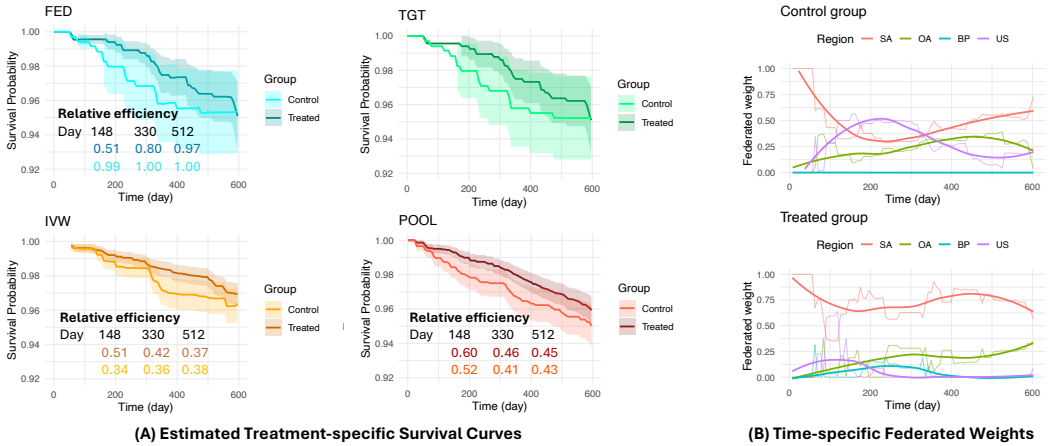


Figure 4: AMP trial results with SA as the target site. **(A):** Relative efficiency, defined as the ratio of the estimated standard deviation to that of the TGT estimator, at 148, 330, and 512 days. **(B):** Time-specific federated weights with locally weighted smoothing (Cleveland & Devlin, 1988).

5 Discussion

We developed a federated learning framework for estimation and inference of treatment-specific survival functions in a target population. By leveraging external data sources with potentially shifted covariate and outcome distributions, while preserving privacy, our method achieves efficiency gains under oracle selection and mild regularity conditions. The writing of this paper was supported by the use of ChatGPT-5.0 for language polishing (see Appendix A).

Limitations and future directions. Several limitations suggest opportunities for future work. First, although Theorem 2.7 and our simulations demonstrate efficiency gains, developing potentially more efficient covariate-adaptive weighting schemes remains crucial. Second, when data sharing is permitted but the CCOD assumption fails, it is unclear whether any method—including the pooled estimator—can outperform the target-only semiparametric efficient estimator (TGT in our simulation) and our federated approach. Third, while our time-specific weights provide flexibility, they may yield non-smooth trajectories and incur computational costs in continuous-time settings; future work should pursue smoothing strategies to capture temporal trends more efficiently. Finally, incorporating time-varying covariates could further improve efficiency by leveraging post-baseline information.

Broader extensions. Our framework naturally connects to several active areas of research. Extensions include surrogate-assisted causal inference (Han et al., 2022; Gao et al., 2024a), dynamic treatment regimes (Zhang et al., 2013), and data-driven selection of external sources (Gao et al., 2024b). It also opens opportunities for constructing two-sided conformalized prediction intervals for event times by leveraging the EIF–based conformal scores for survival outcomes developed (Farina et al., 2025) with federated learning for predicting missing outcomes (Liu et al., 2024). Our approach could be adapted to alternative estimands such as restricted mean survival time (Han, 2023), and under complex regimes such as competing risks (Lok et al., 2018) or left-truncation (Han, 2024; Wang et al., 2024). These directions highlight the broader potential of federated methods for causal survival analysis under distributional shift.

ETHICS STATEMENT

This work complies with the ICLR Code of Ethics. We used only publicly available datasets with appropriate licenses and did not involve human subjects or sensitive personal information. We acknowledge potential risks of misuse (e.g., unfair application, misinterpretation, or unintended deployment beyond the intended research scope) and discuss limitations and safeguards in the paper. All results are reported transparently, and code will be released to support reproducibility.

REPRODUCIBILITY STATEMENT

All simulation studies and real data analyses were performed using the statistical language `R` (version 4.4.2). The dependent `R` packages include: `CFsurvival`, `survSuperLearner`, `superLearner` (version 2.0.29), `glmnet` (version 4.1.8), `caret` (version 6.0.94) and `tidyverse` (version 2.0.0). To enhance computational efficiency, parallel computing packages `foreach` (version 1.5.2) and `doParellel` (version 1.0.17) were employed. The replication of simulations was carried out using 200 CPU cores by a high performance computing cluster.

We provide an anonymous GitHub repository containing all code for our simulations and data analysis: `https://anonymous.4open.science/r/FuseSurvSubmission-3D16/README.md`. All source code and software (R package) will be made publicly available through the author's Github upon acceptance of the paper.

The two real datasets are publicly available. The AMP trial data can be found at `https://atlas.scharp.org/project/HVTN%20Public%20Data/HVTN%20704%20HPTN%20085%20and%20HVTN%20703%20HPTN%20081%20AMP/begin.view`, and the "flchain" data can be found at `https://rdrr.io/cran/survival/man/flchain.html` or by typing command `data(flchain)` in R after loading the `survival` R package.

REFERENCES

Alejandro Almodóvar, Juan Parras, and Santiago Zazo. Propensity weighted federated learning for treatment effect estimation in distributed imbalanced environments. *Computers in Biology and Medicine*, 178:108779, 2024.

Peter C Austin. Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Statistics in medicine*, 31(29):3946–3958, 2012.

Xiaofei Bai, Anastasios A Tsiatis, and Sean M O'Brien. Doubly-robust estimators of treatment-specific survival distributions in observational studies with stratified sampling. *Biometrics*, 69(4): 830–839, 2013.

Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer, 1993.

Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. *Introduction to meta-analysis*. John wiley & sons, 2021.

Kate Bull and David J Spiegelhalter. Tutorial in biostatistics survival analysis in observational studies. *Statistics in medicine*, 16(9):1041–1074, 1997.

Zhiqiang Cao, Youngjoo Cho, and Fan Li. Transporting randomized trial results to estimate counterfactual survival functions in target populations. *Pharmaceutical Statistics*, 2024.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2018.

William S Cleveland and Susan J Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association*, 83(403):596–610, 1988.

Lawrence Corey, Peter B Gilbert, Michal Juraska, David C Montefiori, Lynn Morris, Shelly T Karuna, Srilatha Edupuganti, Nyaradzo M Mgodi, Allan C Decamp, Erika Rudnicki, et al. Two randomized trials of neutralizing antibodies to prevent hiv-1 acquisition. *New England Journal of Medicine*, 384(11):1003–1014, 2021.

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

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

Iván Díaz. *survtmlerct: Efficiency guarantees for covariate adjustment in RCTs with survival outcomes*, 2020. URL `https://github.com/idiazst/survtmlerct`. R package version 1.0.0.

Iván Díaz, Elizabeth Colantuoni, Daniel F Hanley, and Michael Rosenblum. Improved precision in the analysis of randomized trials with survival outcomes, without assuming proportional hazards. *Lifetime data analysis*, 25:439–468, 2019.

Angela Dispenzieri, Jerry A Katzmann, Robert A Kyle, Dirk R Larson, Terry M Therneau, Colin L Colby, Raynell J Clark, Graham P Mead, Shaji Kumar, L Joseph Melton III, et al. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. In *Mayo Clinic Proceedings*, volume 87, pp. 517–523. Elsevier, 2012.

Xianqiu Fan, Jun Cheng, Hailing Wang, Bin Zhang, and Zhenzhen Chen. A fast trans-lasso algorithm with penalized weighted score function. *Computational Statistics & Data Analysis*, 192:107899, 2024.

Rebecca Farina, Eric J Tchetgen Tchetgen, and Arun Kumar Kuchibhotla. Doubly robust and efficient calibration of prediction sets for censored time-to-event outcomes. *arXiv preprint arXiv:2501.04615*, 2025.

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

Chenyin Gao, Shu Yang, Mingyang Shan, Wenyu YE, Ilya Lipkovich, and Douglas Faries. Improving randomized controlled trial analysis via data-adaptive borrowing. *Biometrika*, pp. asae069, 2024b.

Richard D Gill and Soren Johansen. A survey of product-integration with a view toward application in survival analysis. *The annals of statistics*, 18(4):1501–1555, 1990.

Larry Han. Breaking free from the hazard ratio: Embracing the restricted mean survival time in clinical trials, 2023.

Larry Han. Truncated, not forgotten—handling left truncation in time-to-event studies, 2024.

Larry Han, Xuan Wang, and Tianxi Cai. Identifying surrogate markers in real-world comparative effectiveness research. *Statistics in Medicine*, 41(26):5290–5304, 2022.

Larry Han, Zhu Shen, and Jose Zubizarreta. Multiply robust federated estimation of targeted average treatment effects. *Advances in Neural Information Processing Systems*, 36:70453–70482, 2023.

Larry Han, Yige Li, Bijan Niknam, and José R Zubizarreta. Privacy-preserving, communication-efficient, and target-flexible hospital quality measurement. *The Annals of Applied Statistics*, 18(2): 1337–1359, 2024.

Larry Han, Jue Hou, Kelly Cho, Rui Duan, and Tianxi Cai. Federated adaptive causal estimation (face) of target treatment effects. *Journal of the American Statistical Association*, (just-accepted): 1–25, 2025.

Miguel A Hernán. The hazards of hazard ratios. *Epidemiology*, 21(1):13–15, 2010.

Beilin Jia, Donglin Zeng, Jason JZ Liao, Guanghan F Liu, Xianming Tan, Guoqing Diao, and Joseph G Ibrahim. Inferring latent heterogeneity using many feature variables supervised by survival outcome. *Statistics in medicine*, 40(13):3181–3195, 2021.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

Robert A Kyle, Terry M Therneau, S Vincent Rajkumar, Dirk R Larson, Matthew F Plevak, Janice R Offord, Angela Dispenzieri, Jerry A Katzmann, and L Joseph Melton III. Prevalence of monoclonal gammopathy of undetermined significance. *New England Journal of Medicine*, 354(13):1362–1369, 2006.

Dasom Lee, Shu Yang, and Xiaofei Wang. Doubly robust estimators for generalizing treatment effects on survival outcomes from randomized controlled trials to a target population. *Journal of Causal Inference*, 10(1):415–440, 2022.

Fan Li and Fan Li. Using propensity scores for racial disparities analysis. *Observational Studies*, 9 (1):59–68, 2023.

Sai Li, Tianxi Cai, and Rui Duan. Targeting underrepresented populations in precision medicine: A federated transfer learning approach. *The Annals of Applied Statistics*, 17(4):2970–2992, 2023.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in neural information processing systems*, 30, 2017.

Jiajun Liu, Yi Liu, Yunji Zhou, and Roland A Matsouaka. Assessing racial disparities in healthcare expenditure using generalized propensity score weighting. *BMC Medical Research Methodology*, 25(1):64, 2025.

Yi Liu, Alexander W Levis, Sharon-Lise Normand, and Larry Han. Multi-source conformal inference under distribution shift. *Proceedings of machine learning research*, 235:31344, 2024.

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*, 24:201–223, 2018.

Disha Makhija, Joydeep Ghosh, and Yejin Kim. Federated learning for estimating heterogeneous treatment effects. *arXiv preprint arXiv:2402.17705*, 2024.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Chirag Nagpal, Vedant Sanil, and Artur Dubrawski. Recovering sparse and interpretable subgroups with heterogeneous treatment effects with censored time-to-event outcomes. *arXiv preprint arXiv:2302.12504*, 2023.

Xi Ning, Yinghao Pan, Yanqing Sun, and Peter B Gilbert. A semiparametric cox–aalen transformation model with censored data. *Biometrics*, 79(4):3111–3125, 2023.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Mark J van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1):1–21, 2007.

AW Van der Vaart and JA Wellner. *Weak Convergence and Empirical Processes.* Springer & Verlag New York, 1996.

Yuyao Wang, Andrew Ying, and Ronghui Xu. Doubly robust estimation under covariate-induced dependent left truncation. *Biometrika*, pp. asae005, 2024.

Lan Wen, Jon A Steingrimsson, Sarah E Robertson, and Issa J Dahabreh. Multi-source analyses of average treatment effects with failure time outcomes. *Lifetime Data Analysis*, pp. 1–29, 2025.

Ted Westling, Mark J van der Laan, and Marco Carone. Correcting an estimator of a multivariate monotone function with isotonic regression. *Electronic journal of statistics*, 14(2):3032, 2020.

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.

Charles J Wolock, Peter B Gilbert, Noah Simon, and Marco Carone. A framework for leveraging machine learning tools to estimate personalized survival curves. *Journal of Computational and Graphical Statistics*, pp. 1–11, 2024.

Jun Xie and Chaofeng Liu. Adjusted kaplan–meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in medicine*, 24(20):3089–3110, 2005.

Shu Yang and Peng Ding. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 2019.

Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3):681–694, 2013.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

A USE OF LLMS

We acknowledge the use of ChatGPT-5.0 exclusively for language polishing and grammatical corrections. No large language models (LLMs) were used for any other aspects of this work. The research ideas, conceptualization, methodology development, and all experiments are entirely original contributions of the authors.

B SIMULATION DETAILS AND ADDITIONAL RESULTS

B.1 DETAILS OF DATA GENERATING PROCESS

Three covariates X_1, X_2, and X_3 are sampled as transformations of Beta random variables with site-specific parameters:

$$\begin{aligned}
X_1 &\sim 33 \cdot \text{Beta}(1.1 - 0.05\gamma(k), 1.1 + 0.2\gamma(k)) + 9 + 2\gamma(k), \\
X_2 &\sim 52 \cdot \text{Beta}(1.5 + (X_1 + 0.5\gamma(k))/20, 4 + 2\gamma(k)) + 7 + 2\gamma(k), \\
X_3 &\sim (4 + 2\gamma(k)) \cdot \text{Beta}(1.5 + |X_1 - 50 + 3\gamma(k)|/20, 3 + 0.1\gamma(k)),
\end{aligned}$$

where $\gamma(k)$ represents some function of site k, specified later. We then generate the treatment assignment probabilities $\pi(\mathbf{X})$ using the logistic function:

$$\text{logit}(\pi(\mathbf{X})) = -1.05 + \log\left(1.3 + \exp(-12 + X_1/10) + \exp(-2 + X_2/12) + \exp(-2 + X_3/3)\right),$$

and treatments A are sampled as $A \sim \text{Bernoulli}(\pi(\mathbf{X}))$.

Next, we consider the mechanisms of event and censoring times. The hazard rates for event times and censoring times are given by the following $\exp(h_t)$ and $\exp(h_c)$, respectively, where $h_t = -5.02 + 0.1(X_1 - 25) - 0.1(X_2 - 25) + 0.05(X_3 - 2) + D_T(k) \cdot 0.1(X_2 - 25) + A \cdot \delta_T(k) \cdot 0.1(X_1 + X_2 + X_3 - 50)$, and $h_c = -4.87 + 0.01(X_1 - 25) - 0.02(X_2 - 25) + 0.01(X_3 - 2) - D_C(k) \cdot 0.1(X_2 - 25) + A \cdot \delta_C(k) \cdot 0.1(X_1 + X_2 + X_3 - 50)$.

Here, $D_T(k)$, $D_C(k)$, $\delta_T(k)$ and $\delta_C(k)$ are some site-specific indicators, specified later, for varying the treatment effects and trends of survival curves for different sites. Then, event times and censoring times are sampled as:

$$T = \left(-\frac{\log(U_1)}{\exp(h_t) \cdot \lambda}\right)^{1/\rho}, \quad C = \left(-\frac{\log(U_2)}{\exp(h_c) \cdot \lambda}\right)^{1/\rho},$$

with $\rho = 1.2$, $\lambda = 0.6$, and $U_1, U_2 \sim \text{Uniform}(0, 1)$. This technique follows Austin (2012). Thus, the observed times and event indicators are $Y = \min(T, C)$, $\Delta = \mathbb{I}(T \leq C)$, respectively.

Under this data generating process (DGP), the event time is generated to mimic days in a year (365 days), and we truncate the censoring time at $\tau = 200$ days to mimic the end of follow-up in survival analysis. Our DGP allows the following scenarios based on site-specific distributional heterogeneity:

- **Homogeneous**: Homogeneous covariates and hazard rates across sites. We let $\gamma(k) = D_T(k) = D_C(k) = \delta_T(k) = \delta_C(k) = 0$ for $k = 0, 1, \ldots, 4$.
- **Covariate Shift**: Covariates X_1, X_2, and X_3 vary across sites. We let $\gamma(k) = k$ and $D_T(k) = D_C(k) = \delta_T(k) = \delta_C(k) = 0$, for $k = 0, 1, \ldots, 4$.
- **Outcome Shift**: Conditional outcome distribution varies across sites. We assign $\gamma(k) = 0$, $D_T(k) = \delta_T(k) = k$, and $D_C(k) = \delta_C(k) = 0$ for $k = 0, 1, \ldots, 4$.
- **Censoring Shift**: Censoring mechanism varies across sites. We let $\gamma(k) = 0$, $D_T(k) = \delta_T(k) = 0$ and $D_C(k) = \delta_C(k) = k$, for $k = 0, 1, \ldots, 4$.
- **All Shift**: Covariates and both event and censoring effects vary across sites. We let $\gamma(k) = D_T(k) = D_C(k) = \delta_T(k) = \delta_C(k) = k$, for $k = 0, 1, \ldots, 4$.

Figure 5 below plots the true treatment-specific survival curves under the Covariate Shift and Outcome Shift scenarios, as defined by our designed DGPs, to illustrate the effect of site differences on survival outcomes.

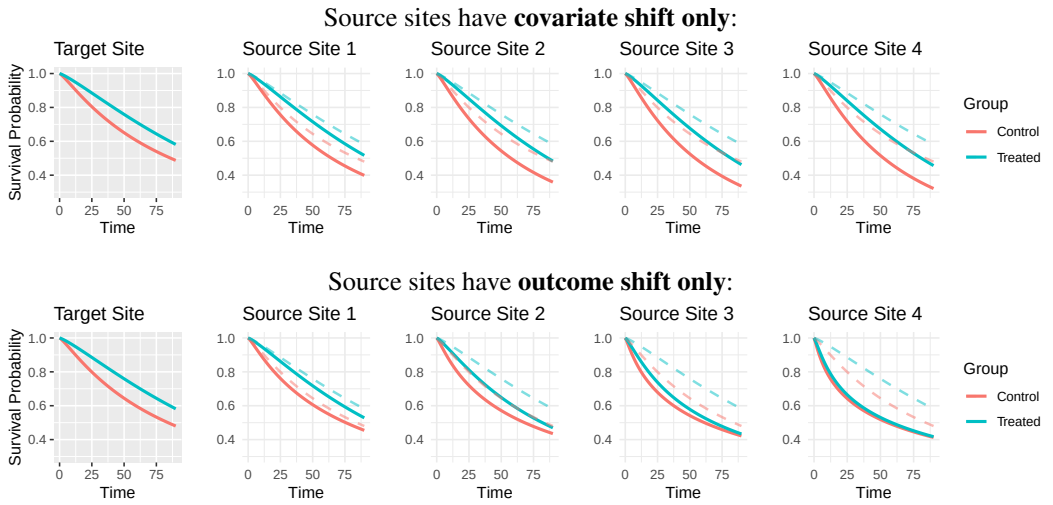


Figure 5: Site- and treatment-specific survival curves, each based on a random sample of $n = 10^4$ from the true DGP of each site. The two dashed curves in each source site panel are the target site survival functions for reference. Under covariate shift, curves preserve their shapes and trends but differ in scale, whereas outcome shift produces marked changes in shape and treatment effects.

B.2 Performance criteria definitions

The simulation performance criteria considered in Section 3.2 with an additional metric 95% confidence interval (CI) width for the complete simulation results are defined as follows.

Let θ denote the true target parameter, and let $\widehat{\theta}_i$ and $\widehat{\sigma}_i$ be the point and standard error estimates, respectively, from the ith Monte Carlo replication of a competing method, $i = 1, \ldots, 500$. Then:

- **Estimation bias:** $\widehat{\theta}_i - \theta$, $i = 1, \ldots, 500$, summarized via boxplots;
- **RRMSE:** the RMSE of a method relative to that of the TGT estimator, where RMSE $= \sqrt{500^{-1} \sum_{i=1}^{500} (\widehat{\theta}_i - \theta)^2}$. By definition, the TGT estimator has RRMSE $= 1$. Smaller RRMSE values indicate higher efficiency relative to TGT;
- **CP%:** the proportion of replications in which the Wald-type CI contains θ: $100\% \times 500^{-1} \sum_{i=1}^{500} \mathbb{I}\{\theta \in [\widehat{\theta}_i - 1.96\widehat{\sigma}_i,\ \widehat{\theta}_i + 1.96\widehat{\sigma}_i]\}$. The closer CP% is to 95, the more reliable the inference based on $\widehat{\sigma}_i$; and
- **95% CI width:** the average CI width across replications, where the CI from the ith replication is $\widehat{\theta}_i \pm 1.96, \widehat{\sigma}_i$. Thus, CI width $= 3.92 \times 500^{-1} \sum_{i=1}^{500} \widehat{\sigma}_i$.

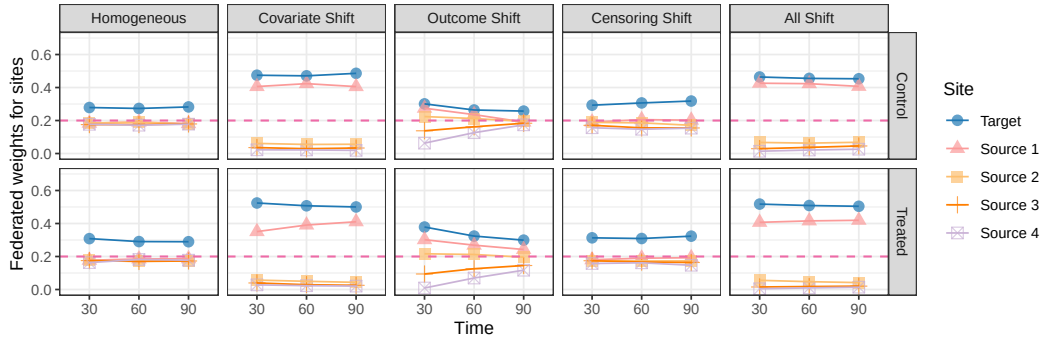


Figure 6: Average federated weights of each site at different time point by site heterogeneity cases. This figure uses the case where $n_k = 300$ ($k \geq 1$) as an illustration for weights.

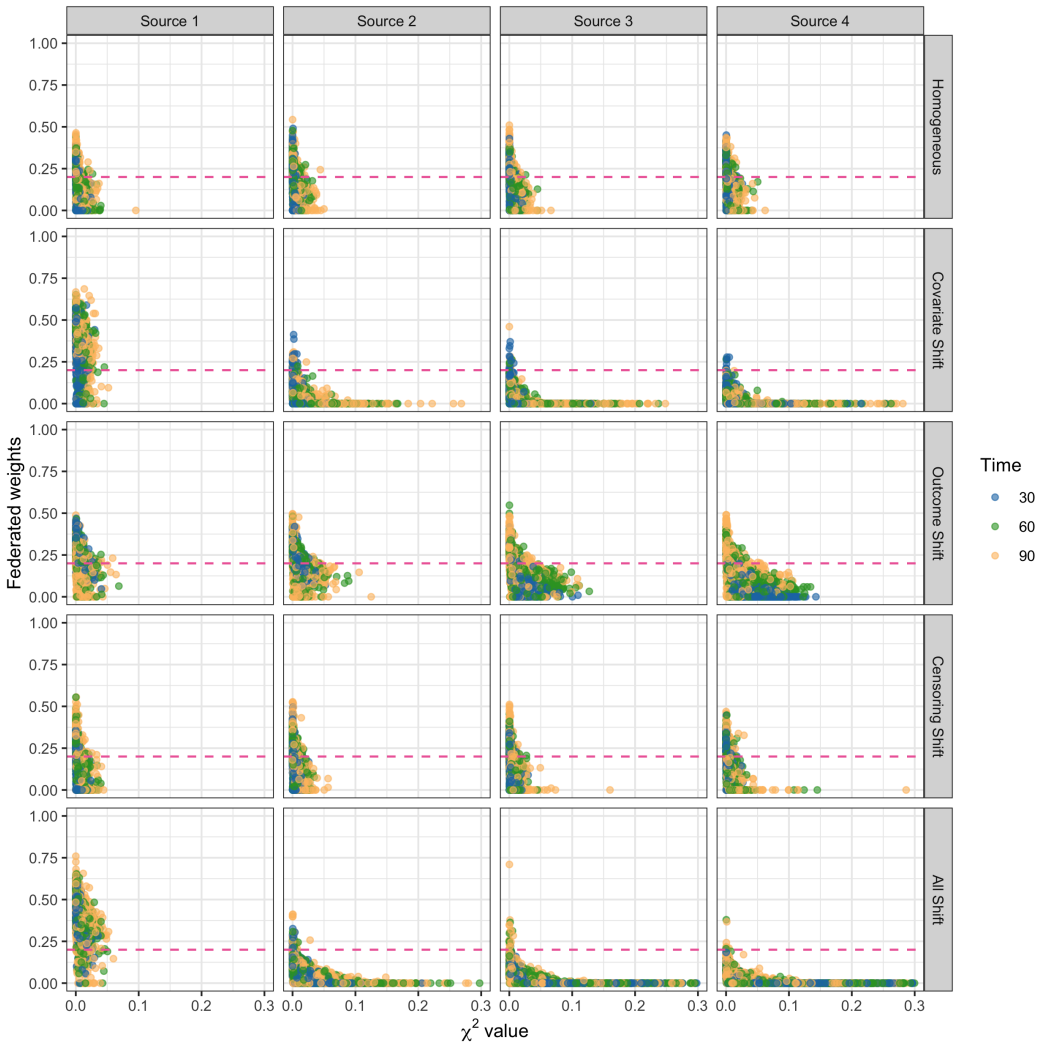


Figure 7: Scatter plots of site-specific federated weights vs. discrepancy measure $(\widehat{\chi}^k_{n,t,a})^2$ values, under 5 scenarios of site heterogeneity and 3 selected time points (days 30, 60 and 90). Sites 2–4 under Covariate Shift and All Shift have more larger $(\widehat{\chi}^k_{n,t,a})^2$ values with clear trends of decreasing weights. The pink dashed lines indicate weight = 1/5, i.e., one over five sites. This figure uses the case where $n_k = 300$ ($k \geq 1$) as an illustration for weights.

B.3 COMPLETE SIMULATION RESULTS

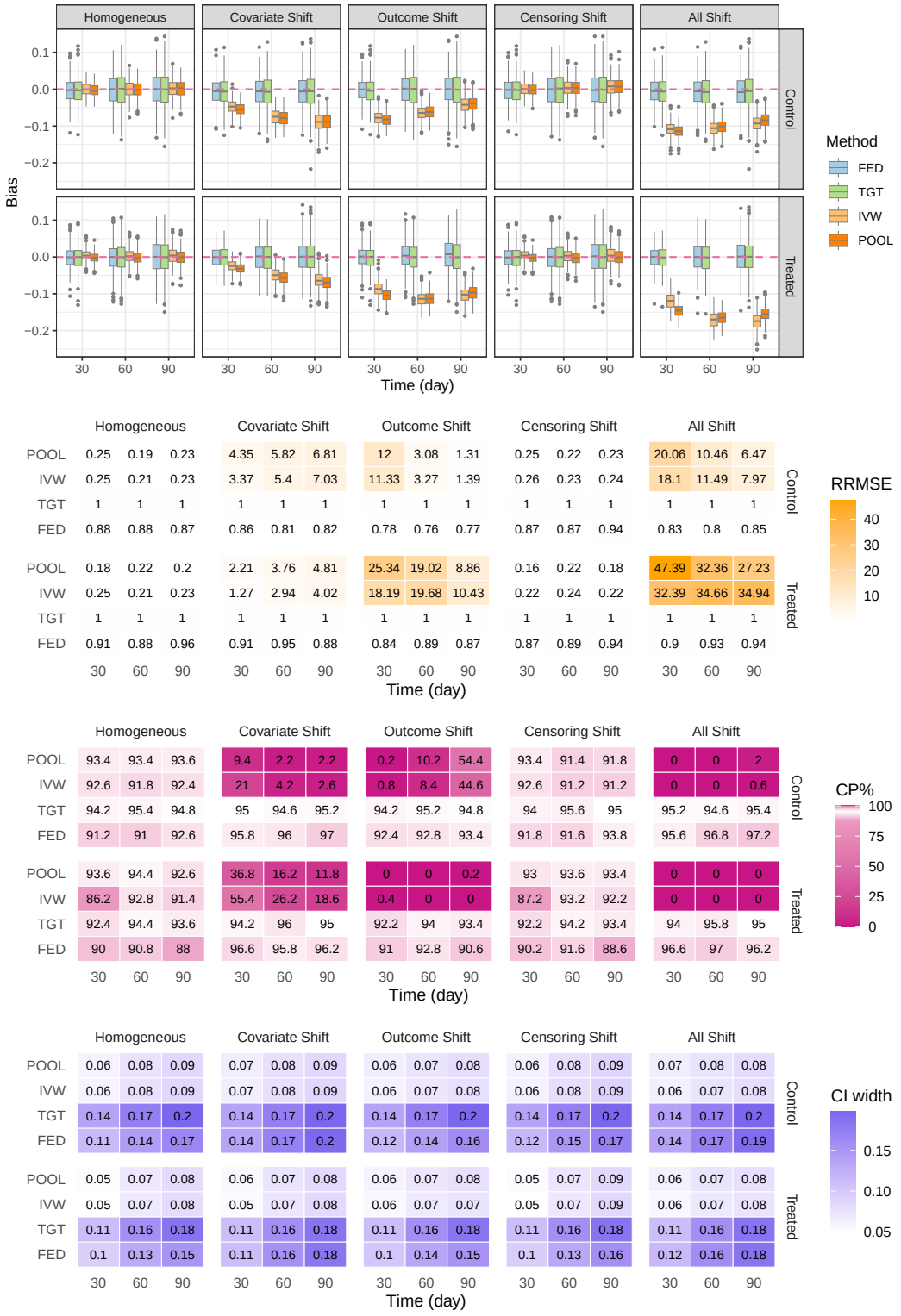


Figure 8: Estimation bias (boxplots), relative root mean square error (RRMSE) compared to TGT, coverage probability (CP%) with 95% nominal coverage level, and width of 95% CI under $n_k = 300$ ($k \geq 1$), evaluated at days 30, 60 and 90 in simulation.

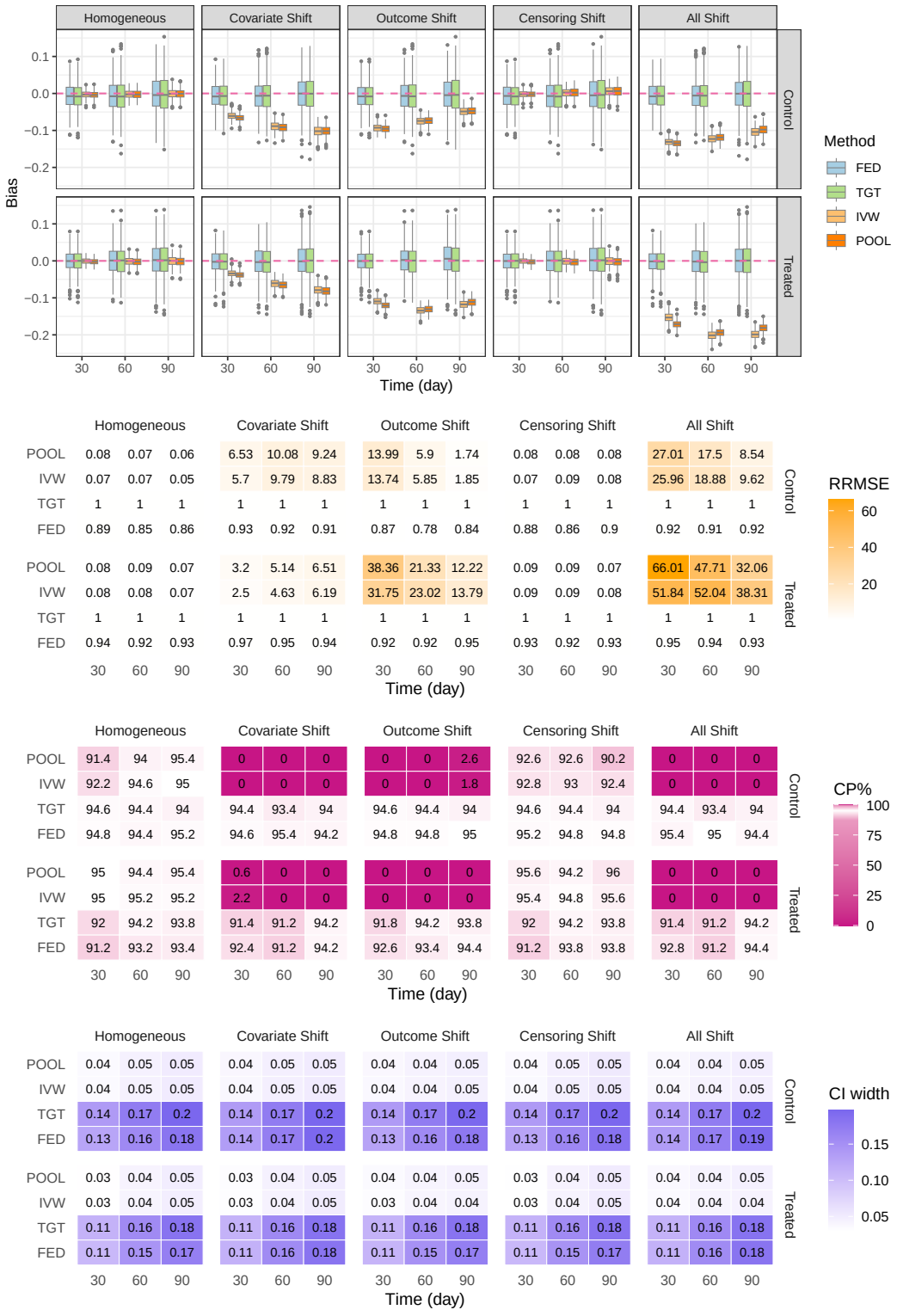


Figure 9: Estimation bias (boxplots), relative root mean square error (RRMSE) compared to TGT, coverage probability (CP%) with 95% nominal coverage level, and width of 95% CI under $n_k = 600$ ($k \geq 1$), evaluated at days 30, 60 and 90 in simulation.

Figure 10: Estimation bias (boxplots), relative root mean square error (RRMSE) compared to TGT, coverage probability (CP%) with 95% nominal coverage level, and width of 95% CI under $n_k = 1000$ ($k \geq 1$), evaluated at days 30, 60 and 90 in simulation.

C Additional Results for Real Data Analysis

C.1 AMP trial data

Table 1 presents summary statistics for baseline covariates and outcomes in the AMP trial data, stratified by region and treatment group. Comparing the treatment groups—both overall and within each region—we observe that the treated group consistently shows a lower average event proportion. Additionally, some covariates appear to shift across regions; for example, among treated participants, the standardized risk scores exhibit notably different means when comparing SA to BP and US.

	Treated (bnAb) group				
	Total ($n = 3,076$)	**SA** ($n = 679$)	**OA** ($n = 608$)	**BP** ($n = 846$)	**US** ($n = 943$)
Age (year) at baseline	25.9 (4.60)	27.0 (5.19)	25.4 (4.59)	25.1 (3.70)	26.2 (4.68)
Standardized risk score	0.0 (1.00)	-0.01 (1.00)	0.02 (1.00)	0.76 (0.67)	-0.68 (0.71)
Weight at baseline (kg)	72.8 (15.64)	68.8 (14.24)	65.2 (12.63)	70.9 (12.42)	82.3 (16.43)
HIV diagnosis by week-80	107 (3.48%)	27 (3.98%)	20 (3.29%)	46 (5.44%)	14 (1.49%)
	Control (placebo) group				
	Total ($n = 1,535$)	**SA** ($n = 340$)	**OA** ($n = 297$)	**BP** ($n = 428$)	**US** ($n = 470$)
Age (year) at baseline	25.9 (4.72)	26.6 (5.28)	25.4 (4.78)	25.2 (3.94)	26.1 (3.79)
Standardized risk score	0.0 (1.00)	0.02 (0.92)	-0.02 (0.98)	0.75 (0.67)	-0.68 (0.73)
Weight at baseline (kg)	72.5 (16.35)	67.6 (14.77)	65.1 (13.64)	71.1 (12.84)	81.8 (17.5)
HIV diagnosis by week-80	67 (4.36%)	16 (4.71%)	13 (4.38%)	29 (6.78%)	9 (1.91%)

Table 1: Summary statistics of AMP trial data by treatment group and region. The standardized risk score is a baseline score built by machine learning models (Corey et al., 2021) that is predictive to the time-to-event outcome. Age, standardized risk score and weight are summarized by mean (standard deviation), while the HIV diagnosis by week-80 is summarized by count (percentage).

In Figure 11, we plot the region-specific survival curves of all the 4 regions we considered (SA, OA, BP and US) for a direct comparison on region heterogeneity, using their target-site-only (TGT) estimators, to showcase the heterogeneous effects of the bnAb antibody treatment on different target populations.

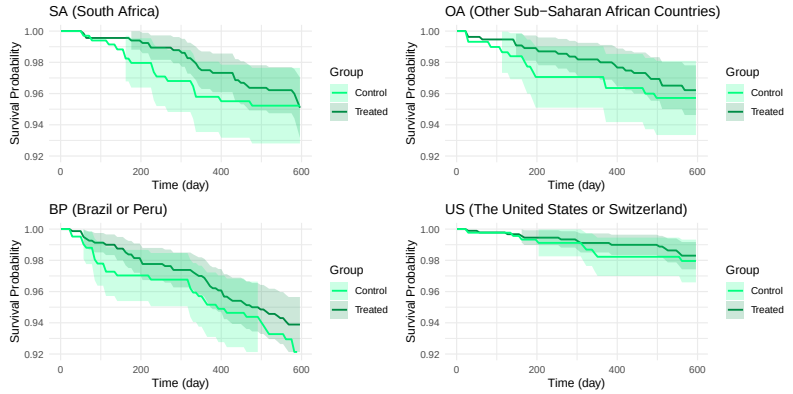


Figure 11: Estimated region-specific survival curves of the HVTN 704/HPTN 085 and HVTN 703/HPTN 081 trials. SA (our target region in the main text) and OA exhibit relatively similar curves, indicating less heterogeneity of these two regions. In contrast, both BP and US regions show significant differences to SA, which also confirms why they often have small or zero federated weights in Panel (B) of Figure 4 in the main text. The BP and US also show a substantial difference on their curves.

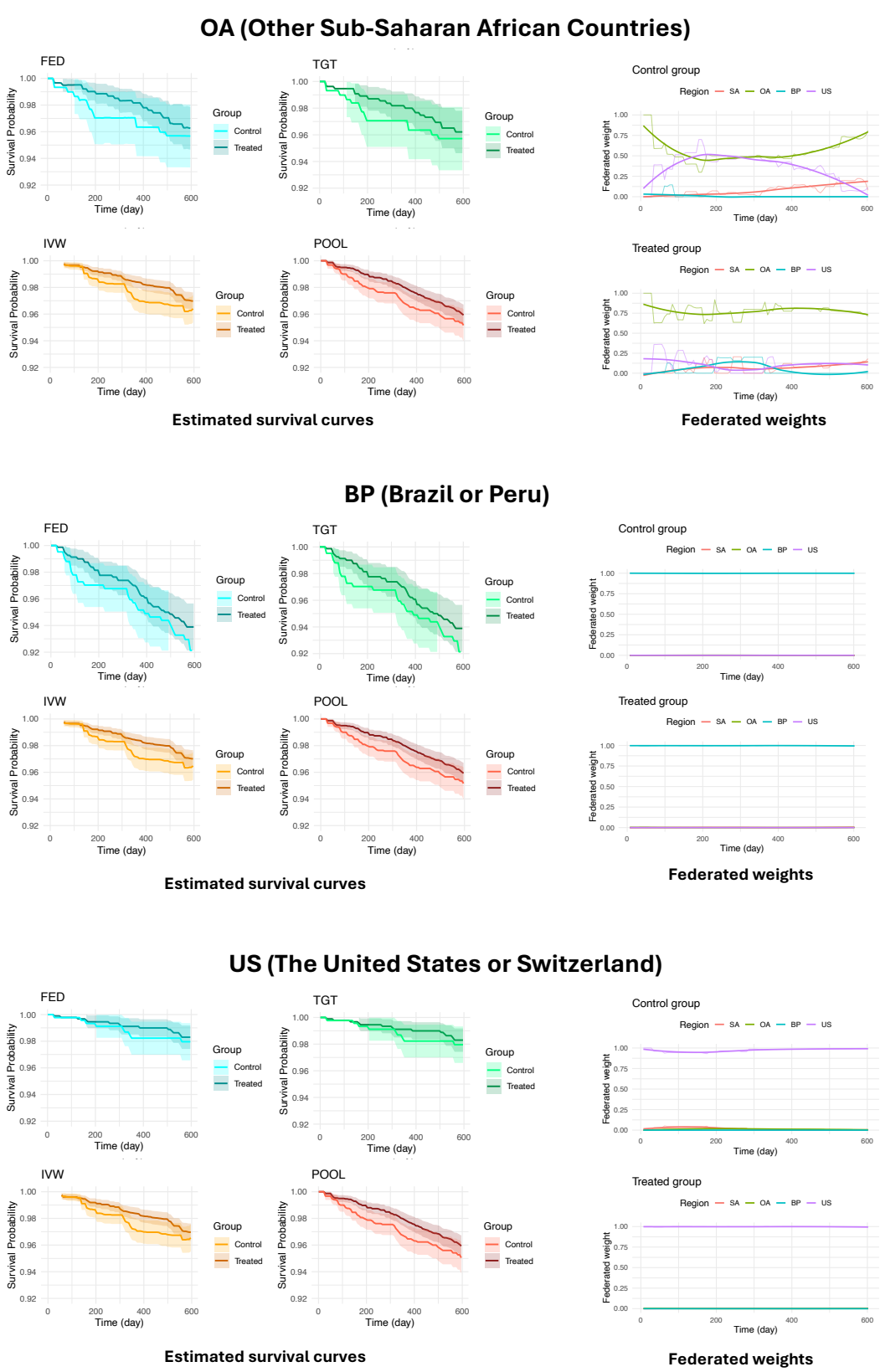


Figure 12: Additional data analysis results when treating the other three regions (OA, BP and US) as the target site.

Furthermore, in Figure 12, we present the results—including survival curve estimations and federated weights—using three regions other than SA as the target population. For the federated weights, similar to Figure 4 in the main text, we applied locally weighted regression (Cleveland & Devlin, 1988) to smooth the observed weights over the study period, providing a clearer visualization of temporal trends in this specific example.

From Figure 12, we observe that for each region, the FED method yields results similar to the TGT estimator, while also recovering some interval estimations at earlier time points. This finding is consistent with the observations made in Figure 4. In contrast, the IVW and POOL methods deviate noticeably from the TGT and FED results—especially for the BP and US regions—indicating potential biases introduced by site heterogeneity.

Finally, regarding federated weights, the results for the OA region resemble those of SA in Figure 4. However, for the BP and US regions, the federated weights are nearly 1 for the target site and 0 for all other sites. This pattern suggests that when targeting the survival curves of BP or US, other sites contribute substantial biases—an observation that corroborates our findings in Figure 11.

C.2 "FLCHAIN" DATASET FROM R PACKAGE `SURVIVAL`

The "flchain" dataset, obtained from the Mayo Clinic Study of Serum Free Light Chain (FLC) and Mortality, comprises data on 7,874 individuals followed between 1995 and 2009 to investigate the prognostic value of serum free light chains for survival (Dispenzieri et al., 2012; Kyle et al., 2006). This dataset is freely available in R package `survival`.

This dataset does not contain a natural treatment variable, but to illustrate and extend the use of our framework, we investigate the sex difference in mortality. Since sex (female vs. male) is assigned at birth, it can be viewed as a "treatment" variable for methodological purposes, as it precedes the occurrence of any outcomes. While not manipulable in the conventional sense, causal inference methods allow us to frame sex as an exposure to quantify disparities in survival outcomes, rather than as an intervention subject to policy or clinical decision-making. Similar approaches have been employed to assess disparities associated with non-manipulable variables such as race (Li & Li, 2023; Liu et al., 2025).

			Male	
	Total ($n = 3,524$)	**Group A** ($n = 972$)	**Group B** ($n = 1,429$)	**Group C** ($n = 1,123$)
Age (year) at baseline	63.1 (9.62)	60.1 (7.80)	62.6 (9.25)	66.4 (10.5)
MGUS	0.01 (0.11)	0.04 (0.20)	0.00 (0.05)	0.00 (0.00)
Sample year	1996.9 (1.84)	1996.7 (1.72)	1996.9 (1.87)	1996.9 (1.90)
Concentration of κ light chain	1.5 (1.01)	0.9 (0.34)	1.4 (0.45)	2.2 (1.44)
Concentration of λ light chain	1.8 (1.19)	1.1 (0.35)	1.6 (0.47)	2.5 (1.77)
Mortality	1,004 (28.5%)	159 (16.4%)	372 (26.0%)	473 (42.1%)
			Female	
	Total ($n = 4,350$)	**Group A** ($n = 1,399$)	**Group B** ($n = 1,771$)	**Group C** ($n = 1,180$)
Age (year) at baseline	65.2 (11.01)	62.2 (9.57)	65.0 (10.8)	69.1 (11.8)
MGUS	0.02 (0.12)	0.05 (0.21)	0.00 (0.05)	0.0 (0.00)
Sample year	1996.7 (1.70)	1996.6 (1.55)	1996.7 (1.68)	1996.9 (1.87)
Concentration of κ light chain	1.4 (0.78)	0.9 (0.34)	1.3 (0.43)	2.1 (1.03)
Concentration of λ light chain	1.6 (0.88)	1.1 (0.35)	1.6 (0.46)	2.4 (1.22)
Mortality	1,165 (26.8%)	231 (16.5%)	455 (25.7%)	479 (40.6%)

Table 2: Summary statistics of "flchain" data by sex group and the site variable we defined. All baseline covariates are summarized by mean (standard deviation), while the mortality is summarized by count (percentage).

We include age, the presence of monoclonal gammopathy of undetermined significance (MGUS) and sample year as baseline covariates for nuisance models. The primary outcome consists of follow-up time in days and an event indicator for all-causes death (mortality).

A categorical variable (`flc.grp`, taking values $1, 2, \ldots, 10$) related to κ and λ concentration levels is available in the data. We construct the "site" variable (R in our notation) based on `flc.grp` as follows: (i) Group A for `flc.grp` $\in \{1, 2, 5\}$; (ii) Group B for `flc.grp` $\in \{3, 4, 6, 9\}$; and (iii) Group C for `flc.grp` $\in \{7, 8, 10\}$. Several categories were merged in this way to ensure a sufficient sample size within each group, allowing 5-fold cross-fitting to train different nuisance functions reliably. In addition, we allow the groups to share nearby values of `flc.grp` (e.g., 5 in Group A, 6 in Group B, and 7 in Group C) so that each site retains comparable information, enabling borrowing across groups. We emphasize that this grouping method is adopted solely for illustrative purposes in demonstrating our framework.

Table 2 presents the summary statistics of baseline covariates and mortality for the "flchain" data. Across Groups A, B, and C, we observe clear covariate shifts, accompanied by differences in the marginal death rates. In contrast, when comparing the two "treatment" (sex) groups, the distributions of baseline covariates and mortality appear overall similar.

We analyzed the sex-specific survival curves over the first 10 years for the three groups in Figure 13. We used a 5-fold cross-fitting, and estimated conditional survival for both event and censoring processes by an ensemble of Kaplan–Meier, Cox regression and survival random forest models via the `survSuperLearner` package (Westling et al., 2024). The propensity score and density ratio (used in federated method) models were fitted by the ensemble of logistic regression and LASSO using the Super Learner (van der Laan et al., 2007).

Overall, the FED method yields point estimates that closely track those of the TGT estimator, while producing slightly narrower confidence bands. By calculations, the efficiency gain (by estimated standard error of FED to that of TGT) can achieve 3%–10%, consistent with the findings from both our simulation studies and the AMP trial data. By contrast, the IVW and POOL estimators exhibit noticeably different survival curve patterns relative to TGT and FED when Groups A and C are regarded as targets, suggesting potential biases.

D Implementation Details

In the following Algorithm 2, we detail the double machine learning procedure for fitting and predicting nuisance functions in Algorithm 1.

Remark D.1. To ensure the monotonicity of the estimated survival curves, we invoke isotonic regression techniques (Westling et al., 2020), which enforce a non-increasing constraint on the site-specific survival and censoring estimates $\widehat{S}^k$ and $\widehat{G}^k$, for $k = 0, 1, \ldots, K-1$, thereby maintaining their logical consistency over time.

E Technical Proofs

We adopt the following notation throughout this appendix: (i) $\mathbb{P}_\infty$ denotes a general probability limit, and the nuisance functions under $\mathbb{P}_\infty$ are denoted with subscript ∞, e.g., S^0_∞ for the limit of $\widehat{S}^0$; (ii) $\widehat{\mathbb{P}}$ means the corresponding nuisance functions are replaced by their estimates, and $\widehat{\mathbb{P}}$ may converge to a general limit $\mathbb{P}_\infty$; (iii) $\mathbb{P}_n^m[f(\mathcal{O})] = |\mathcal{V}_m|^{-1} \sum_{i \in \mathcal{I}_m} f(\mathcal{O}_i)$ to denote the empirical average on the m-th validation set $\mathcal{V}_m$ by cross-fitting, $m = 1, \ldots, M$.

Furthermore, we distinguish notation $\mathbb{P}(f)$ and $\mathbb{E}_{\mathbb{P}}(f)$: $\mathbb{P}(f) = \int f(\mathcal{O}) d\mathbb{P}$ denotes an integral over a new observation $\mathcal{O} \sim \mathbb{P}$, treating f, which possibly depends on training data (e.g., some estimated parameters for nuisance functions), as fixed. In contrast, $\mathbb{E}_{\mathbb{P}}(f)$ is the usual mathematical expectation of random variable/element f under distribution $\mathbb{P}$, a fixed value without randomness.

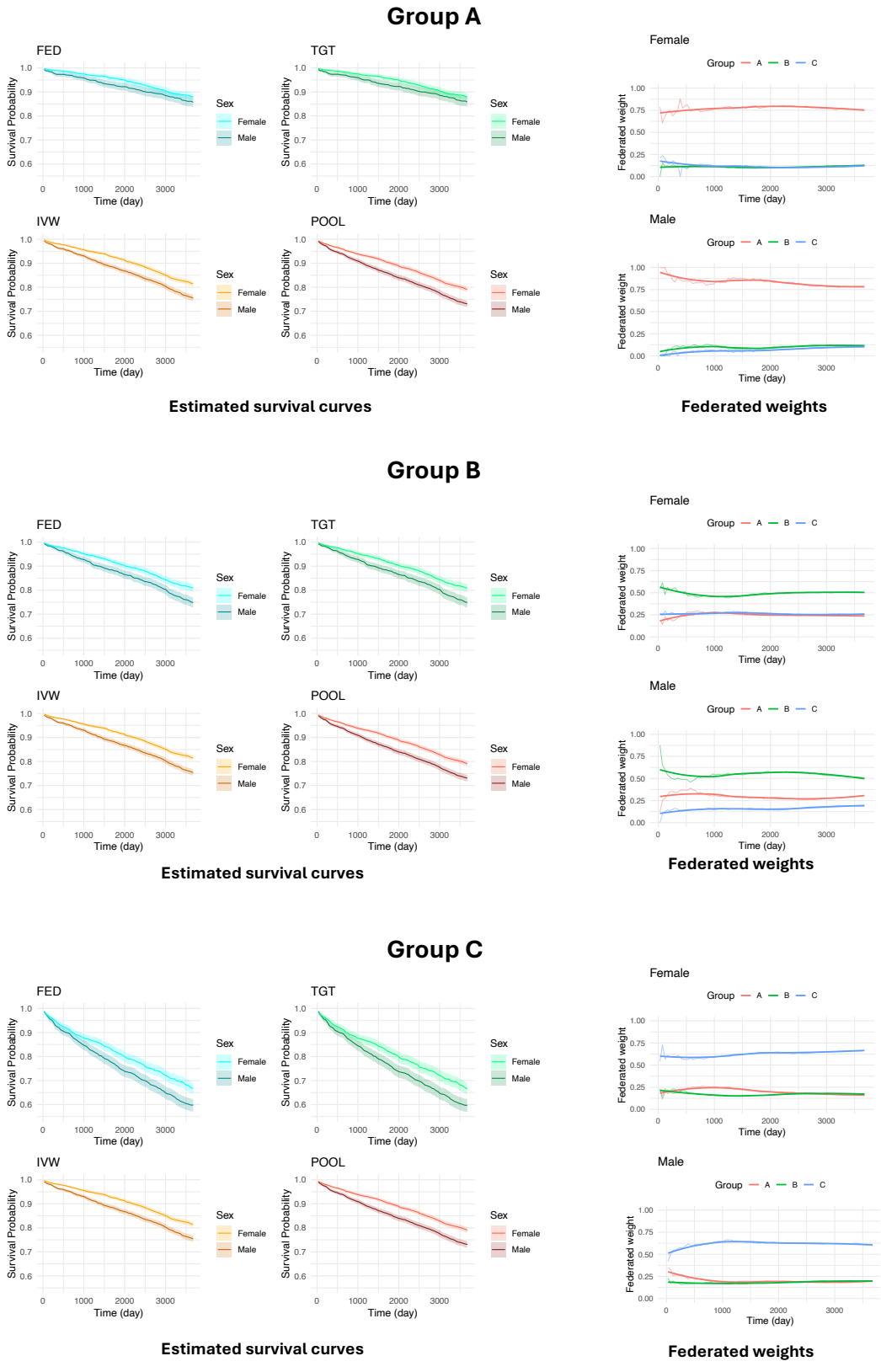


Figure 13: "flchain" data analysis results. Estimated sex-specific survival curves and federated weights for sites (Groups A, B and C defined by `flc.grp` variable).

Algorithm 2 Double/debiased machine learning algorithm for nuisance function estimations and influence function calculations in Algorithm 1 at a given time point and treatment.

Input: Observed multi-source right-censored data $\mathcal{O} = \{\mathcal{O}_i = (\mathbf{X}_i, A_i, Y_i, \Delta_i, R_i), i = 1, \ldots, n\} = \mathcal{O}^0 \cup \mathcal{O}^1 \cup \cdots \cup \mathcal{O}^{K-1}$, where $R_i \in \{0, 1, \ldots, K-1\}$ and $\mathcal{O}^k$ represents the data for site $R = k$; Given treatment group $A = a$ and a specific time point t; The number of disjoint folds into which the data are split, M, where $M \in \{2, 3, \ldots, \lfloor n^*/2 \rfloor\}$ with $n^* = \min\{n, n_1, \ldots, n_{K-1}\}$.
Output: Estimated influence functions for each individual.
Partition $\mathcal{O}^0$ into M approximately equal-sized, disjoint validation folds $\mathcal{V}_1^0, \ldots, \mathcal{V}_M^0$, allowing a size difference of at most ± 1 between folds.
for $m = 1, \ldots, M$ **do**
    Define the training set $\mathcal{T}_m^0 = \mathcal{O}^0 \backslash \mathcal{V}_m^0$;
    Fit nuisance functions S^0, G^0, π^0 on $\mathcal{T}_m^0$, using some methods ensemble from `survSuperLearner` and `SuperLearner`;
    Predict nuisance functions on $\mathcal{V}_m^0$ as $\widehat{S}_m^0, \widehat{G}_m^0$ and $\widehat{\pi}_m^0$.
end for
Train a model of S^0 by the entire data of the target site $\mathcal{O}^0$, denoted as $S^{0,\text{full}}$, using chosen methods ensemble from `survSuperLearner`.
for $k = 1, \ldots, K-1$ **do**
    Partition $\mathcal{O}^k$ into M approximately equal-sized, disjoint validation folds $\mathcal{V}_1^k, \ldots, \mathcal{V}_M^k$, allowing a size difference of at most ± 1 between folds.
    for $m = 1, \ldots, M$ **do**
        Define the training set $\mathcal{T}_m^k = \mathcal{O}^k \backslash \mathcal{V}_m^k$;
        Fit the density ratio $\omega^{k,0}$ using only covariate data of $\mathcal{T}_m^0 \cup \mathcal{T}_m^k$, or by just passing through some coarsening level summary statistics;
        Fit nuisance functions G^k, π^k on $\mathcal{T}_m^k$, using chosen methods ensembles from `survSuperLearner` and `SuperLearner`;
        Predict above nuisance functions on $\mathcal{V}_m^k$ as $\widehat{G}_m^k$, $\widehat{\omega}_m^{k,0}$ and $\widehat{\pi}_m^k$;
        Predict nuisance function S^k on $\mathcal{V}_m^k$ using the pre-trained $S^{0,\text{full}}$ model, and denote the predicted value by $\widehat{S}_m^k$.
    end for
    Aggregate all predicted nuisance functions over M folds as $\widehat{S}^k, \widehat{G}^k, \widehat{\omega}^{k,0}$ and $\widehat{\pi}^k$;
end for
Return: The estimated EIFs, by plugging-in their predicted nuisance function values, $\widehat{\varphi}_{t,a}^{*k,0}(\mathcal{O}; \widehat{\mathbb{P}}) = \widehat{\varphi}_{t,a}^{*k,0}(\mathcal{O}; \widehat{S}^k, \widehat{S}^0, \widehat{G}^k, \widehat{\pi}^k, \widehat{\omega}^{k,0})$, for all $k \in \{0, 1, \ldots, K-1\}$.

E.1 THEORY OF THE LOCAL ESTIMATOR

E.1.1 PROOF OF THEOREM 2.5

Recall that a mean zero, finite variance function $\varphi^{*0}_{t,a}(\mathcal{O};\mathbb{P})$ is called an *influence function* of the target estimand (a functional) $\theta^0(t,a) = \theta^0(t,a;\mathbb{P})$ at $\mathbb{P}$ if, for any one-dimensional regular parametric submodel $\{\mathbb{P}_\epsilon : \epsilon \in [0,1)\}$ through $\mathbb{P} \equiv \mathbb{P}_0$,

$$\frac{\partial}{\partial \epsilon}\theta^0(t,a;\mathbb{P}_\epsilon)\bigg|_{\epsilon=0} = \mathbb{E}_{\mathbb{P}}[\varphi^{*0}_{t,a}(\mathcal{O};\mathbb{P})\dot{\ell}(\mathcal{O})],$$

where $\dot{\ell}(\mathcal{O})$ is the score function of the submodel at $\epsilon = 0$ (i.e., typically, $\dot{\ell}(\mathcal{O}) = \partial \log\{p_\epsilon(\mathcal{O})\}/\partial\epsilon\,|_{\epsilon=0}$), where $p_\epsilon(\cdot)$ denotes the probability density (likelihood) function under submodel $\mathbb{P}_\epsilon$ (Bickel et al., 1993).

Recall the partial CCOD assumption made in Theorem 2.5, $S^0(t \mid a, \mathbf{X}) = S^0(t \mid a, \mathbf{X})$ almost surely. To find the EIF, we begin by writing the following equation:

$$\begin{aligned}
0 &= \frac{\partial}{\partial\epsilon}\theta^0(t,a)\bigg|_{\epsilon=0} = \frac{\partial}{\partial\epsilon}\mathbb{E}\{S^0_\epsilon(t \mid a,\mathbf{X}) \mid R=0\}\bigg|_{\epsilon=0} \\
&= \mathbb{E}\{[S^0(t \mid a,\mathbf{X}) - \theta^0(t,a)]\dot{\ell}_{\mathbf{X}|R=0} \mid R=0\} + \mathbb{E}\left\{\int \frac{\partial}{\partial\epsilon}S^0_\epsilon(t \mid a,\mathbf{x})\bigg|_{\epsilon=0}\mu(d\mathbf{x}) \;\middle|\; R=0\right\} \\
&= \mathbb{E}\{[S^0(t \mid a,\mathbf{X}) - \theta^0(t,a)]\dot{\ell}_{\mathbf{X}|R=0} \mid R=0\} + \mathbb{E}\left\{\int \frac{\partial}{\partial\epsilon}S^k_\epsilon(t \mid a,\mathbf{x})\bigg|_{\epsilon=0}\mu(d\mathbf{x}) \;\middle|\; R=0\right\},
\end{aligned} \tag{3}$$

where $\mu(\cdot)$ denotes the distribution of $\mathbf{X}$ induced by $\mathbb{P}$ and, for any sets of variables V and W, $\dot{\ell}_{V|W}$ denotes the conditional score function of V given W, i.e., typically $\partial \log\{p_\epsilon(V \mid W)\}/\partial\epsilon\,|_{\epsilon=0}$—note that such scores always satisfy $\mathbb{E}_{\mathbb{P}}(\dot{\ell}_{V|W} \mid W) = 0$ (Bickel et al., 1993).

For the derivative of S^k_ϵ with respect to ϵ, by the chain rule, we decompose it as $(\partial S^k_\epsilon/\partial\Lambda^k_\epsilon) \times (\partial\Lambda^k_\epsilon/\partial\epsilon)$. For the first part $\partial S^k_\epsilon/\partial\Lambda^k_\epsilon$, we leverage Theorem 8 in Gill & Johansen (1990). Specifically, the mapping $H \mapsto S^k(t;H) := \prod_{(0,t]}\{1 + H(du)\}$ is Hadamard differentiable at H relative to the supremum norm with derivative

$$\alpha \mapsto S^k(t;H)\int_0^t \frac{S^k(u-;H)}{S^k(u;H)}\alpha(du).$$

Thus, by letting $H(t) = \Lambda^k_\epsilon(t \mid a,\mathbf{x})$ and the chain rule, the integrand in the second term becomes

$$\frac{\partial}{\partial\epsilon}\prod_{(0,t]}\{1-\Lambda^k_\epsilon(du \mid a,\mathbf{x})\}\bigg|_{\epsilon=0} = -S^k(t \mid a,\mathbf{x})\int_0^t \frac{S^k(u- \mid a,\mathbf{x})}{S^k(u \mid a,\mathbf{x})}\frac{\partial}{\partial\epsilon}\Lambda^k_\epsilon(du \mid a,\mathbf{x})\bigg|_{\epsilon=0}.$$

Furthermore, recall that

$$\Lambda^k(t \mid a,\mathbf{X}) = \int_0^t \frac{N^k_1(du \mid a,\mathbf{X})}{D^k(u \mid a,\mathbf{X})},$$

where $N^k_1(t \mid a,\mathbf{X}) = \mathbb{P}(Y \le t, \Delta = 1 \mid A = a, \mathbf{X}, R = k)$ and $D^k(t \mid a,\mathbf{X}) = \mathbb{P}(Y \ge t \mid A = a,\mathbf{X}, R = k)$. Hence,

$$\frac{\partial}{\partial\epsilon}\Lambda^k_\epsilon(du \mid a,\mathbf{x})\bigg|_{\epsilon=0} = \frac{\frac{\partial}{\partial\epsilon}N^k_{1,\epsilon}(du \mid a,\mathbf{x})\,|_{\epsilon=0}}{D^k(u \mid a,\mathbf{x})} - \frac{\frac{\partial}{\partial\epsilon}D^k_\epsilon(u \mid a,\mathbf{x})\,|_{\epsilon=0}\,N^k_{1,\epsilon}(du \mid a,\mathbf{x})}{D^k(u \mid a,\mathbf{x})^2}.$$

In addition,

$$\begin{aligned}\frac{\partial}{\partial\epsilon}N^k_{1,\epsilon}(du \mid a,\mathbf{x})\bigg|_{\epsilon=0} &= \frac{\partial}{\partial\epsilon}\mathbb{P}_\epsilon(Y\le u,\Delta=1 \mid A=a,\mathbf{X}=\mathbf{x},R=k)\bigg|_{\epsilon=0}\\ &= \frac{\partial}{\partial\epsilon}\iint \mathbb{I}(y\le u,\delta=1)\mathbb{P}_\epsilon(dy,d\delta \mid a,\mathbf{x},k)\bigg|_{\epsilon=0}\\ &= \iint \mathbb{I}(y\le u,\delta=1)\dot{\ell}(y,\delta \mid a,\mathbf{x})\mathbb{P}(dy,d\delta \mid a,\mathbf{x},k)\\ &= \int_\delta \mathbb{I}(\delta=1)\dot{\ell}(u,\delta \mid a,\mathbf{x})\mathbb{P}(du,d\delta \mid a,\mathbf{x},k),\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial\epsilon}D^k_\epsilon(u \mid a,\mathbf{x})\bigg|_{\epsilon=0} &= \frac{\partial}{\partial\epsilon}\mathbb{P}_\epsilon(Y\ge u \mid A=a,\mathbf{X}=\mathbf{x},R=k)\bigg|_{\epsilon=0}\\ &= \frac{\partial}{\partial\epsilon}\iint \mathbb{I}(y\ge u)\mathbb{P}_\epsilon(dy,d\delta \mid a,\mathbf{x},k)\bigg|_{\epsilon=0}\\ &= \iint \mathbb{I}(y\le u)\dot{\ell}(y,\delta \mid a,\mathbf{x})\mathbb{P}(dy,d\delta \mid a,\mathbf{x},k).\end{aligned}$$

We can then express the integrand of (3) as

$$\begin{aligned}&\frac{\partial}{\partial\epsilon}\iint\prod_{(0,t]}\{1-\Lambda^k_\epsilon(du \mid a,\mathbf{x})\}\mu(d\mathbf{x})\bigg|_{\epsilon=0}\\
&= \iiint -\mathbb{I}(y\le t,\delta=1)\frac{S^k(t \mid a,\mathbf{x})S^k(y- \mid a,\mathbf{x})}{S^k(y \mid a,\mathbf{x})D^k(y \mid \mathbf{x})}\dot{\ell}(y,\delta \mid a,\mathbf{x},k)\mathbb{P}(dy,d\delta \mid a,\mathbf{x},k)\mu(d\mathbf{x})\\
&\quad+\iiiint \mathbb{I}(u\le t,u\le y)\frac{S^k(t \mid a,\mathbf{x})S^k(u- \mid a,\mathbf{x})}{S^k(u \mid a,\mathbf{x})D^k(u \mid \mathbf{x})}\\
&\qquad\times\dot{\ell}(y,\delta \mid a,\mathbf{x},k)\mathbb{P}(dy,d\delta \mid a,\mathbf{x},k)N^k_1(du \mid a,\mathbf{x})\mu(d\mathbf{x})\\
&= \iiint -\mathbb{I}(y\le t,\delta=1)\frac{S^k(t \mid a,\mathbf{x})S^k(y- \mid a,\mathbf{x})}{S^k(y \mid a,\mathbf{x})D^k(y \mid \mathbf{x})}\dot{\ell}(y,\delta \mid a,\mathbf{x},k)\mathbb{P}(dy,d\delta \mid a,\mathbf{x},k)\mu(d\mathbf{x})\\
&\quad+\iiint S^k(t \mid a,\mathbf{x})\int_0^{t\wedge y}\frac{S^k(u- \mid a,\mathbf{x})}{S^k(u \mid a,\mathbf{x})D^k(u \mid \mathbf{x})^2}N^k_1(du \mid a,\mathbf{x})\\
&\qquad\times\dot{\ell}(y,\delta \mid a,\mathbf{x},k)\mathbb{P}(dy,d\delta \mid a,\mathbf{x},k)\mu(d\mathbf{x})\\
&= \mathbb{E}\bigg[S^k(t \mid a,\mathbf{X})\frac{\mathbb{I}(A=a)}{\pi^k(a \mid \mathbf{X})}\left\{H^k(t\wedge Y,a,\mathbf{X})-\frac{\mathbb{I}(Y\le t,\Delta=1)S^k(Y- \mid a,\mathbf{X})}{S^k(Y \mid a,\mathbf{X})D^k(Y \mid a,\mathbf{X})}\right\}\\
&\qquad\times\dot{\ell}(Y,\Delta \mid a,\mathbf{X},R=k)\bigg],\end{aligned}$$

where

$$H^k(t,a,\mathbf{x})=\int_0^t\frac{S^k(u- \mid a,\mathbf{x})N^k_1(du \mid a,\mathbf{x})}{S^k(u \mid a,\mathbf{x})D^k(u \mid a,\mathbf{x})^2}.$$

Now, we note that

$$\mathbb{E}\left[\frac{\mathbb{I}(Y\le t,\Delta=1)S^k(Y- \mid A,\mathbf{X})}{S^k(Y \mid A,\mathbf{X})D^k(Y \mid A,\mathbf{X})}\,\middle|\,A=a,\mathbf{X}=\mathbf{x},R=k\right]=\int_0^t\frac{S^k(y- \mid a,\mathbf{x})N^k_1(dy \mid a,\mathbf{x})}{S^k(y \mid a,\mathbf{x})D^k(y \mid a,\mathbf{x})},$$

and

$$\begin{aligned}
&\mathbb{E}\{H^k(t\wedge Y, A, \mathbf{X}) \mid A=a, \mathbf{X}=\mathbf{x}, R=k\} \\
&= \iint^t \mathbb{I}(u \le y) \frac{S^k(u- \mid a,\mathbf{x}) N_1^k(du \mid a,\mathbf{x})}{S^k(u \mid a,\mathbf{x}) D^k(u \mid a,\mathbf{x})^2} \mathbb{P}(dy \mid a,\mathbf{x},k) \\
&= \int_0^t \mathbb{P}(Y \ge u \mid A=a, \mathbf{X}=\mathbf{x}, R=k) \frac{S^k(u- \mid a,\mathbf{x}) N_1^k(du \mid a,\mathbf{x})}{S^k(u \mid a,\mathbf{x}) D^k(u \mid a,\mathbf{x})^2} \mathbb{P}(dy \mid a,\mathbf{x},k) \\
&= \int_0^t \frac{S^k(u- \mid a,\mathbf{x}) N_1^k(du \mid a,\mathbf{x})}{S^k(u \mid a,\mathbf{x}) D^k(u \mid a,\mathbf{x})}.
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[H^k(t\wedge Y, A, \mathbf{X}) - \frac{\mathbb{I}(Y \le t, \Delta = 1) S^k(Y- \mid A, \mathbf{X})}{S^k(Y \mid A, \mathbf{X}) D^k(Y \mid A, \mathbf{X})} \;\middle|\; A, \mathbf{X}, R=k \right] = 0$$

almost surely. By properties of score functions and the tower property, the above implies that

$$\begin{aligned}
&\frac{\partial}{\partial \epsilon} \iint \prod_{(0,t]} \{1 - \Lambda_\epsilon^k(du \mid a,\mathbf{x})\} \mu(d\mathbf{x}) \Big|_{\epsilon=0} \\
&= \mathbb{E}\Bigg[S^k(t \mid a, \mathbf{X}) \frac{\mathbb{I}(R=k)}{\mathbb{P}(R=k \mid \mathbf{X})} \frac{\mathbb{I}(A=a)}{\pi^k(a \mid \mathbf{X})} \\
&\qquad \times \left\{ H^k(t\wedge Y, A, \mathbf{X}) - \frac{\mathbb{I}(Y \le t, \Delta=1) S^k(Y- \mid A, \mathbf{X})}{S^k(Y \mid A, \mathbf{X}) D^k(Y \mid A, \mathbf{X})} \right\} \dot{\ell}(\mathcal{O}) \Bigg].
\end{aligned}$$

Combining these results with the facts that $N_1^k(du \mid a,\mathbf{x})/D^k(u \mid a,\mathbf{x}) = \Lambda^k(du \mid a,\mathbf{x})$ and $D^k(u \mid a,\mathbf{x}) = S^k(u- \mid \mathbf{x}) G^k(u \mid a,\mathbf{x})$, we can rewrite (3) at the beginning as follows:

$$\begin{aligned}
&\frac{\partial}{\partial \epsilon} \theta^0(t,a) \Big|_{\epsilon=0} \\
&= \mathbb{E}\Bigg[\frac{\mathbb{I}(R=0)}{\mathbb{P}(R=0)} [S^k(t \mid a, \mathbf{X}) - \theta^0(t,a)] \dot{\ell}(\mathcal{O}) - \frac{\mathbb{I}(R=0)}{\mathbb{P}(R=0)} \mathbb{E}\Bigg\{ S^k(t \mid a, \mathbf{X}) \frac{\mathbb{I}(R=k)}{\mathbb{P}(R=k \mid \mathbf{X})} \\
&\qquad \times \frac{\mathbb{I}(A=a)}{\pi^k(a \mid \mathbf{X})} \left\{ \frac{\mathbb{I}(Y \le t, \Delta=1)}{S^k(y \mid \mathbf{X}) G^k(y \mid a, \mathbf{X})} - \int_0^{t\wedge y} \frac{\Lambda^k(du \mid a, \mathbf{X})}{S^k(u \mid \mathbf{X}) G^k(u \mid a, \mathbf{X})} \right\} \dot{\ell}(\mathcal{O}) \;\Bigg|\; \mathbf{X} \Bigg\} \Bigg] \\
&= \mathbb{E}\left[\frac{\mathbb{I}(R=0)}{\mathbb{P}(R=0)} \{S^k(t \mid a, \mathbf{X}) - \theta^0(t,a)\} \dot{\ell}(\mathcal{O}) \right] - \mathbb{E}\Bigg[\frac{\mathbb{I}(R=k)}{\mathbb{P}(R=0)} \frac{\mathbb{P}(R=0 \mid \mathbf{X})}{\mathbb{P}(R=k \mid \mathbf{X})} S^k(t \mid a, \mathbf{X}) \\
&\qquad \times \frac{\mathbb{I}(A=a)}{\pi^k(a \mid \mathbf{X})} \left\{ \frac{\mathbb{I}(Y \le t, \Delta=1)}{S^k(y \mid \mathbf{X}) G^k(y \mid a, \mathbf{X})} - \int_0^{t\wedge y} \frac{\Lambda^k(du \mid a, \mathbf{X})}{S^k(u \mid \mathbf{X}) G^k(u \mid a, \mathbf{X})} \right\} \dot{\ell}(\mathcal{O}) \Bigg].
\end{aligned}$$

Therefore, an EIF of $\theta^0(t,a)$ at $\mathbb{P}$ is found as

$$\begin{aligned}
&\varphi_{t,a}^{*k,0}(\mathcal{O}; \mathbb{P}) \\
&= \frac{\mathbb{I}(R=0)}{\mathbb{P}(R=0)} \{S^0(t \mid a, \mathbf{X}) - \theta^0(t,a)\} - \frac{\mathbb{I}(R=k)\mathbb{P}(R=0 \mid \mathbf{X})}{\mathbb{P}(R=0)\mathbb{P}(R=k \mid \mathbf{X})} S^k(t \mid a, \mathbf{X}) \\
&\qquad \times \frac{\mathbb{I}(A=a)}{\pi^k(a \mid \mathbf{X})} \left[\frac{\mathbb{I}(Y \le t, \Delta=1)}{S^k(Y \mid a, \mathbf{X}) G^k(Y \mid a, \mathbf{X})} - \int_0^{t\wedge Y} \frac{\Lambda^k(du \mid a, \mathbf{X})}{S^k(u \mid a, \mathbf{X}) G^k(u \mid a, \mathbf{X})} \right].
\end{aligned}$$

Observe that, by Bayes's rule,

$$\frac{\mathbb{P}(R=0 \mid \mathbf{X})}{\mathbb{P}(R=k \mid \mathbf{X})} = \underbrace{\frac{\mathbb{P}(\mathbf{X} \mid R=0)}{\mathbb{P}(\mathbf{X} \mid R=k)}}_{\omega^{k,0}(\mathbf{X})} \cdot \frac{\mathbb{P}(R=0)}{\mathbb{P}(R=k)},$$

where $\omega^{k,0}(\mathbf{X})$ is a covariates density ratio function. We then find that the EIF form in Theorem 2.5:

$$\begin{aligned}
\varphi_{t,a}^{*k,0}(\mathcal{O}; \mathbb{P}) = &\frac{\mathbb{I}(R=0)}{\mathbb{P}(R=0)} \{S^0(t \mid a, \mathbf{X}) - \theta^0(t,a)\} - \frac{\mathbb{I}(R=k)}{\mathbb{P}(R=k)} \omega^{k,0}(\mathbf{X}) S^k(t \mid a, \mathbf{X}) \\
&\times \frac{\mathbb{I}(A=a)}{\pi^k(a \mid \mathbf{X})} \left[\frac{\mathbb{I}(Y \le t, \Delta=1)}{S^k(Y \mid a, \mathbf{X}) G^k(Y \mid a, \mathbf{X})} - \int_0^{t\wedge Y} \frac{\Lambda^k(du \mid a, \mathbf{X})}{S^k(u \mid a, \mathbf{X}) G^k(u \mid a, \mathbf{X})} \right].
\end{aligned}$$

E.1.2 Regularity conditions and the RAL property of the local estimator

For site $R = k$, we denote π^k, G^k, $\omega^{k,0}$, Λ^k and S^k the truths of nuisance functions. We use π^k_∞, $\omega^{k,0}_\infty$, G^k_∞, Λ^k_∞ and S^k_∞ to denote some general probability limits for nuisance function estimators.

Condition E.1. *There exist π^k_∞, $\omega^{k,0}_\infty$, G^k_∞, Λ^k_∞ and S^k_∞ such that*

(a) $\max_m \mathbb{P}\left[\frac{1}{\widehat{\pi}^k_m(a \mid \mathbf{X})} - \frac{1}{\pi^k_\infty(a \mid \mathbf{X})}\right]^2 \to_p 0$;

(b) $\max_m \mathbb{P}\left[\widehat{\omega}^{k,0}_m(\mathbf{X}) - \omega^{k,0}_\infty(\mathbf{X})\right]^2 \to_p 0$;

(c) $\max_m \mathbb{P}\left[\sup_{u\in[0,t]}\left|\frac{1}{\widehat{G}^k_m(u \mid a, \mathbf{X})} - \frac{1}{G^k_\infty(u \mid a, \mathbf{X})}\right|\right]^2 \to_p 0$;

(d) $\max_m \mathbb{P}\left[\sup_{u\in[0,t]}\left|\frac{\widehat{S}^k_m(t \mid a, \mathbf{X})}{\widehat{S}^k_m(u \mid a, \mathbf{X})} - \frac{S^k_\infty(t \mid a, \mathbf{X})}{S^k_\infty(u \mid a, \mathbf{X})}\right|\right]^2 \to_p 0$.

Condition E.2. *There exists an $\eta \in (0, \infty)$ such that for $\mathbb{P}$-almost all $\mathbf{x}$, $\widehat{\pi}^k_m(a \mid \mathbf{x}) \geq 1/\eta$, $\pi^k_\infty(a \mid \mathbf{x}) \geq 1/\eta$, $\widehat{\omega}^{k,0}_m(\mathbf{x}) \leq \eta$, $\omega^{k,0}_\infty(\mathbf{x}) \leq \eta$, $\widehat{G}^k_m(t \mid a, \mathbf{x}) \geq 1/\eta$, and $G^k_\infty(t \mid a, \mathbf{x}) \geq 1/\eta$ with probability tending to 1.*

Condition E.3. *Define*

$$
\begin{aligned}
r^k_{n,t,a,1} &= \max_m \mathbb{P}\left|\{\widehat{\pi}^k_m(a \mid \mathbf{X}) - \pi^k(a \mid \mathbf{X})\}\{\widehat{S}^k_m(t \mid a, \mathbf{X}) - S^k(t \mid a, \mathbf{X})\}\right|, \\
r^k_{n,t,a,2} &= \max_m \mathbb{P}\left|\{\widehat{\omega}^{k,0}_m(\mathbf{X}) - \omega^{k,0}(\mathbf{X})\}\{\widehat{S}^k_m(t \mid a, \mathbf{X}) - S^k(t \mid a, \mathbf{X})\}\right|, \text{ and} \\
r^k_{n,t,a,3} &= \max_m \mathbb{P}\left|\widehat{S}^k_m(t \mid a, \mathbf{X}) \int_0^t \left\{\frac{G^k(u \mid a, \mathbf{X})}{\widehat{G}^k_m(u \mid a, \mathbf{X})} - 1\right\}\left(\frac{S^k}{\widehat{S}^k_m} - 1\right)(du \mid a, \mathbf{X})\right|.
\end{aligned}
$$

Then, it holds that $r^k_{n,t,a,1} = o_p(n^{-1/2})$, $r^k_{n,t,a,2} = o_p(n^{-1/2})$ and $r^k_{n,t,a,3} = o_p(n^{-1/2})$.

The following theorem formally establishes the RAL property of the local estimator. For simplicity of notation, we write an EIF $\varphi(\mathcal{O}; \mathbb{P})$ as φ, omitting its dependence on $\mathcal{O}$ and $\mathbb{P}$ without loss of clarity.

Theorem E.1. *If Conditions E.1–E.3 hold, with $\pi^k_\infty = \pi^k$, $\omega^{k,0}_\infty = \omega^{k,0}$, $G^k_\infty = G^k$, and $S^k_\infty = S^k$, then $\widehat{\theta}^{k,0}_n(t, a) = \theta^0(t, a) + \mathbb{P}_n(\varphi^{*k,0}_{t,a}) + o_p(n^{-1/2})$. In particular, for each $t \in [0, \tau]$ and $a \in \{0, 1\}$,*

$$
n^{1/2}(\widehat{\theta}^{k,0}_n(t, a) - \theta^0(t, a)) \to_d \mathcal{N}(0, \sigma^2), \qquad \text{where } \sigma^2 = \mathbb{P}[(\varphi^{*k,0}_{t,a})^2].
$$

To prove Theorem E.1, we first introduce some useful results and lemmata in the next section.

E.1.3 Useful lemmata for the local estimator

We start by examining the difference $\widehat{\theta}^{k,0}_n(t, a) - \theta^0(t, a)$. Recall $\mathbb{P}^m_n$ is the empirical distribution corresponding to the m-th validation set $\mathcal{V}_m$ from the entire data $\mathcal{O}$, and denote $\mathbb{G}^m_n$ the corresponding empirical process. A result exactly following Westling et al. (2024) is that

$$
\begin{aligned}
\widehat{\theta}^{k,0}_n(t, a) - \theta^0(t, a) = \mathbb{P}_n[\varphi^{*k,0}_{\infty,t,a}] &+ \frac{1}{M}\sum_{m=1}^M \frac{M n_m^{1/2}}{n}\mathbb{G}^m_n\left[\widehat{\varphi}^{k,0}_{n,m,t,a} - \varphi^{k,0}_{\infty,t,a}\right] \\
&+ \frac{1}{M}\sum_{m=1}^M \frac{M n_m}{n}\mathbb{P}\left[\widehat{\varphi}^{k,0}_{t,a} - \theta^0(t, a)\right]. \qquad (4)
\end{aligned}
$$

We then establish the $L_2(\mathbb{P})$ norm distance (bound) between the estimated EIF and its underlying limit for the local estimator by the following lemma.

Lemma E.2. *Under Condition E.2, there exists a universal constant $C = C(\eta)$ such that for each k, m, n, t, and a,*

$$\mathbb{P}[\widehat{\varphi}^{k,0}_{t,a} - \varphi^{k,0}_{\infty,t,a}]^2 \leq C(\eta) \sum_{j=1}^{6} A^k_{j,n,m,t,a},$$

where

$$\begin{aligned}
A^k_{1,n,m,t,a} &= \mathbb{P}\left[\frac{1}{\mathbb{P}^m_n(R=0)} - \frac{1}{\mathbb{P}(R=0)}\right]^2, \\
A^k_{2,n,m,t,a} &= \mathbb{P}\left[\frac{1}{\mathbb{P}^m_n(R=k)} - \frac{1}{\mathbb{P}(R=k)}\right]^2, \\
A^k_{3,n,m,t,a} &= \mathbb{P}\left[\widehat{\omega}^{k,0}_m(a \mid \mathbf{X}) - \omega^{k,0}_\infty(a \mid \mathbf{X})\right]^2, \\
A^k_{4,n,m,t,a} &= \mathbb{P}\left[\frac{1}{\widehat{\pi}^k_m(a \mid \mathbf{X})} - \frac{1}{\pi^k_\infty(a \mid \mathbf{X})}\right]^2, \\
A^k_{5,n,m,t,a} &= \mathbb{P}\left[\sup_{u\in[0,t]}\left|\frac{1}{\widehat{G}^k_m(u \mid a, \mathbf{X})} - \frac{1}{G^k_\infty(u \mid a, \mathbf{X})}\right|\right]^2, \\
A^k_{6,n,m,t,a} &= \mathbb{P}\left[\sup_{u\in[0,t]}\left|\frac{\widehat{S}^k_m(t \mid a, \mathbf{X})}{\widehat{S}^k_m(u \mid a, \mathbf{X})} - \frac{S^k_\infty(t \mid a, \mathbf{X})}{S^k_\infty(u \mid a, \mathbf{X})}\right|\right]^2.
\end{aligned}$$

Proof. We first denote

$$\begin{aligned}
B^k(\mathcal{V}_m) &= \frac{\mathbb{I}(A=a)}{\pi^k(a \mid \mathbf{X})} S^k(t \mid a, \mathbf{X}) \\
&\quad \times \left[\frac{\mathbb{I}(Y \leq t, \Delta = 1)}{S^k(Y \mid a, \mathbf{X}) G^k(Y \mid a, \mathbf{X})} - \int_0^{t\wedge Y} \frac{\Lambda^k(du \mid a, \mathbf{X})}{S^k(u \mid a, \mathbf{X}) G^k(u \mid a, \mathbf{X})}\right], \\
C^k(\mathcal{V}_m) &= B^k(\mathcal{V}_m)\omega^{k,0}(\mathbf{X}).
\end{aligned}$$

Then, we first have the following decomposition:

$$\widehat{\varphi}^{k,0}_{t,a} - \varphi^{k,0}_{\infty,t,a} = \sum_{j=1}^{4} U^k_{j,n,m,t,a},$$

where

$$\begin{aligned}
U^k_{1,n,m,t,a} &= \left[\frac{\mathbb{I}(R=0)}{\mathbb{P}^m_n(R=0)} - \frac{\mathbb{I}(R=0)}{\mathbb{P}(R=0)}\right] \widehat{S}^0_m(t \mid a, \mathbf{x}), \\
U^k_{2,n,m,t,a} &= \frac{\mathbb{I}(R=0)}{\mathbb{P}(R=0)}\left[\widehat{S}^0_m(t \mid a, \mathbf{x}) - S^0_\infty(t \mid a, \mathbf{x})\right], \\
U^k_{3,n,m,t,a} &= \left[\frac{\mathbb{I}(R=k)}{\mathbb{P}^m_n(R=k)} - \frac{\mathbb{I}(R=k)}{\mathbb{P}(R=k)}\right] \widehat{C}^k_m(\mathcal{V}_m), \\
U^k_{4,n,m,t,a} &= \frac{\mathbb{I}(R=k)}{\mathbb{P}(R=k)}\left[\widehat{C}^k_m(\mathcal{V}_m) - C^k_\infty(\mathcal{V}_m)\right].
\end{aligned}$$

Now, for $U^k_{4,n,m,t,a}$, we further decompose it as

$$U^k_{4,n,m,t,a} = \frac{\mathbb{I}(R=k)}{\mathbb{P}(R=k)} \sum_{j=1}^{2} V^k_{j,n,m,t,a},$$

where

$$\begin{aligned}
V^k_{1,n,m,t,a} &= B^k_\infty(\mathcal{V}_m)\left[\widehat{\omega}^{k,0}_m(\mathbf{X}) - \omega^{k,0}_m(\mathbf{X})\right], \\
V^k_{2,n,m,t,a} &= \widehat{\omega}^{k,0}_m(\mathbf{X})\left[\widehat{B}^k_m(\mathcal{V}_m) - B^k_\infty(\mathcal{V}_m)\right].
\end{aligned}$$

The expression of $\widehat{B}_m^k(\mathcal{V}_m) - B_\infty^k(\mathcal{V}_m)$ is exactly the same as the Lemma 3 in Westling et al. (2024), while we only need to replace the corresponding nuisance functions by the site-k version here, so the detail is omitted. By the triangle inequality, we have $\mathbb{P}[\widehat{\varphi}_{t,a}^{k,0} - \varphi_{\infty,t,a}^{k,0}]^2 \leq \left\{\sum_{j=1}^4 \{\mathbb{P}[(U_{j,n,m,t,a}^k)^2]\}^{1/2}\right\}^2$. Therefore, under Assumption 2.3 and Condition E.2, there exists a universal constant $C = C(\eta)$ such that the result in the statement holds. □

Furthermore, we need to bound the empirical process term $\mathbb{G}_n^m\left[\widehat{\varphi}_{n,m,t,a}^{k,0} - \varphi_{\infty,t,a_0}^{k,0}\right]$ by $o_p(n^{-1/2})$. This is formally shown below in Lemma E.3.

Lemma E.3. *If Conditions E.1–E.2 hold,* $M^{-1}\sum_{m=1}^M n^{-1} M n_m^{1/2} \mathbb{G}_n^m\left[\widehat{\varphi}_{n,m,t,a}^{k,0} - \varphi_{\infty,t,a_0}^{k,0}\right] = o_p(n^{-1/2})$.

Proof. We follow notation in Lemma E.2. First, we note that

$$\begin{aligned}\frac{M n_m^{1/2}}{n} &\leq \frac{M(|n_m - n/M| + n/M)^{1/2}}{n} \\ &\leq \frac{M|n_m - n/M|^{1/2} + M|n/M|^{1/2}}{n} \\ &\leq \left(\frac{M}{n}\right)^{1/2} + \frac{M}{n},\end{aligned}$$

for all m since $|n_m - n/M| \leq 1$ by assumption on n_m. Then, we have that

$$\begin{aligned}&\frac{1}{M}\sum_{m=1}^M \frac{M n_m^{1/2}}{n} \sup_{u\in[0,t]}\left|\mathbb{G}_n^m\left[\widehat{\varphi}_{n,m,t,a}^{k,0} - \varphi_{\infty,t,a}^{k,0}\right]\right| \\ &\leq O(n^{-1/2})\frac{1}{M}\sum_{m=1}^M \sup_{u\in[0,t]}\left|\mathbb{G}_n^m\left[\widehat{\varphi}_{n,m,t,a}^{k,0} - \varphi_{\infty,t,a}^{k,0}\right]\right|,\end{aligned}$$

since $K = O(1)$.

Therefore, we turn to show $M^{-1}\sum_{m=1}^M \left|\mathbb{G}_n^m\left[\widehat{\varphi}_{n,m,t,a}^{k,0} - \varphi_{\infty,t,a}^{k,0}\right]\right| = o_p(1)$. Using conditional argument, we write

$$\mathbb{E}\left|\mathbb{G}_n^m\left[\widehat{\varphi}_{n,m,t,a}^{k,0} - \varphi_{\infty,t,a}^{k,0}\right]\right| = \mathbb{E}\left[\mathbb{E}\left|\mathbb{G}_n^m\left[\widehat{\varphi}_{n,m,t,a}^{k,0} - \varphi_{\infty,t,a}^{k,0}\right]\right| \mid \mathcal{T}_m\right],$$

where $\mathcal{T}_m = \mathcal{O}\backslash\mathcal{V}_m$ is the m-th training set. Note that the randomness in the inner expectation of the right-hand-side above, by conditioning on the training set, is only induced from $\mathbb{G}_n^m$ by averaging over the observations on the validation set. Therefore,

$$\mathbb{E}\left[\mathbb{E}\left|\mathbb{G}_n^m\left[\widehat{\varphi}_{n,m,t,a}^{k,0} - \varphi_{\infty,t,a}^{k,0}\right]\right| \mid \mathcal{T}_m\right] = \mathbb{P}\left|\mathbb{G}_n^m(\widehat{\varphi}_{n,m,t,a}^{k,0} - \varphi_{\infty,t,a}^{k,0})\right|.$$

Defining $\mathcal{F}_{n,m,t,a}^{k,0}$ as the singleton class of functions $\widehat{\varphi}_{n,m,t,a}^{k,0} - \varphi_{\infty,t,a}^{k,0}$, we further have

$$\mathbb{P}\left|\mathbb{G}_n^m(\widehat{\varphi}_{n,m,t,a}^{k,0} - \varphi_{\infty,t,a}^{k,0})\right| = \mathbb{P}\left[\sup_{f\in\mathcal{F}_{n,m,t,a}^{k,0}} |\mathbb{G}_n^m(f)|\right].$$

By Theorem 2.1.14 in Van der Vaart & Wellner (1996), the covering number of $\mathcal{F}_{n,m,t,a}^{k,0}$ is 1 for all ε, so the uniform entropy integral $J(1, \mathcal{F}_{n,m,t,a}^{k,0})$ is 1 relative to the natural envelope $|\widehat{\varphi}_{n,m,t,a}^{k,0} - \varphi_{\infty,t,a}^{k,0}|$. Therefore, there is a universal constant C' such that

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}_{n,m,t,a}^{k,0}} |\mathbb{G}_n^m(f)|\right] \leq C'\left\{\mathbb{P}(\widehat{\varphi}_{n,m,t,a}^{k,0} - \varphi_{\infty,t,a}^{k,0})^2\right\}^{1/2} \leq C''\sum_{j=1}^6 \bar{A}_{j,n,m,t,a},$$

following definition of $\bar{A}_{j,n,m,t,a}$ terms in Lemma E.2, so that $M^{-1}\sum_{m=1}^{M}\mathbb{E}\left|\mathbb{G}_n^m\left[\widehat{\varphi}_{n,m,t,a}^{k,0}-\varphi_{\infty,t,a}^{k,0}\right]\right|$ is bounded up to $C'''\sum_{j=1}^{6}\mathbb{E}[\max_m(\bar{A}_{j,n,m,t,a})]$ for some constant C'''. It is straightforward that by Conditions E.1 and E.2, this upper bound tends to zero, so $M^{-1}\sum_{m=1}^{M}\left|\mathbb{G}_n^m\left[\widehat{\varphi}_{n,m,t,a}^{k,0}-\varphi_{\infty,t,a}^{k,0}\right]\right|=o_p(1)$. □

Finally, the only difference we have not characterized in (4) is $\mathbb{P}[\varphi_{t,a}^{k,0}(\mathcal{O};\mathbb{P}_\infty)]-\theta^0(t,a)$, which we show it below.

Lemma E.4. *Consider some general nuisance functions under $\mathbb{P}_\infty$, denoted by S_∞^0, S_∞^k, G_∞^k, π_∞^k, and $\omega_\infty^{k,0}$ (equals 1 if $k=0$). Then, $\mathbb{P}[\varphi_{t,a}^{k,0}(\mathcal{O};\mathbb{P}_\infty)]-\theta^0(t,a)$ equals*

$$\mathbb{E}\Bigg[\frac{q^0(\mathbf{X})}{\mathbb{P}(R=0)}S_\infty^k(t\mid a,\mathbf{X})\int_0^t\frac{S^k(y-\mid a,\mathbf{X})}{S_\infty^k(y\mid a,\mathbf{X})}\times\left\{\frac{\omega_\infty^{k,0}(\mathbf{X})G^k(y\mid a,\mathbf{X})\pi^k(a\mid\mathbf{X})}{\omega^{k,0}(\mathbf{X})G_\infty^k(y\mid a,\mathbf{X})\pi_\infty^k(a\mid\mathbf{X})}-1\right\}(\Lambda_\infty^k-\Lambda^k)(dy\mid a,\mathbf{X})\Bigg].$$

Proof. By direct calculations, $\mathbb{P}[\varphi_{t,a}^{k,0}(\mathcal{O};\mathbb{P}_\infty)]-\theta^0(t,a)$ equals

$$\begin{aligned}&\mathbb{E}\Bigg[\frac{\mathbb{I}(R=0)}{\mathbb{P}(R=0)}\{S_\infty^0(t\mid a,\mathbf{X})-S^0(t\mid a,\mathbf{X})\}+\frac{q^k(\mathbf{X})}{\mathbb{P}(R=k)}\omega_\infty^{k,0}(\mathbf{X})S_\infty^k(t\mid a,\mathbf{X})\frac{\pi^k(a\mid\mathbf{X})}{\pi_\infty^k(a\mid\mathbf{X})}\\&\qquad\times\int_0^t\frac{S^k(y-\mid a,\mathbf{X})G^k(y\mid a,\mathbf{X})}{S_\infty^k(y\mid a,\mathbf{X})G_\infty^k(y\mid a,\mathbf{X})}(\Lambda_\infty^k-\Lambda^k)(dy\mid a,\mathbf{X})\Bigg]\\&=\mathbb{E}\Bigg[\frac{q^0(\mathbf{X})}{\mathbb{P}(R=0)}\{S_\infty^0(t\mid a,\mathbf{X})-S^0(t\mid a,\mathbf{X})\}+\frac{q^0(\mathbf{X})}{\mathbb{P}(R=0)}\frac{\omega_\infty^{k,0}(\mathbf{X})}{\omega^{k,0}(\mathbf{X})}S_\infty^k(t\mid a,\mathbf{X})\frac{\pi^k(a\mid\mathbf{X})}{\pi_\infty^k(a\mid\mathbf{X})}\\&\qquad\times\int_0^t\frac{S^k(y-\mid a,\mathbf{X})G^k(y\mid a,\mathbf{X})}{S_\infty^k(y\mid a,\mathbf{X})G_\infty^k(y\mid a,\mathbf{X})}(\Lambda_\infty^k-\Lambda^k)(dy\mid a,\mathbf{X})\Bigg].\end{aligned}$$

In the second "$\mathbb{E}$" after "$=$", we used the following relationship:

$$\frac{q^0(\mathbf{X})}{q^k(\mathbf{X})}=\omega^{k,0}(\mathbf{X})\frac{\mathbb{P}(R=0)}{\mathbb{P}(R=k)}$$

by Bayes's rule. Furthermore, by Duhamel equation in Gill & Johansen (1990),

$$\begin{aligned}&\mathbb{P}[\varphi_{t,a}^{k,0}(\mathcal{O};\mathbb{P}_\infty)]-\theta^0(t,a)\\&=\mathbb{E}\Bigg[\frac{q^0(\mathbf{X})}{\mathbb{P}(R=0)}S_\infty^k(t\mid a,\mathbf{X})\int_0^t\frac{S^k(y-\mid a,\mathbf{X})}{S_\infty^k(y\mid a,\mathbf{X})}\\&\qquad\times\left\{\frac{\omega_\infty^{k,0}(\mathbf{X})G^k(y\mid a,\mathbf{X})\pi^k(a\mid\mathbf{X})}{\omega^{k,0}(\mathbf{X})G_\infty^k(y\mid a,\mathbf{X})\pi_\infty^k(a\mid\mathbf{X})}-1\right\}(\Lambda_\infty^k-\Lambda^k)(dy\mid a,\mathbf{X})\Bigg].\end{aligned}\tag{5}$$

□

E.1.4 Proof of Theorem E.1

By (4) with $\pi_\infty^k=\pi^k$, $\omega_\infty^{k,0}=\omega^{k,0}$, $G_\infty^k=G^k$, and $S_\infty^k=S^k$,

$$\begin{aligned}\widehat{\theta}_n^{k,0}(t,a)-\theta^0(t,a)=\mathbb{P}_n[\varphi_{t,a}^{*k,0}]&+\frac{1}{M}\sum_{m=1}^{M}\frac{Mn_m^{1/2}}{n}\mathbb{G}_n^m\left[\widehat{\varphi}_{n,m,t,a}^{k,0}-\varphi_{t,a}^{k,0}\right]\\&+\frac{1}{M}\sum_{m=1}^{M}\frac{Mn_m}{n}\mathbb{P}\left[\widehat{\varphi}_{t,a}^{k,0}-\theta^0(t,a)\right].\end{aligned}$$

By Conditions E.1 and E.2, the second summand on the right-hand-side is $o_p(n^{-1/2})$ by Lemma E.3. By Lemma E.4, $\mathbb{P}[\widehat{\varphi}^{k,0}_{t,a}] - \theta^0(t,a)$ equals

$$\mathbb{E}\Bigg[\frac{q^0(\mathbf{X})}{\mathbb{P}(R=0)}\widehat{S}^k_m(t \mid a, \mathbf{X}) \int_0^t \frac{S^k(y- \mid a, \mathbf{X})}{\widehat{S}^k_m(y \mid a, \mathbf{X})} \times \left\{\frac{\widehat{\omega}^{k,0}_m(\mathbf{X}) G^k(y \mid a, \mathbf{X})\pi^k(a \mid \mathbf{X})}{\omega^{k,0}(\mathbf{X})\widehat{G}^k_m(y \mid a, \mathbf{X})\widehat{\pi}^k_m(a \mid \mathbf{X})} - 1\right\} (\widehat{\Lambda}^k_m - \Lambda^k)(dy \mid a, \mathbf{X})\Bigg].$$

By Duhamel equation in Gill & Johansen (1990) and Condition E.3, we find that the above bias term can be bounded by $\eta^2\{r^k_{n,t,a,1} + r^k_{n,t,a,2} + r^k_{n,t,a,3}\}$ over m. Since $M^{-1}\sum_{m=1}^M n^{-1} M n_m \leq 2$, we have

$$\left|\frac{1}{M}\sum_{m=1}^M \frac{M n_m}{n}\mathbb{P}\left[\widehat{\varphi}^{k,0}_{t,a} - \theta^0(t,a)\right]\right| \leq 2\eta^2\left\{r^k_{n,t,a,1} + r^k_{n,t,a,2} + r^k_{n,t,a,3}\right\} = o_p(n^{-1/2}),$$

by Condition E.3. This established the pointwise RAL property: $\widehat{\theta}^{k,0}_n(t,a) = \theta^0(t,a) + \mathbb{P}_n(\varphi^{*k,0}_{t,a}) + o_p(n^{-1/2})$. Since $\varphi^{*k,0}_{t,a}$ is uniformly bounded, $\mathbb{P}\{(\varphi^{*k,0}_{t,a})^2\} < \infty$ and since $\mathbb{P}\{\varphi^{*k,0}_{t,a}\} = 0$, then

$$n^{1/2}\mathbb{P}_n(\widehat{\varphi}^{*k,0}_{t,a}) \to_d \mathcal{N}(0, \mathbb{P}\{(\varphi^{*k,0}_{t,a})^2\}).$$

Remark E.5 (Double robustness of the local estimator). If we only need the consistency of $\widehat{\theta}^k_n(t,a)$, then condition $\pi^k_\infty = \pi^k$, $\omega^{k,0}_\infty = \omega^{k,0}$, $G^k_\infty = G^k$, and $S^k_\infty = S^k$ can be replaced by the following statement: For $\mathbb{P}$-almost all $\mathbf{X}$, there exist measurable sets $\mathcal{S}^k_x, \mathcal{G}^k_x \subseteq [0,t]$ such that $\mathcal{S}^k_x \cup \mathcal{G}^k_x = [0,t]$ and $\Lambda^k(u \mid a, \mathbf{X}) = \Lambda^k_\infty(u \mid a, \mathbf{X})$ for all $u \in S^k_x$ and $G(u \mid a, \mathbf{X}) = G^k_\infty(u \mid a, \mathbf{X})$ for all $u \in \mathcal{G}^k_x$. In addition, if $\mathcal{S}^k_x$ is a strict subset of $[0,t]$, then $\pi^k(a \mid \mathbf{X}) = \pi^k_\infty(a \mid \mathbf{X})$ and $\omega^{k,0}(\mathbf{X}) = \omega^{k,0}_\infty(\mathbf{X})$ as well. Then, $\widehat{\theta}^k_n(t,a)$ is consistent if Conditions E.1 and E.2 hold. This statement could be interpreted as that at a given time t, if either (i) the conditional survival model S^k; or (ii) all other nuisance functions G^k, π^k and $\omega^{k,0}$ are correctly specified (with other conditions above), $\widehat{\theta}^k_n(t,a)$ is consistent.

E.2 Theory for the federated estimator

In this section, we present the properties of the federated estimator. Given that our proposed weights, $\boldsymbol{\eta}_{t,a}$, are both time- and treatment-specific, we focus on the pointwise convergence properties.

Let the set of all source site indices be $\mathcal{S} = \{1, \ldots, K-1\}$. We then define the oracle selection space for $\boldsymbol{\eta}_{t,a}$, and the corresponding weight space as:

$$\mathcal{S}^*_{t,a} = \{k \in \mathcal{S} : \theta^k(t,a) = \theta^0(t,a)\}, \text{ and } \mathbb{R}^{S^*_{t,a}} = \{\boldsymbol{\eta}_{t,a} \in \mathbb{R}^{K-1} : \eta^j_{t,a} = 0, \forall j \notin \mathcal{S}^*_{t,a}\},$$

respectively.

The space $\mathcal{S}^*_{t,a}$ is both time- and treatment-varying, indicating that a source site may not consistently be useful or unhelpful across different time points or treatments. However, it offers the advantage of increased flexibility and adaptivity, allowing for more effective borrowing of information at different points along the survival functions. Based on the theory presented in Section E.1, for $k \in \mathcal{S}^*_{t,a}$, the site-specific estimator $\widehat{\theta}^{k,0}_n(t,a)$ is consistent for $\theta^0(t,a)$ for any given $t \in [0,\tau]$ and $a \in \{0,1\}$.

We begin by assuming fixed $\boldsymbol{\eta}_{t,a} = (\eta^0_{t,a}, \eta^1_{t,a}, \ldots, \eta^{K-1}_{t,a})$. We invoke Lemmata 4 and 5 in Han et al. (2025), which state that the proposed adaptive estimation for $\eta^k_{t,a}$ as shown in (2) allows for (i) the recovery of the optimal $\eta^k_{t,a}$ by the estimator $\widehat{\eta}^k_{t,a}$, and (ii) the uncertainty induced by $\widehat{\eta}^k_{t,a}$ is negligible when estimating $\theta^0(t,a)$. We require regularity Conditions E.1, E.2 and E.3 for the pointwise convergence result in Theorem E.1 hold. Let us denote the federated estimator by plugging-in the fixed $\boldsymbol{\eta}_{t,a}$ as

$$\widehat{\theta}^{\text{fed}}_n(t,a;\boldsymbol{\eta}_{t,a}) = \left(1 - \sum_{k\in\mathcal{S}}\eta^k_{t,a}\right)\widehat{\theta}^0_n(t,a) + \sum_{k\in\mathcal{S}}\eta^k_{t,a}\widehat{\theta}^{k,0}_n(t,a).$$

Recall that notation $\mathcal{H}_{t,a}$ defined in (1):

$$\mathcal{H}_{t,a}(\mathcal{O}; S, G) = \frac{\mathbb{I}(Y \le t, \delta = 1)}{S(Y \mid a, \mathbf{X}) G(Y \mid a, \mathbf{X})} - \int_0^{t \wedge Y} \frac{\Lambda(du \mid a, \mathbf{X})}{S(u \mid a, \mathbf{X}) G(u \mid a, \mathbf{X})}.$$

Let us then write

$$\begin{aligned}
\xi^{0,(1)}(\mathcal{O}) &= S^0(t \mid a, \mathbf{X}) \frac{\mathbb{I}(A = a)}{\pi^0(a \mid \mathbf{X})} \mathcal{H}_{t,a}(\mathcal{O}; S^0, \Lambda^0, G^0), \\
\xi^{k,0,(1)}(\mathcal{O}) &= \omega^{k,0}(\mathbf{X}) S^k(t \mid a, \mathbf{X}) \frac{\mathbb{I}(A = a)}{\pi^k(a \mid \mathbf{X})} \mathcal{H}_{t,a}(\mathcal{O}; S^k, \Lambda^k, G^k), \\
\xi^{0,(2)}(\mathcal{O}) &= S^0(t \mid a, \mathbf{X}) - \theta^0(t, a),
\end{aligned}$$

and $n_k = \sum_{i=1}^n \mathbb{I}(R_i = k)$ for $k = 0, 1, \ldots, K-1$.

Then,

$$\begin{aligned}
&\widehat{\theta}_n^{\text{fed}}(t, a; \boldsymbol{\eta}_{t,a}) - \theta^0(t, a) \\
&= \left(1 - \sum_{k \in \mathcal{S}} \eta_{t,a}^k\right) \left\{\widehat{\theta}_n^0(t, a) - \theta^0(t, a)\right\} + \sum_{k \in \mathcal{S}} \eta_{t,a}^k \left\{\widehat{\theta}_n^{k,0}(t, a) - \theta^0(t, a)\right\} \\
&= \left(1 - \sum_{k \in \mathcal{S}} \eta_{t,a}^k\right) \frac{1}{n_0} \sum_{i=1}^n \mathbb{I}(R_i = 0) \left\{\widehat{\xi}^{0,(2)}(\mathcal{O}_i) - \widehat{\xi}^{0,(1)}(\mathcal{O}_i)\right\} \\
&\quad + \sum_{k \in \mathcal{S}} \frac{1}{n_0} \sum_{i=1}^n \mathbb{I}(R_i = 0) \eta_{t,a}^k \widehat{\xi}^{0,(2)}(\mathcal{O}_i) - \sum_{k \in \mathcal{S}} \frac{1}{n_k} \sum_{i=1}^n \mathbb{I}(R_i = k) \eta_{t,a}^k \widehat{\xi}^{k,0,(1)}(\mathcal{O}_i) \\
&= \frac{1}{n} \sum_{i=1}^n \left(1 - \sum_{k \in \mathcal{S}} \eta_{t,a}^k\right) \mathbb{I}(R_i = 0) \frac{\widehat{\xi}^{0,(2)}(\mathcal{O}_i) - \widehat{\xi}^{0,(1)}(\mathcal{O}_i)}{\widehat{\mathbb{P}}(R_i = 0)} \\
&\quad + \frac{1}{n} \sum_{i=1}^n \mathbb{I}(R_i = 0) \left(\sum_{k \in \mathcal{S}} \eta_{t,a}^k\right) \frac{\widehat{\xi}^{0,(2)}(\mathcal{O}_i)}{\widehat{\mathbb{P}}(R_i = 0)} - \frac{1}{n} \sum_{k \in \mathcal{S}} \sum_{i=1}^n \mathbb{I}(R_i = k) \eta_{t,a}^k \frac{\widehat{\xi}^{k,0,(1)}(\mathcal{O}_i)}{\widehat{\mathbb{P}}(R_i = k)}. \qquad (6)
\end{aligned}$$

The asymptotic variance of $\widehat{\theta}_n^{\text{fed}}(t, a; \boldsymbol{\eta}_{t,a})$ equals the variance of the influence function of (6). Let us denote it as $\mathcal{V}_{t,a}^{\text{fed}} = \mathcal{V}_{t,a}^{\text{fed}}(\boldsymbol{\eta}_{t,a})$. We highlight its dependence to the federated weights vector $\boldsymbol{\eta}_{t,a}$ here because in the below (8), we consider an optimization program for deriving the weights based on minimizing the (estimated) asymptotic variance.

Under the assumption of i.i.d. participants within each site, we have

$$\begin{aligned}
\mathcal{V}_{t,a}^{\text{fed}} &= \left(1 - \sum_{k \in \mathcal{S}} \eta_{t,a}^k\right)^2 \frac{\mathbb{V}\{\xi^{0,(2)}(\mathcal{O}_i) - \xi^{0,(1)}(\mathcal{O}_i) \mid R_i = 0\}}{\mathbb{P}(R_i = 0)} \\
&\quad + \left(\sum_{k \in \mathcal{S}} \eta_{t,a}^k\right)^2 \frac{\mathbb{V}\{\xi^{0,(2)}(\mathcal{O}_i) \mid R_i = 0\}}{\mathbb{P}(R_i = 0)} \\
&\quad + 2 \left(1 - \sum_{k \in \mathcal{S}} \eta_{t,a}^k\right) \left(\sum_{k \in \mathcal{S}} \eta_{t,a}^k\right) \frac{\text{Cov}\{\xi^{0,(2)}(\mathcal{O}_i) - \xi^{0,(1)}(\mathcal{O}_i), \xi^{0,(2)}(\mathcal{O}_i) \mid R_i = 0\}}{\mathbb{P}(R_i = 0)} \\
&\quad + \sum_{k \in \mathcal{S}} (\eta_{t,a}^k)^2 \frac{\mathbb{V}\{\xi^{k,0,(1)}(\mathcal{O}_i) \mid R_i = k\}}{\mathbb{P}(R_i = k)}. \qquad (7)
\end{aligned}$$

With appropriate boundedness conditions on conditional variance and covariance terms above, $\mathcal{V}_{t,a}^{\text{fed}} < \infty$ (see Lemma E.7). Consequently, the asymptotic distribution of $\widehat{\theta}_n^{\text{fed}}(t, a; \boldsymbol{\eta}_{t,a})$ is given by

$$\sqrt{n} \left\{\widehat{\theta}_n^{\text{fed}}(t, a; \boldsymbol{\eta}_{t,a}) - \theta^0(t, a)\right\} \to_d \mathcal{N}(0, \mathcal{V}_{t,a}^{\text{fed}}).$$

Remark E.6. Based on the derivations in (6) and (7), an influence-function–based asymptotic variance estimator of $\widehat{\theta}_n^{\text{fed}}(t,a)$ ($\widehat{\mathcal{V}}_{t,a}^{\text{fed}}$ in Theorem 2.7), is obtained by replacing the population proportions, variances, and covariances in (7) with their sample (empirical) counterparts and plugging in the estimated weight vector $\widehat{\boldsymbol{\eta}}_{t,a}$.

We further define the optimal adaptive weights $\bar{\boldsymbol{\eta}}_{t,a}$ as follows:

$$\bar{\boldsymbol{\eta}}_{t,a} = \underset{\eta_{t,a}^k=0,\forall k\notin\mathcal{S}_{t,a}^*}{\arg\min}\ \mathcal{V}_{t,a}^{\text{fed}}(\boldsymbol{\eta}_{t,a}). \tag{8}$$

We adapt two lemmata from Han et al. (2025) for recovering the optimal weights $\bar{\boldsymbol{\eta}}_{t,a}$ with negligible uncertainty for estimating $\theta^0(t,a)$ if we estimate $\boldsymbol{\eta}_{t,a}$ using (2), akin to adaptive Lasso (Zou, 2006; Fan et al., 2024).

Lemma E.7 (adapted from Lemma 4 in Han et al. (2025))**.** *Under Conditions E.1—E.3, along with the following mild conditions on covariates support and covariances: (i) The covariates* $\mathbf{X}$ *and density ratio* $\omega^{k,0}(\mathbf{X})$ *are in compact sets* $\mathbf{X}\in[-B,B]^p$ *and* $\omega^{k,0}(\mathbf{X})\in[-B,B]$ *for all* $k=1,\ldots,K-1$ *with probability* 1*; and (ii) The variance of* $\xi^{k,0,(1)}(\mathcal{O})\in[\varepsilon,M]$*, and the variance-covariance matrix* $\mathcal{V}\big[(\xi^{0,(1)},\xi^{0,(2)})'\mid R=0\big]$ *has eigenvalues in* $[\varepsilon,B]$ *for some positive constants* ε *and* B*. Then, it holds that*

$$\lim_{n\to\infty}\mathbb{P}(\widehat{\boldsymbol{\eta}}_{t,a}\in\mathbb{R}^{S_{t,a}^*})=1,\quad \|\widehat{\boldsymbol{\eta}}_{t,a}-\bar{\boldsymbol{\eta}}_{t,a}\|=O_p(n^{-1/2}),$$

for all $(t,a)\in[0,\tau]\times\{0,1\}$.

Lemma E.8 (adapted from Lemma 5 in Han et al. (2025))**.** *Under conditions in Lemma E.7,*

$$\sqrt{n}\left(\widehat{\theta}^{fed}(t,a;\widehat{\boldsymbol{\eta}}_{t,a})-\theta^0(t,a)\right)\to_d\mathcal{N}\big(0,\mathcal{V}_{t,a}^{fed}(\bar{\boldsymbol{\eta}}_{t,a})\big),$$

for all $(t,a)\in[0,\tau]\times\{0,1\}$.

The consistency of $\widehat{\mathcal{V}}_{t,a}^{\text{fed}}=\widehat{\mathcal{V}}_{t,a}^{\text{fed}}(\widehat{\boldsymbol{\eta}}_{t,a})$ follows when we can effectively approximate $\mathcal{V}_{t,a}^{\text{fed}}(\bar{\boldsymbol{\eta}}_{t,a})$ with $\widehat{\mathcal{V}}_{t,a}^{\text{fed}}$. Thus,

$$\sqrt{n/\widehat{\mathcal{V}}_{t,a}^{\text{fed}}}\left\{\widehat{\theta}_n^{\text{fed}}(t,a)-\theta^0(t,a)\right\}\to_d\mathcal{N}(0,1).$$

We now analyze the efficiency gain resulting from the federation process. The estimator relies only on the target data is denoted as $\widehat{\theta}_n^0(t,a)=\widehat{\theta}_n^{\text{fed}}(t,a;\boldsymbol{\eta}_{t,a}^0)$, where $\boldsymbol{\eta}_{t,a}^0$ assigns all weights to the target and none to the source. In contrast, the estimator that leverages the proposed adaptive ensemble approach is denoted as $\widehat{\theta}_n^{\text{fed}}(t,a;\widehat{\boldsymbol{\eta}}_{t,a})$. Here $\widehat{\boldsymbol{\eta}}_{t,a}$ can recover the optimal weights $\bar{\boldsymbol{\eta}}_{t,a}$ that are associated with the minimum asymptotic variance. Consequently, the variance of $\widehat{\theta}_n^{\text{fed}}(t,a;\widehat{\boldsymbol{\eta}}_{t,a})$ is no larger than that of the estimator relying solely on the target data since $\boldsymbol{\eta}_{t,a}^0$ is generally not the variance minimizer.

To establish that the asymptotic variance of $\widehat{\theta}_n^{\text{fed}}(t,a;\widehat{\boldsymbol{\eta}}_{t,a})$ is strictly smaller than that of the estimator based solely on the target data $\widehat{\theta}_n^0(t,a)$, we adopt Proposition 1 in Han et al. (2025) with a modified informative source condition (modified Assumption 3(b) in Han et al. (2025)).

Specifically, for each source site $s\in\mathcal{S}_{t,a}^*$, we define $\widehat{\theta}_n^{\text{fed}}(t,a;\eta_{t,a}^s)$ a federated estimator where $\eta_{t,a}^s$ is the optimal ensemble weight of site s if we only consider target site and this source site s for the federation. Then, the modified informative source condition is given as

$$\left|\text{Cov}\left[\sqrt{n}\widehat{\theta}_n^0(t,a),\sqrt{n}\left\{\widehat{\theta}_n^{\text{fed}}(t,a;\eta_{t,a}^s)-\widehat{\theta}_n^0(t,a)\right\}\right]\right|\geq\varepsilon,$$

for some $\varepsilon > 0$, where $\widehat{\theta}_n^{\text{fed}}(t,a;\eta_{t,a}^s) - \widehat{\theta}_n^0(t,a)$ can be expressed as

$$
\begin{aligned}
&\widehat{\theta}_n^{\text{fed}}(t,a;\eta_{t,a}^s) - \widehat{\theta}_n^0(t,a) \\
&= \left\{\widehat{\theta}_n^{\text{fed}}(t,a;\eta_{t,a}^s) - \theta^0(t,a)\right\} - \left\{\widehat{\theta}_n^0(t,a) - \theta^0(t,a)\right\} \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{I}(R_i = 0)(1-\eta_{t,a}^s)\frac{\widehat{\xi}^{0,(2)}(\mathcal{O}_i) - \widehat{\xi}^{0,(1)}(\mathcal{O}_i)}{\widehat{\mathbb{P}}(R_i = 0)} + \frac{1}{n}\sum_{i=1}^n \mathbb{I}(R_i = 0)\eta_{t,a}^s \frac{\widehat{\xi}^{0,(2)}(\mathcal{O}_i)}{\widehat{\mathbb{P}}(R_i = 0)} \\
&\quad - \frac{1}{n}\sum_{i=1}^n \mathbb{I}(R_i = s)\eta_{t,a}^s \frac{\widehat{\xi}^{s,0,(1)}(\mathcal{O}_i)}{\widehat{\mathbb{P}}(R_i = s)} - \frac{1}{n}\sum_{i=1}^n \mathbb{I}(R_i = 0)\frac{\widehat{\xi}^{0,(2)}(\mathcal{O}_i) - \widehat{\xi}^{0,(1)}(\mathcal{O}_i)}{\widehat{\mathbb{P}}(R_i = 0)} \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{I}(R_i = 0)\eta_{t,a}^s \frac{\widehat{\xi}^{0,(1)}(\mathcal{O}_i)}{\widehat{\mathbb{P}}(R_i = 0)} - \frac{1}{n}\sum_{i=1}^n \mathbb{I}(R_i = s)\eta_{t,a}^s \frac{\widehat{\xi}^{s,0,(1)}(\mathcal{O}_i)}{\widehat{\mathbb{P}}(R_i = s)}.
\end{aligned}
$$

Therefore, it is straightforward to see that the modified condition can be achieved if $\eta_{t,a}^s > 0$.