Efficient data fusion for time-to-event outcomes with right-censored and current status data

Xiudi Li

Division of Biostatistics
University of California, Berkeley

June 24, 2026
IMSI Workshop New Horizons on Model Transportability and Data Integration

Acknowledgement

Sijia (Lucy) Li
Assistant Professor
Department of Biostatistics, UCLA

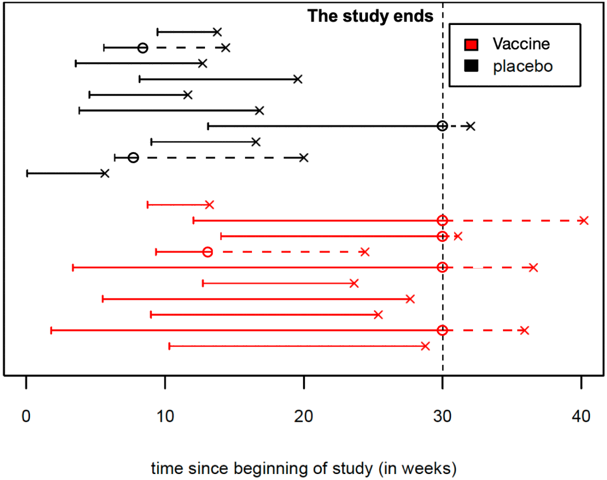
Outline

1. Background and motivation
2. Proposed Data fusion estimators
   - Doubly robust estimator
   - Efficient estimator
3. Numerical results
4. Extension and discussion

Data fusion under mixed censoring types

- Time-to-event outcomes are often partially observed (coarsened)
- common forms of censoring include
  - right censoring
  - interval censoring (with current status as a special case)

Illustration of right-censored data

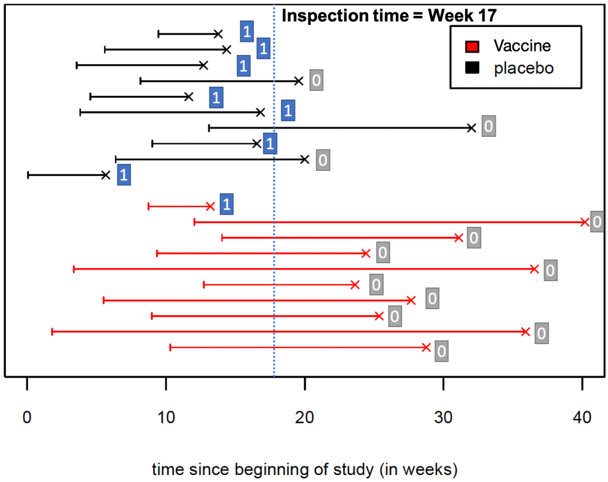


Source: modified from Dr. Marco Carone (BIOST 537 slides)

Illustration of right-censored data

- Event time is either exactly observed, or known to be larger than some censoring time
  - observe $Y = \min\{T, R\}$, $\Delta_R = 1\{T \leq R\}$
  - T is the event time of interest, R is the censoring time
- Right censoring is often seen under longitudinal follow-up
  - can happen due to loss of follow-up, end of study, etc

Illustration of current status data

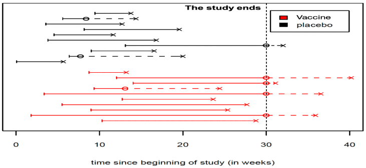


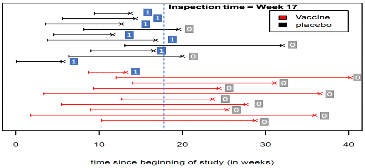
Source: modified from Dr. Marco Carone (BIOST 537 slides)

Illustration of current status data

- Each participant is assessed only **once** at some inspection time
  - observe inspection time C, and whether event has occurred $\Delta_C = 1\{T \leq C\}$
- Coarser outcome information, but potentially cheaper to acquire
- often encountered in cross-sectional surveys or screening programs

Data fusion under mixed censoring types





An example study on dementia onset

- the Alzheimer's Disease Neuroimaging Initiative (ADNI) study
  - designed to emulate an interventional trial cohort
  - assessments every 6 months, longitudinal follow-up spanning 20 years
- the National Alzheimer's Coordinating Center (NACC) dataset
  - larger sample, but assessments at coarser temporal resolution

In infectious diseases: prospective studies + periodic seroprevalence surveys

Data fusion under mixed censoring types

How can we integrate multiple data sources, even those with mixed censoring types, to obtain more precise, more generalizable inference?

- general data integration frameworks for fully observed data (Li and Luedtke, 2023; Graham et al., 2024)
- data integration methods for time-to-event data subject to right censoring (Gao et al., 2024; Wen et al., 2025; Liu et al., 2025)

Potential alternatives using existing methods

Estimands of interest (e.g., survival probability, RMST) can often be estimated using either right-censored or current status data

- combine via inverse variance weighting?

For survival probability, estimator based on current status data behaves differently

- converge at cubic-root-n rate, with a Chernoff limiting distribution (Groeneboom and Wellner, 2001; Wolock et al., 2025)
- asymptotically, weight may converge to 0

Potential alternatives using existing methods

Pool data sources + apply methods for mixed censoring

- require marginal independent censoring (Turnbull, 1976; Huang, 1999; Yu et al., 2001);
- do not account for covariate shift;
- or focus inference on β in Cox or AFT models (Kim, 2003; Zhu et al., 2021; Zhou et al., 2021; Ninga et al., 2025)

Data fusion under mixed censoring types

How can we integrate multiple data sources, even those with mixed censoring types, to obtain more precise, more generalizable inference?

An ideal strategy should be

- efficient (*more data will never hurt and will in general improve*)
- flexible (*not relying on parametric assumptions*).

Notation

Data: $O = (S, W, (1-S)\Delta_C, (1-S)C, S\Delta_R, SY) \sim P$

- $S \in \{0, 1\}$: data source indicator ($S = 1$ for right-censored data; $S = 0$ for current status data)
- W: covariate vector
- T: time to event of interest
- C: inspection time in current status data
- $\Delta_C = 1\{T \leq C\}$
- $Y = \min\{T, R\}$, where R is the censoring time in right-censored data
- $\Delta_R = 1\{T \leq R\}$

Assumptions

Assumption 1 (Exchangeability)

The conditional event time distribution is the same across data sources, that is, $T \perp S \mid W$.

- allowing for covariate shift

Assumption 2 (Conditional independent censoring)

$T \perp C \mid W, S = 0$ *and* $T \perp R \mid W, S = 1$.

- allowing for covariate-dependent censoring

Estimand: target survival probability at a given time point

$$\Phi_{j,t^*}(P) = \mathbb{E}[\mathbb{E}[1\{T > t^*\}|W]|S = j], \quad j \in \{0, 1\}$$

One-step estimation

Given an initial estimator $\hat{P}$ of P and a gradient $D_{\hat{P}}$ of Φ at $\hat{P}$, we construct a one-step estimator

$$\hat{\phi} := \Phi(\hat{P}) + \frac{1}{n}\sum_{i=1}^{n} D_{\hat{P}}(O_i)$$

- corrects the bias of the plug-in estimator
- under conditions, regular and asymptotically linear with influence function D_P
- semiparametrically efficient if D is the canonical gradient (minimal variance)
- often is the empirical mean of the estimated, un-centered gradient

A valid gradient

A valid gradient of Φ_{j,t^*} with respect to the data fusion model at P is

$$\tau_{j,P}(o) = s \int_0^\infty \frac{h^*(u,w) - \mathbb{E}_{T|W}[h^*(T,W) \mid T \geq u, W = w]}{\Gamma(u \mid w)} \, dM_T^{(r)}(u)$$
$$+ \frac{(1-s)\{\delta_C - F_{T|W}(c \mid w)\}}{F_{T|W}(c \mid w)\{1 - F_{T|W}(c \mid w)\}} H^*(c,w)$$
$$+ \frac{1\{s=j\}}{\Pi(S=j)} \left\{ S_{T|W}(t^* \mid w) - \phi_{j,t^*} \right\}.$$

Right-censored contribution

Current-status contribution

Target-population covariate distribution

$$H^*(c,w) = \mathbb{E}[h^*(T,W)1\{T \leq c\} \mid W = w],$$
$$M_T^{(r)}(t) = 1\{y \leq t, \delta_R = 1\} - \int_0^t 1\{y \geq u\} \, d\Lambda_{T|W}(u \mid w).$$

A valid gradient: the function h^*

The function $h^*(t, w)$ refines the usual $1\{t > t^*\}$ term in a right-censored-data influence function by incorporating information from the current-status sample.

For almost every (t, w), it solves

$$
\begin{aligned}
1\{t > t^*\} - S_{T|W}(t^* \mid w) &= \pi \frac{dP_{1,W}(w)}{dP_{j,W}(w)} h^*(t, w) \\
&+ (1-\pi) \frac{dP_{0,W}(w)}{dP_{j,W}(w)} \left[\mathbb{E}\left\{ \frac{1\{C \geq t\} H^*(C, W)}{F_{T|W}(C \mid W)\{1 - F_{T|W}(C \mid W)\}} \,\middle|\, W = w \right\} - \gamma(w) \right],
\end{aligned}
$$

where

$$
\gamma(w) = \mathbb{E}\left[\frac{1\{C \geq T\} H^*(C, W)}{F_{T|W}(C \mid W)\{1 - F_{T|W}(C \mid W)\}} \,\middle|\, W = w \right].
$$

The integral equation determines how information is shared across the two coarsening schemes.

Solving the bridge equation

Fix w, the equation for h^* is a Fredholm equation of the second kind:

$$h^*(t, w) = \frac{dP_{j,W}(w)}{dP_{1,W}(w)} \frac{1\{t > t^*\} - \mu(w)}{\pi} + \int h^*(s, w) K(t, s \mid w)\, ds,$$

and

$$K(t, s \mid w) = -\frac{1-\pi}{\pi} \frac{dP_{0,W}(w)}{dP_{1,W}(w)} f_{T|W}(s \mid w) \int_s^\infty \frac{1\{c \geq t\} - F_{T|W}(c \mid w)}{F_{T|W}(c \mid w)\{1 - F_{T|W}(c \mid w)\}} dG(c \mid w).$$

For each fixed w, solve a one-dimensional Fredholm equation in t.

Practical approaches:

- discretize time on a dense grid, approximate integrals by sums, and solve a linear system;
- approximate $h^*(t, w)$ by basis functions in t, with separate bases on either side of t^*

Doubly robust data-fusion estimator

$$\widehat{\phi}_{j,t^*} = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\underbrace{S_i \int_0^{\infty} \frac{\widehat{h}^*(u, W_i) - \widehat{\mathbb{E}}_{T|W}\left[\widehat{h}^*(T, W) \mid T \geq u, W = W_i\right]}{\widehat{\Gamma}(u \mid W_i)}\, d\widehat{M}_T^{(i)}(u)}_{\text{right-censored contribution}}$$

$$+ \underbrace{\frac{(1 - S_i)\{\Delta_{C,i} - \widehat{F}_{T|W}(C_i \mid W_i)\}}{\widehat{F}_{T|W}(C_i \mid W_i)\{1 - \widehat{F}_{T|W}(C_i \mid W_i)\}} \int_0^{C_i} \widehat{h}^*(u, W_i)\, d\widehat{F}_{T|W}(u \mid W_i)}_{\text{current-status contribution}}$$

$$+ \underbrace{\frac{1\{S_i = j\}}{\Pi(S = j)} \widehat{S}_{T|W}(t^* \mid W_i)}_{\text{target covariate distribution}}\Bigg].$$

$$\widehat{M}_T^{(i)}(t) = 1\{Y_i \leq t, \Delta_{R,i} = 1\} - \int_0^t 1\{Y_i \geq u\}\, d\widehat{\Lambda}_{T|W}(u \mid W_i).$$

Doubly robust data-fusion estimator

To construct the estimator, we estimate:

- $\widehat{F}_{T|W}$, $\widehat{S}_{T|W}$, and $\widehat{\Lambda}_{T|W}$: conditional CDF, survival function, and cumulative hazard of $T \mid W$;
- $\widehat{\Gamma}(\cdot \mid W)$: conditional censoring survival function in the right-censored sample;
- $\widehat{G}(\cdot \mid W)$: conditional distribution of the inspection time in the current-status sample;
- density ratios between the covariate distributions;
- $\widehat{h}^*$: obtained by solving the empirical version of the integral equation.

All nuisance functions can be estimated flexibly, with cross-fitting used for inference.

What does data fusion change?

Right-censored-only AIPCW estimator

$$\widetilde{\phi}^{\mathrm{AIPCW}}_{j,t^*} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{S_i}{\Pi(S=1)}\widehat{\frac{dP_{j,W}}{dP_{1,W}}}(W_i)\int_0^\infty \frac{1\{u>t^*\} - \widehat{\mathbb{E}}_{T|W}[1\{T>t^*\} \mid T\geq u, W=W_i]}{\widehat{\Gamma}(u \mid W_i)}\, d\widehat{M}_T^{(i)}(u) + \frac{1\{S_i=j\}}{\Pi(S=j)}\widehat{S}_{T|W}(t^* \mid W_i)\right].$$

Data-fusion estimator

$$\widehat{\phi}_{j,t^*} = \frac{1}{n}\sum_{i=1}^{n}\left[S_i\int_0^\infty \frac{\widehat{h}^*(u,W_i) - \widehat{\mathbb{E}}_{T|W}\left[\widehat{h}^*(T,W) \mid T\geq u, W=W_i\right]}{\widehat{\Gamma}(u \mid W_i)}\, d\widehat{M}_T^{(i)}(u) + \frac{(1-S_i)\{\Delta_{C,i} - \widehat{F}_{T|W}(C_i \mid W_i)\}}{\widehat{F}_{T|W}(C_i \mid W_i)\{1-\widehat{F}_{T|W}(C_i \mid W_i)\}}\int_0^{C_i}\widehat{h}^*(u,W_i)\, d\widehat{F}_{T|W}(u \mid W_i) + \frac{1\{S_i=j\}}{\Pi(S=j)}\widehat{S}_{T|W}(t^* \mid W_i)\right].$$

$1\{u>t^*\}$ is replaced by $\widehat{h}^*(u,W)$ and current-status outcomes contribute directly.

Asymptotic normality of the doubly robust estimator

Suppose the support of the inspection time satisfies $\mathcal{C} \subseteq [c_l, c_u]$.

Theorem 1 (Asymptotic behavior of $\widehat{\phi}_{j,t^*}$)

Under Assumptions 1–2 and other regularity conditions, suppose nuisance estimators are obtained by cross-fitting (or belong to a fixed P-Donsker class), $\|\tau_{j,\widehat{P}} - \tau_{j,P}\|_{L_2(P)} = o_P(1)$, and $\Gamma\{\max(c_u, t^*) \mid W\} > \epsilon$ with probability one for some constant $\epsilon > 0$. Further suppose

1. $\sup_{c \in [c_l, c_u]} \|\widehat{F}_{T|W}(c \mid \cdot) - F_{T|W}(c \mid \cdot)\|_{L_2(P_W)} \|\widehat{g}(c \mid \cdot) - g(c \mid \cdot)\|_{L_2(P_W)} = o_P(n^{-1/2})$;
2. $$\sup_{u \leq \max(c_u, t^*)} \left\| \widehat{\mathbb{E}}_{T|W}[\widehat{h}^*(T, W) \mid T \geq u, W = \cdot] - \mathbb{E}_{T|W}[\widehat{h}^*(T, W) \mid T \geq u, W = \cdot] \right\|_{L_2(P_W)}$$ $$\times \|\widehat{\lambda}_{R|W}(u \mid \cdot) - \lambda_{R|W}(u \mid \cdot)\|_{L_2(P_W)} = o_P(n^{-1/2});$$
3. $\sup_{c \in [c_l, c_u]} \|\widehat{F}_{T|W}(c \mid \cdot) - F_{T|W}(c \mid \cdot)\|_{L_2(P_W)} \|\widehat{\rho}(\cdot) - \rho(\cdot)\|_{L_2(P_W)} = o_P(n^{-1/2})$.

Then,

$$\sqrt{n}\left(\widehat{\phi}_{j,t^*} - \phi_{j,t^*}\right) \rightsquigarrow N(0, \mathrm{var}\{\tau_{j,P}(O)\}).$$

The nuisance errors enter through products; each nuisance estimator may converge more slowly than $n^{-1/2}$.

Efficient influence function

The gradient $\tau_{j,P}$ is valid, but generally not the canonical gradient.

- Exchangeability requires the same perturbation of $T \mid W$ in both data sources.
- Hence, the tangent space is smaller than in an unconstrained observed-data model.
- $\tau_{j,P}$ generally does not lie in this restricted tangent space.

Efficient influence function

The canonical gradient of Φ_{j,t^*} is

$$\tau^{\text{eff}}_{j,P}(o) = s\left\{\delta_R \eta^*(y,w) - \int_0^y \eta^*(u,w)\lambda_{T|W}(u \mid w)\,du\right\}$$
$$+ (1-s)\left\{\frac{\delta_C - F_{T|W}(c \mid w)}{F_{T|W}(c \mid w)}\right\}\Theta^*(c,w)$$
$$+ \frac{1\{s=j\}}{\Pi(S=j)}\left\{S_{T|W}(t^* \mid w) - \phi_{j,t^*}\right\},$$

where

$$\Theta^*(c,w) = \int_0^c \eta^*(u,w)\lambda_{T|W}(u \mid w)\,du.$$

For almost every (t,w), η^* solves

$$0 = \pi \frac{dP_{1,W}(w)}{dP_{j,W}(w)}\eta^*(t,w)\Gamma(t \mid w)S_{T|W}(t \mid w) + S_{T|W}(t^* \mid w)1\{t \le t^*\}$$
$$+ (1-\pi)\frac{dP_{0,W}(w)}{dP_{j,W}(w)}\int_t^\infty \frac{S_{T|W}(c \mid w)\Theta^*(c,w)}{F_{T|W}(c \mid w)} g(c \mid w)\,dc.$$

The censoring survival $\Gamma(t \mid w)$ now enters the integral equation itself, determining the optimal allocation of information across sources.

Efficient estimator and its asymptotic behavior

Let $\widehat{\phi}^{\text{eff}}_{j,t^*}$ denote the (cross-fitted) one-step estimator based on $\tau^{\text{eff}}_{j,\widehat{P}}$.

Theorem 2 (Asymptotic behavior of $\widehat{\phi}^{\text{eff}}_{j,t^*}$)

Under Assumptions 1–2 and regularity conditions, suppose nuisance estimators are obtained by cross-fitting (or belong to a fixed P-Donsker class), and $\Gamma\{\max(c_u, t^*) \mid W\} > \epsilon$ *with probability one for some* $\epsilon > 0$. *A representative rate condition is*

$$\sup_{u \le \max(c_u, t^*)} \|\widehat{\lambda}_{T|W}(u \mid \cdot) - \lambda_{T|W}(u \mid \cdot)\|_{L_2(P_W)} \times \Bigg[\sup_{u \le \max(c_u, t^*)} \|\widehat{\lambda}_{T|W}(u \mid \cdot) - \lambda_{T|W}(u \mid \cdot)\|_{L_2(P_W)}$$
$$+ \sup_{\substack{u' \le \max(c_u, t^*) \\ c \in [c_l, c_u]}} \left\{ \|\widehat{g}(c \mid \cdot) - g(c \mid \cdot)\|_{L_2(P_W)} + \|\widehat{\lambda}_{R|W}(u' \mid \cdot) - \lambda_{R|W}(u' \mid \cdot)\|_{L_2(P_W)} + \|\widehat{\rho} - \rho\|_{L_2(P_W)} \right\} \Bigg]$$
$$= o_P(n^{-1/2}).$$

Then,

$$\sqrt{n}\left(\widehat{\phi}^{\text{eff}}_{j,t^*} - \phi_{j,t^*}\right) \rightsquigarrow N\Big(0, \text{var}\{\tau^{\text{eff}}_{j,P}(O)\}\Big).$$

Because $\tau^{\text{eff}}_{j,P}$ is the canonical gradient, $\widehat{\phi}^{\text{eff}}_{j,t^*}$ attains the semiparametric efficiency bound.

Simulation: setup

We aim to estimate the survival probabilities at time points $t^* \in \{0.2, 0.7, 0.9\}$ in the current status population and compare four estimators:

- current status only estimator (CS) (Wolock et al., 2025)
- right-censored only AIPCW estimator (adjusting for covariate shift)
- data fusion estimator
- efficient data fusion estimator

under the following data generating mechanism

- $W_1 \sim \mathrm{Beta}(1, 1 + (1 - S)/4)$, $W_2 \sim \mathrm{Bernoulli}(0.5 - 0.1(1 - S))$.
- $T \sim \mathrm{Exp}(0.8 + 0.4W_1 + 0.2W_1W_2)$
- $R \sim \mathrm{Exp}(1.5 - 0.2W_1 - 0.5W_2)$
- $C \sim 0.5\mathrm{Beta}(1, 0.75 + 0.5W_1 + 0.1W_2) + 0.5$

Simulation: results

$n = 300,\ t^* = 0.2$				$n = 1500,\ t^* = 0.9$			
Method	Est	CI length	Coverage	Method	Est	CI length	Coverage
CS	·	·	·	CS	0.385	0.157	0.942
AIPCW	0.810	0.186	0.954	AIPCW	0.395	0.126	0.929
Fusion	0.808	0.175	0.952	Fusion	0.393	0.080	0.945
Efficient Fusion	0.810	0.174	0.962	Efficient Fusion	0.394	0.076	0.954

The sample size ratio of CS : RC is 2:1.

Naive inverse-variance weighting yields CI length of 0.098.

Simulation: results

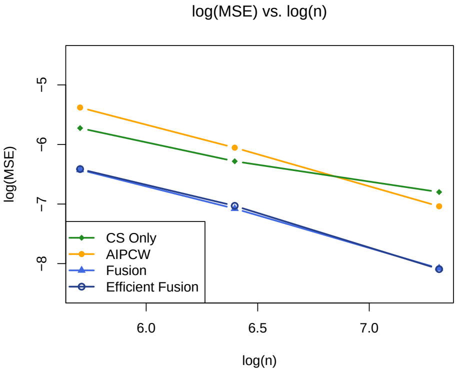


Figure: log(MSE) vs. log(N) when $t^* = 0.7$ for current status only estimator (CS only), right censored only AIPCW estimator, and the proposed data fusion estimators (Fusion and Efficient Fusion).

- **Steeper slope**: fusion estimator achieves root-n rate.
- **Lower intercept**: fusion estimator has smaller asymptotic variance.

Data fusion both recovers parametric-rate inference and improves precision.

Data application: dementia onset

- Dementia affects tens of millions of people worldwide, with prevalence projected to rise substantially as populations age (World Health Organization, 2017).
- A key question: how quickly do individuals with normal cognition or mild cognitive impairment (MCI) progress to dementia?
- We study adults aged 55 years or older with normal cognition or MCI at baseline by integrating:
  - **ADNI**: a smaller cohort with visits approximately every 6 months;
  - **NACC**: a larger cohort with coarser (annual) assessments.
- Target population: the ADNI cohort; NACC contributes complementary information after accounting for covariate shift.
- For the analysis, we randomly select one NACC visit per participant, treating it as current status data.

Data application: dementia onset

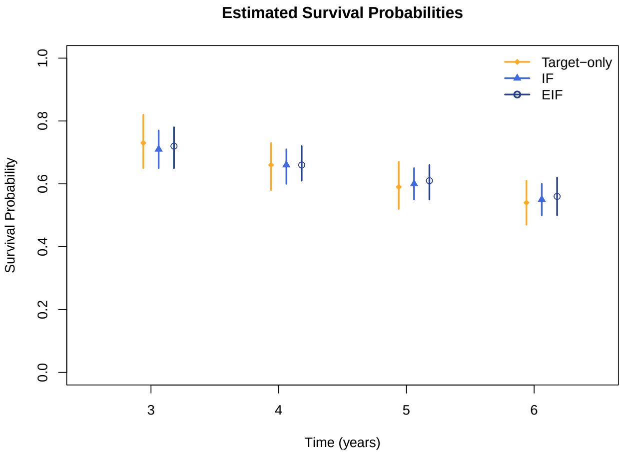


Figure: Estimated survival probabilities at $t^* = 3, 4, 5, 6$ years with 95% CI for AIPCW estimator that only uses ADNI data (Target-only), and the proposed data fusion estimators (IF and EIF).

Extension: valid data fusion gradient for a general estimand

Suppose the full-data gradient admits the decomposition

$$\tau^{\mathrm{full}}(t,w) = \tau_1(t,w) + \tau_2(w), \qquad \mathbb{E}[\tau_1(T,W) \mid W = w] = 0.$$

- Restricted mean survival time (RMST), assumption-lean Cox regression (Vansteelandt et al., 2024)

Then a valid gradient under the data-fusion model is

$$\begin{aligned}\tau_P(o) = {} & s\int_0^\infty \frac{h^*(u,w) - \mathbb{E}_{T|W}[h^*(T,W) \mid T \geq u, W = w]}{\Gamma(u \mid w)}\, dM_T^{(r)}(u) \\ & + \frac{(1-s)\{\delta_C - F_{T|W}(c \mid w)\}}{F_{T|W}(c \mid w)\{1 - F_{T|W}(c \mid w)\}} H^*(c,w) \\ & + \frac{1\{s = j\}}{\Pi(S = j)} \tau_2(w),\end{aligned}$$

Extension: valid data fusion gradient for a general estimand

Suppose the full-data gradient admits the decomposition

$$\tau^{\text{full}}(t, w) = \tau_1(t, w) + \tau_2(w), \qquad \mathbb{E}[\tau_1(T, W) \mid W = w] = 0.$$

For almost every (t, w), h^* solves

$$\tau_1(t, w) = \pi \frac{dP_{1,W}(w)}{dP_{j,W}(w)} h^*(t, w)$$
$$+(1-\pi) \frac{dP_{0,W}(w)}{dP_{j,W}(w)} \left[\mathbb{E}\left\{ \frac{1\{C \geq t\} H^*(C, W)}{F_{T|W}(C \mid W)\{1 - F_{T|W}(C \mid W)\}} \,\middle|\, W = w \right\} - \gamma(w) \right],$$

Extension: general interval censoring

Let $\mathbf{V} = (V_1, \ldots, V_K)$ denote inspection times, with $V_0 = 0$, $V_{K+1} = \infty$, and let J indicate the interval $(V_{J-1}, V_J]$ containing T.

A valid observed-data gradient is

$$\begin{aligned}\tau_P(o) &= s \int_0^{\infty} \frac{h^*(u,w) - \mathbb{E}_{T|W}[h^*(T,W) \mid T \geq u, W = w]}{\Gamma(u \mid w)} \, dM_T^{(r)}(u) \\ &\quad + (1-s) \sum_{k=1}^{K+1} \frac{1\{J = k\}}{F_{T|W}(V_k \mid w) - F_{T|W}(V_{k-1} \mid w)} H^*(V_{k-1}, V_k, w) \\ &\quad + \frac{1\{s = j\}}{\Pi(S = j)} \tau_2(w).\end{aligned}$$

Current-status data arise as a special case with one inspection time.

Extension: general interval censoring

Let $\mathbf{V} = (V_1, \ldots, V_K)$ denote inspection times, with $V_0 = 0$, $V_{K+1} = \infty$, and let J indicate the interval $(V_{J-1}, V_J]$ containing T.

For general interval censored data, we obtain h^* by solving

$$\tau_1(t, w) = \pi \frac{dP_{1,W}(w)}{dP_{j,W}(w)} h^*(t, w)$$

$$+(1-\pi)\frac{dP_{0,W}(w)}{dP_{j,W}(w)} \left\{ \int \sum_{k=1}^{K+1} \frac{1\{v_{k-1} < t \leq v_k\} H^*(v_{k-1}, v_k, w)}{F_{T|W}(v_k \mid w) - F_{T|W}(v_{k-1} \mid w)} g(\mathbf{v} \mid w)\, d\mathbf{v} - \gamma(w) \right\},$$

where

$$H^*(v, \tilde{v}, w) = \int_v^{\tilde{v}} f_{T|W}(u \mid w) h^*(u, w)\, du.$$

Summary

- We develop a semiparametric data-fusion framework for right-censored and current-status data under covariate shift.
- Current-status data can provide first-order efficiency gains, despite sometimes being insufficient for root-n estimation on their own.
- The framework yields doubly robust and semiparametrically efficient estimators, and extends to general interval censoring.
- Under a limited follow-up or assessment budget, how should we adaptively choose inspection times or visit schedules to maximize efficiency for a target survival estimand?

Thank you for your time!

https://arxiv.org/abs/2508.10357

References I

Gao, C., Yang, S., Shan, M., Ye, W. W., Lipkovich, I., and Faries, D. (2024). Doubly protected estimation for survival outcomes utilizing external controls for randomized clinical trials. *arXiv preprint arXiv:2410.18409*.

Graham, E., Carone, M., and Rotnitzky, A. (2024). Towards a unified theory for semiparametric data fusion with individual-level data. *arXiv preprint arXiv:2409.09973*.

Groeneboom, P. and Wellner, J. A. (2001). Computing chernoff's distribution. *Journal of Computational and Graphical Statistics*, 10(2):388–400.

Huang, J. (1999). Asymptotic properties of nonparametric estimation based on partly interval-censored data. *Statistica Sinica*, pages 501–519.

Kim, J. S. (2003). Maximum likelihood estimation for the proportional hazards model with partly interval-censored data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(2):489–502.

Li, S. and Luedtke, A. (2023). Efficient estimation under data fusion. *Biometrika*, 110(4):1041–1054.

References II

Liu, Y., Levis, A. W., Zhu, K., Yang, S., Gilbert, P. B., and Han, L. (2025). Targeted data fusion for causal survival analysis under distribution shift. *arXiv preprint arXiv:2501.18798*.

Ninga, X., Sun, Y., Pan, Y., and Gilbert, P. B. (2025). Regression analysis of semiparametric cox-aalen transformation models with partly interval-censored data. *Electronic journal of statistics*, 19(1):240.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(3):290–295.

Vansteelandt, S., Dukes, O., Van Lancker, K., and Martinussen, T. (2024). Assumption-lean cox regression. *Journal of the American Statistical Association*, 119(545):475–484.

Wen, L., Steingrimsson, J. A., Robertson, S. E., and Dahabreh, I. J. (2025). Multi-source analyses of average treatment effects with failure time outcomes: L. wen et al. *Lifetime Data Analysis*, pages 1–29.

References III

Wolock, C. J., Jacob, S., Bennett, J. C., Elias-Warren, A., O'Hanlon, J., Kenny, A., Jewell, N. P., Rotnitzky, A., Cole, S. R., Weil, A. A., et al. (2025). Investigating symptom duration using current status data: A case study of postacute covid-19 syndrome. *Epidemiology*, pages 10–1097.

World Health Organization (2017). Global action plan on the public health response to dementia 2017–2025. In *Global action plan on the public health response to dementia 2017–2025*.

Yu, Q., Wong, G. Y., and Li, L. (2001). Asymptotic properties of self-consistent estimators with mixed interval-censored data. *Annals of the Institute of Statistical Mathematics*, 53(3):469–486.

Zhou, Q., Sun, Y., and Gilbert, P. B. (2021). Semiparametric regression analysis of partly interval-censored failure time data with application to an aids clinical trial. *Statistics in medicine*, 40(20):4376–4394.

Zhu, L., Tong, X., Cai, D., Li, Y., Sun, R., Srivastava, D. K., and Hudson, M. M. (2021). Maximum likelihood estimation for the proportional odds model with mixed interval-censored failure time data. *Journal of applied statistics*, 48(8):1496–1512.

Why is $\mathcal{P}$ semiparametric?

Under Assumptions 1-4, the likelihood function is given by

$$L(x) = p_W(w)\left\{\pi\Gamma(y|w)S_{T|W}(y|w)\lambda_{R|W}(y|w)^{1-\delta_R}\lambda_{T|W}(y|w)^{\delta_R}\right\}^{s}\times$$
$$\left\{(1-\pi)g(c|w)F_{T|W}(c|w)^{\delta_C}S_{T|W}(c|w)^{1-\delta_C}\right\}^{1-s},$$

where $\pi = P(S=1)$, $\Gamma(t|w) = P_{R|W}(R \geq t|W=w)$, $g(c \mid w)$ is the conditional density of inspection time given $W=w$,
$S_{T|W}(t|w) = 1 - F_{T|W}(t \mid w) = P_{T|W}(T \geq t|W=w)$, $\lambda_{R|W}$ and $\lambda_{T|W}$ denote the hazard function T and R given W respectively.

More simulation results

	$t^* = 0.2$			$t^* = 0.7$			$t^* = 0.9$		
	Est	CI length	Coverage	Est	CI length	Coverage	Est	CI length	Coverage
$n = 300$									
CS	.	.	.	0.484	0.197	0.920	0.361	0.287	0.892
AIPCW	0.810	0.186	0.954	0.484	0.266	0.954	0.394	0.285	0.937
Fusion	0.808	0.175	0.952	0.480	0.158	0.965	0.387	0.185	0.939
Efficient Fusion	0.810	0.174	0.962	0.486	0.158	0.967	0.394	0.168	0.970
$n = 600$									
CS	.	.	.	0.482	0.158	0.936	0.376	0.221	0.940
AIPCW	0.813	0.122	0.939	0.483	0.190	0.912	0.393	0.199	0.927
Fusion	0.811	0.120	0.949	0.481	0.114	0.951	0.390	0.128	0.932
Efficient Fusion	0.812	0.119	0.954	0.483	0.117	0.934	0.392	0.124	0.964
$n = 1500$									
CS	.	.	.	0.483	0.116	0.922	0.385	0.157	0.942
AIPCW	0.811	0.079	0.954	0.482	0.116	0.943	0.395	0.126	0.929
Fusion	0.810	0.073	0.945	0.481	0.069	0.943	0.393	0.080	0.945
Efficient Fusion	0.810	0.074	0.945	0.482	0.069	0.945	0.394	0.076	0.954

CS is unavailable at $t^* = 0.2$ because this time point lies outside the support of the inspection times.