Causal Inference with Survival Outcomes

Fan Li

Department of Statistical Science
Duke University

Survival/Time-to-Event Outcomes

(Part of the slides is written by Laine Thomas)

- Survival outcomes are common in comparative effectiveness research
- Often incompletely observed due to right-censoring, and require unique handling
- Causal inference: combine standard survival estimators with propensity score weighting
- Standard (but questionable) practice: weighted Cox regression to estimate the causal hazard ratio (HR)
  - Assumes proportional hazards in the target population
  - Coefficient of treatment variable usually does not have a causal interpretation (Hernán, 2010)
- Other estimands besides HR, e.g. survival curve, restricted mean survival time (Mao et al. 2018)

Notation

- Consider a sample of N units drawn from a population, for each unit
  - $Z_i \in \{0, 1\}$: treatment
  - $\mathbf{X}_i$: pre-treatment covariates
  - $\{T_i\}$: failure time
  - $\{C_i\}$: censoring time
  - $Y_i = T_i \wedge C_i$: observed failure time
  - $\delta_i = \mathbf{1}\{T_i \leq C_i\}$: observed censoring indicator
- Observed data for each unit $O_i = (Z_i, \mathbf{X}_i, Y_i, \delta_i)$

Cox proportional hazards regression

- Suppose we have a randomized trial
- We often model the hazard of the true outcome T as

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta \cdot Z)$$

  where:
  - $\lambda(t|Z)$ is the hazard rate of failure time T at time t.
  - $\lambda_0(t)$ is the baseline hazard rate.
  - $\exp(\beta)$ is the hazard ratio associated with Z.
- With a randomized trial there are no covariates in this model
- A key implication: The hazard ratio is constant over time (proportional hazards)

Causation of hazard ratio?

- Interpretation of β (Matinussen et al. 2020)

$$\begin{aligned}\exp(\beta) &= \frac{\lim_{h\to 0}\Pr(t \le T \le t+h \mid T \ge t, Z=1)}{\lim_{h\to 0}\Pr(t \le T \le t+h \mid T \ge t, Z=0)} \\ &= \frac{\lim_{h\to 0}\Pr(t \le T(1) \le t+h \mid T(1) \ge t)}{\lim_{h\to 0}\Pr(t \le T(0) \le t+h \mid T(0) \ge t)}\end{aligned}$$

- The second equation holds only in randomized trials
- Causal interpretation of β?
  - The probabilities are conditional on two different, post-treatment risk sets based on $\{i : T_i(1) \ge t\}$ in the numerator (treatment group) and $\{i : T_i(0) \ge t\}$ in the denominator (control group)
  - The risk sets are the same only under null, i.e. treatment effect is zero throughout
  - When there is non-zero treatment effect, hazard ratio (i.e. $\exp(\beta)$) is NOT a causal estimand, because it compares two different groups

Causation of hazard ratio?

- Causal interpretation of β is ill-defined except under the null
- The hazard ratio has built-in post-treatment selection bias (Hernán, 2010)
  - The hazard at time (t) is defined among a subset of the starting population: those that remain at risk
  - Increasingly over time, people who experience an event, drop out of the risk set
  - Depletion of susceptibles
  - Depletion happens differentially by treatment creating bias in the comparison of hazards by treatment
  - Proportional hazards will not hold (with enough data or time)
- The causal interpretation of the hazard ratio is problematic even in a randomized study!

Hazard ratio: For and Against

- Clinical Trials: The hazard ratio remains widely used to estimate treatment effects
  - Any single summary measure will lose temporal information
  - The hazard ratio represents a type of average over time, even though the contributing increments are increasingly non-causal
  - Value in comparison to prior results
- Observational comparisons: The hazard ratio is widely used in practice and widely discouraged in methodological publications
  - Hazard ratios are often desired when we emulate a trial or prior study
  - Causal inference methodology for observational data involves rigorous specification of the causal estimand. Attention to the causal estimand excludes the hazard ratio.
  - Better alternatives: cumulative probabilities (survival curves), mean survival time and restricted mean survival time

Estimating the Marginal Hazard Ratio

- The target of inference is $\exp(\beta_{\text{MHR}})$ in a misspecified model

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta_{\text{MHR}} \cdot Z)$$

  where MHR stands for "marginal hazard ratio" (Austin, 2013)

- The underlying outcome process is expected to involve more covariates:

$$\lambda(t|Z, \mathbf{X}) = \lambda_0(t)\exp(\beta_Z \cdot Z + f(\beta_{\mathbf{X}} \cdot \mathbf{X}))$$

- An even more complex outcome model is likely

- The target of inference: The marginal hazard ratio that we would obtain if we had a gigantic, randomized clinical trial

Cox Regression with weighting

- Derive weights exactly as before, by modeling $P(Z = 1|\mathbf{X})$ and constructing IPW, OW, or any balancing weight
- Fit the Cox proportion hazard model to the weighted data with Z as the only independent variable
- The estimated parameter for Z is $\hat{\beta}_{\text{MHR}}$
- This has been shown by simulation to be unbiased for β_{MHR}, with the limitations that are inherent to β_{MHR}
- Make sure to obtain the variance by bootstrapping, or sandwich variance

(Austin, 2013), (Mao et al. 2018)

Notation: Potential Outcomes

- Cultural difference between causal inference and survival analysis
  - Causal inference: starting from causal estimands (not model)
  - Survival analysis: starting from an outcome model (e.g. Cox model)
- For causal inference, we need potential (counterfactual) outcomes (assuming SUTVA)
  - $\{T_i(1), T_i(0)\}$: potential failure times
  - $\{C_i(1), C_i(0)\}$: potential censoring times
- Under SUTVA, the relation between counterfactual and factual data:

$$T_i = Z_i T_i(1) + (1 - Z_i) T_i(0)$$

$$C_i = Z_i C_i(1) + (1 - Z_i) C_i(0)$$

- Causal estimands are functions of the potential outcomes

Causal Estimands in Scale of Survival Function

- For simplicity, we will first focus on the sampled target population (ATE)
- Counterfactual survival functions for $z = 0, 1$
  - $S^{(z)}(t|\mathbf{X}) = \Pr(T(z) \geq t|\mathbf{X})$
  - $S^{(z)}(t) = \Pr(T(z) \geq t) = \mathbb{E}_X[S^{(z)}(t|\mathbf{X})]$
- Survival probability causal effect (SPCE) at t (Mao et al. 2018):

$$\Delta^{SPCE}(t) = S^{(1)}(t) - S^{(0)}(t)$$

  - Comparison of two survival probabilities: $\tau(t^*)$, at time t^*, OR
  - Comparison of survival curves (KM) over time $0 \leq t \leq t^{max}$ where t^{max} is the maximum available follow-up

Causal Estimands in Scale of Survival Function

- Average causal effect (ACE):

$$\begin{aligned}\Delta^{ACE} &= \int_0^\infty S^{(1)}(t)dt - \int_0^\infty S^{(0)}(t)dt \\ &= \int_0^\infty \Delta^{SPCE}(t)dt \\ &= \mathbb{E}[T(1)] - \mathbb{E}[T(0)]\end{aligned}$$

  - Integral of the SPCE
  - Difference in mean survival times when the entire population is treated versus untreated
  - Rarely identifiable due to right censoring

Causal Estimands in Scale of Survival Function

- Restricted average causal effect (RACE):

$$\begin{aligned}\Delta^{RACE} &= \int_0^{t^*} S^{(1)}(t)dt - \int_0^{t^*} S^{(0)}(t)dt \\ &= \int_0^{t^*} \Delta^{SPCE}(t)dt \\ &= \mathbb{E}[\min\{T(1), t^*\}] - \mathbb{E}[\min\{T(0), t^*\}]\end{aligned}$$

  - Difference in restricted mean survival times (RMST) when the entire population is treated versus untreated
  - Interpretation: t^* life expectancy, e.g. the average event time over the next $t^* = 5$ years would be 6 months greater if everyone were treated versus no one were treated

Summary of Survival Causal Estimands

- Note that each of the causal estimands is defined on the entire population that could initiate either treatment (nothing is being conditioned on)

$$\begin{aligned}\Delta^{SPCE}(t) &= \Pr(T(1) \geq t) - \Pr(T(0) \geq t)\\ \Delta^{ACE} &= \mathbb{E}[T(1)] - \mathbb{E}[T(0)]\\ \Delta^{RACE} &= \mathbb{E}[\min\{T(1), t^*\}] - \mathbb{E}[\min\{T(0), t^*\}]\end{aligned}$$

- These extend to any weighted population by denoting that expectations be taken with respect to the weighted distribution (Mao et al. 2018)

$$\begin{aligned}\Delta_w^{SPCE}(t) &= \Pr_w(T(1) \geq t) - \Pr_w(T(0) \geq t)\\ \Delta_w^{ACE} &= \mathbb{E}_w[T(1)] - \mathbb{E}_w[T(0)]\\ \Delta_w^{RACE} &= \mathbb{E}_w[\min\{T(1), t^*\}] - \mathbb{E}_w[\min\{T(0), t^*\}]\end{aligned}$$

Assumptions for estimation

- (A1) Unconfoundedness: $\{T(1), T(0)\} \perp Z|\mathbf{X}$
- (A2) Overlap: $0 < e(\mathbf{X}) = \Pr(Z = 1|\mathbf{X}) < 1$
- (A3) Independent/non-informative censoring:

$$T(z) \perp C(z)$$

  - (A3) Will be relaxed to allow for conditionally independent censoring $T(z) \perp C(z)|\mathbf{X}_C, Z = z$ and further to allow for dependent censoring conditional on measured time-varying covariates.

Independent/Non-informative censoring

- Simplified graphic with only 1 time point (generalizes)
- Analysis is restricted to those who are uncensored ($\delta = 0$)
- Here, there is censoring, but it occurs either totally at random or at random within a treatment group (nothing else involved on the DAG)
- Examples: Administrative censoring, staggered study entry, arbitrary reasons for drop-out

Weighted Estimators

- $\hat{\Delta}_w^{SPCE}(t)$ is obtained by weighted Kaplan-Meier methods using any weight that balances covariates (IPW, ATT, OW)

  (Cole and Hernan 2004, Xie and Liu 2005, Conner et al. 2019)

$$\widehat{S}^{(1)}(t) = 1 - \frac{\sum_{i=1}^{n} Z_i w(X_i) \mathbf{1}(Y_i \leq t) \delta_i \Big/ \widehat{Pr}(C_i \geq Y_i | Z_i)}{\sum_{i=1}^{n} Z_i w(X_i)}$$

  where $w(X_i)$ are balancing weights (Cheng et al 2022). Repeat for $\widehat{S}^{(0)}(t)$.

- $\hat{\Delta}^{RACE} = \int_0^{t^*} \hat{\Delta}_w^{SPCE}(t) dt$

  (Conner et al. 2019)

- These are illustrated in the lab section

Alternative Assumption A3

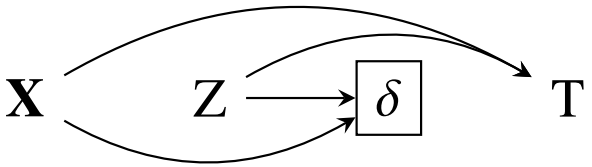
- (A1) Unconfoundedness: $\{T(1), T(0)\} \perp Z|\mathbf{X}$
- (A2) Overlap: $0 < e(\mathbf{X}) = \Pr(Z = 1|\mathbf{X}) < 1$
- (A3) Conditionally independent censoring:

$$T(z) \perp C(z)|\mathbf{X}_C, Z = z$$

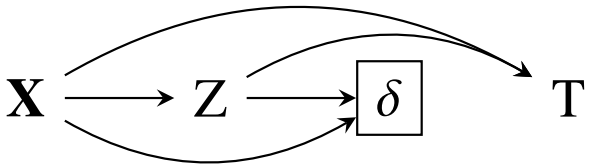
  - Conditional on measured covariates $\mathbf{X}_C$
  - $\mathbf{X}_C$ is not generally the same as $\mathbf{X}$! They can be the same, but it depends on the causal relationships (DAG)
  - Requires new methodology to incorporate censoring weights
  - How to identify $\mathbf{X}_C$?

Conditionally independent censoring

Randomized: Censoring but no confounding



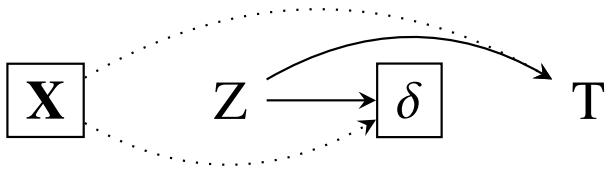
Non-randomized: Confounding and Censoring



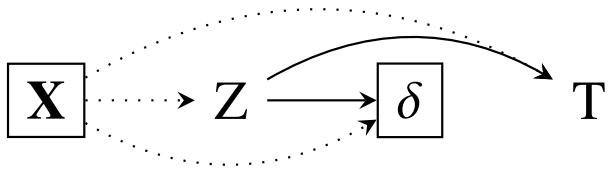
- Heuristically $\mathbf{X}_C$ are common causes of censoring and outcome (more technical definition Chapter 8 Hernan and Robins; What if?)
- Here $\mathbf{X} = \mathbf{X}_C$. The variables involved in confounding are the same as the variables involved in censoring.
- Here $\mathbf{X} = \mathbf{X}_C$ arise and are measured pre-treatment
- Examples: Age, sex, SES, access to transportation ...

Conditionally independent censoring

Randomized: Censoring but no confounding



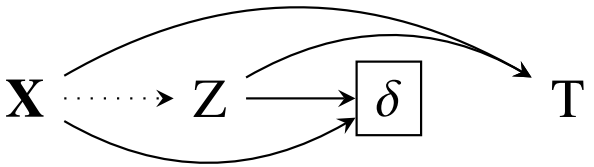
Non-randomized: Confounding and Censoring



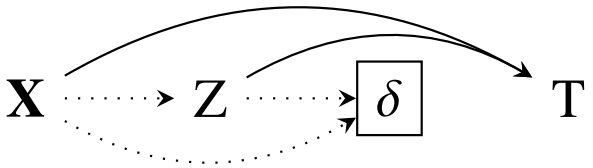
- Conditioning on **X** takes care of confounding and bias due to censoring. Remember, conditioning on **X** means conducting the analysis within fixed levels of **X** (stratification, regression).
- Weighting works a little bit differently but involves the same **X**

Conditionally independent censoring

IPTW psuedo-randomizes Z



IPTW + IPCW
psuedo-randomizes δ or C



- If we are using weighting (IPTW) to address confounding we need to add a second weight (IPCW) to address bias due to censoring
- Even though IPTW and IPCW require the same **X** variables we still need two sets of weights. IPTW is not enough!

Identification: IPTW+ IPCW

(Cheng et al, 2022, AJE)

- Define $K_c^{(z)}(t, \mathbf{X}) = \Pr(C(z) \geq t|X) = \Pr(C \geq t|\mathbf{X}, Z = z)$ as the conditional survival function of censoring time for $z = 0, 1$
- Identification: inverse probability of treatment weights (IPTW) + inverse probability of censoring weights (IPCW) given at observed failure times

$$S^{(1)}(t) = 1 - E\left[\frac{Z_i \delta_i \mathbf{1}\{Y_i \leq t\}}{e(\mathbf{X}_i) K_c^{(1)}(Y_i, \mathbf{X}_i)}\right] \Big/ E\left[\frac{Z_i}{e(\mathbf{X}_i)}\right]$$

  - essentially 1 minus the weighted CDF
  - Identification for $S^{(0)}(t)$ is analogous, replacing Z by $1 - Z$ and e by $(1 - e)$
- $K_c^{(z)}(t, X) = \Pr(C \geq t|\mathbf{X}, Z = z)$ is estimated by parametric survival models (e.g. Weibull) or Cox models
- $e(\mathbf{X})$ is estimated as usual

Estimation: IPTW + IPCW

(Cheng et al, 2022, AJE)

- Estimation: plug the empirical estimates of $e(\mathbf{X})$ and $K_c^{(z)}(t, \mathbf{X})$ into the identification formula

$$\hat{S}^{(1)}(t) = 1 - \frac{\sum_{i=1}^{n} Z_i w(\mathbf{X}_i) \mathbf{1}(Y_i \leq t) \delta_i / \hat{K}_c^1(Y_i, \mathbf{X}_i)}{\sum_{i=1}^{n} Z_i w(\mathbf{X}_i)}$$

- $K_c^{(z)}(t, \mathbf{X}) = \Pr(C \geq t | \mathbf{X}, Z = z)$ differs from the traditional IPTW weighted KM curve via the inclusion of $\mathbf{X}$ in the censoring weight
- Estimator is consistent and asymptotically normal estimators for the target estimands for any balancing weights
- Closed-form sandwich variance estimators are available (via M-estimation), or bootstrap can be used

Extention to other balancing weights

- Recall that IPTW (or IPW) is a special case of the family of balancing weights
- The target population of IPTW is the population represented by the study sample
- Other target population might be more appropriate, e.g. the overlap population (OW) or the treated population (ATT)
- We can use the balancing weight framework to generalized IPW to target population (weighting)
- Denote the sample population by $f(x)$, target population by $g(x)$, then $h(X) \propto g(X)/f(X)$ is a tilting function
- We can re-weight the observed sample to represent the target population g (equivalent h)

$$\begin{cases} w_1(x) \propto \frac{h(x)}{e(x)}, \\ w_0(x) \propto \frac{h(x)}{1-e(x)}. \end{cases}$$

Estimands on target populations

- Conditional counterfactual survival function: $S^{(z)}(t|X) = \Pr(T(z) \geq t|\mathbf{X})$, for $z = 0, 1$
- Survival probability causal effect (SPCE) at t on target population (denote by tilting function h):

$$\tau^h(t) = \frac{\int \{S^{(1)}(t|\mathbf{X}) - S^{(0)}(t|\mathbf{X})\} f(\mathbf{X}) h(\mathbf{X}) \mu(d\mathbf{X})}{\int f(\mathbf{X}) h(\mathbf{X}) \mu(d\mathbf{X})}$$
$$= S_h^{(1)}(t) - S_h^{(0)}(t)$$

- $\Pr(T(z) \geq t \mid g) = S_h^{(z)}(t)$: counterfactual survival function in target population g (equivalently h)

Identification and Estimation

(Cheng et al, 2022)

- Identification and Estimation of $S_h^{(z)}(t)$ remains the same, just add $h(x)$ in the definition of $w(\mathbf{X}_i)$

$$\hat{S}_h^{(1)}(t) = \frac{\sum_{i=1}^N Z_i w(\mathbf{X}_i) \mathbf{1}(Y_i \geq t) \delta_i / K_c^1(Y_i, \mathbf{X}_i)}{\sum_{i=1}^N Z_i w(\mathbf{X}_i)}$$

- For the estimator $\hat{S}_h^{(1)}(t)$, OW remains to be the optimal member in the family of balancing weights (under some conditions)
- Simulations show substantial efficiency improvement for OW compared to IPW when overlap is moderate or poor
- Closed-form sandwich variance estimators are available (via M-estimation), or bootstrap can be used

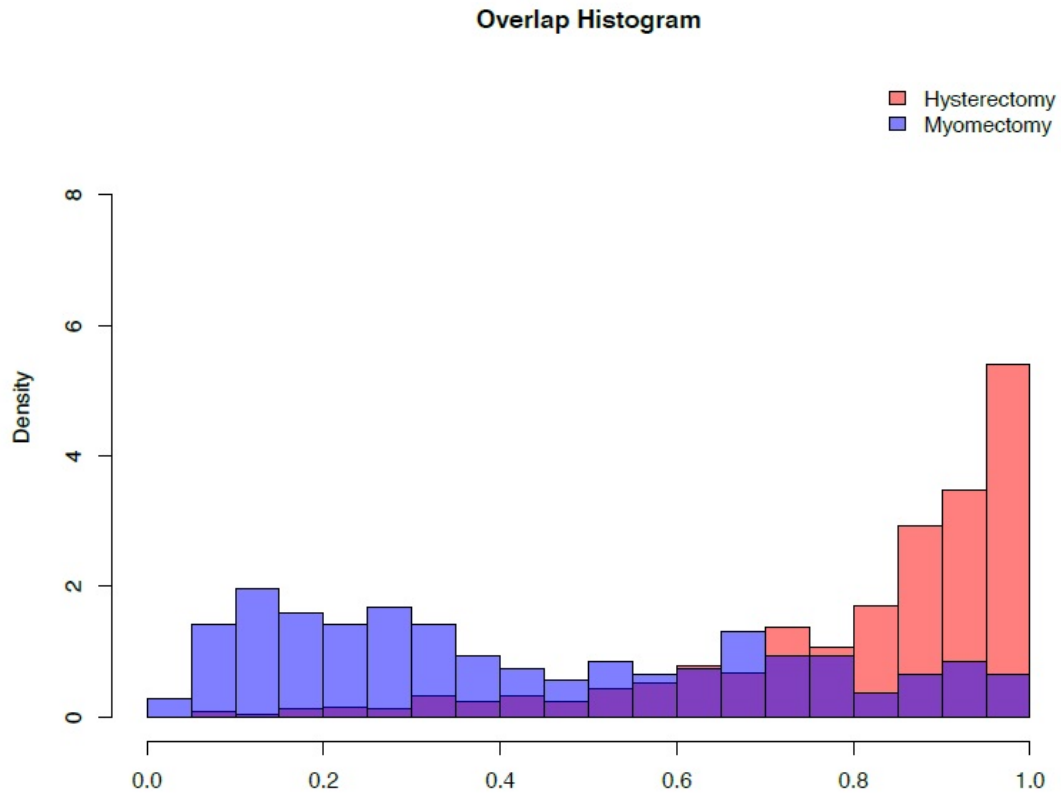
Example: COMPARE-UF

Cheng et al. (2022, AJE)

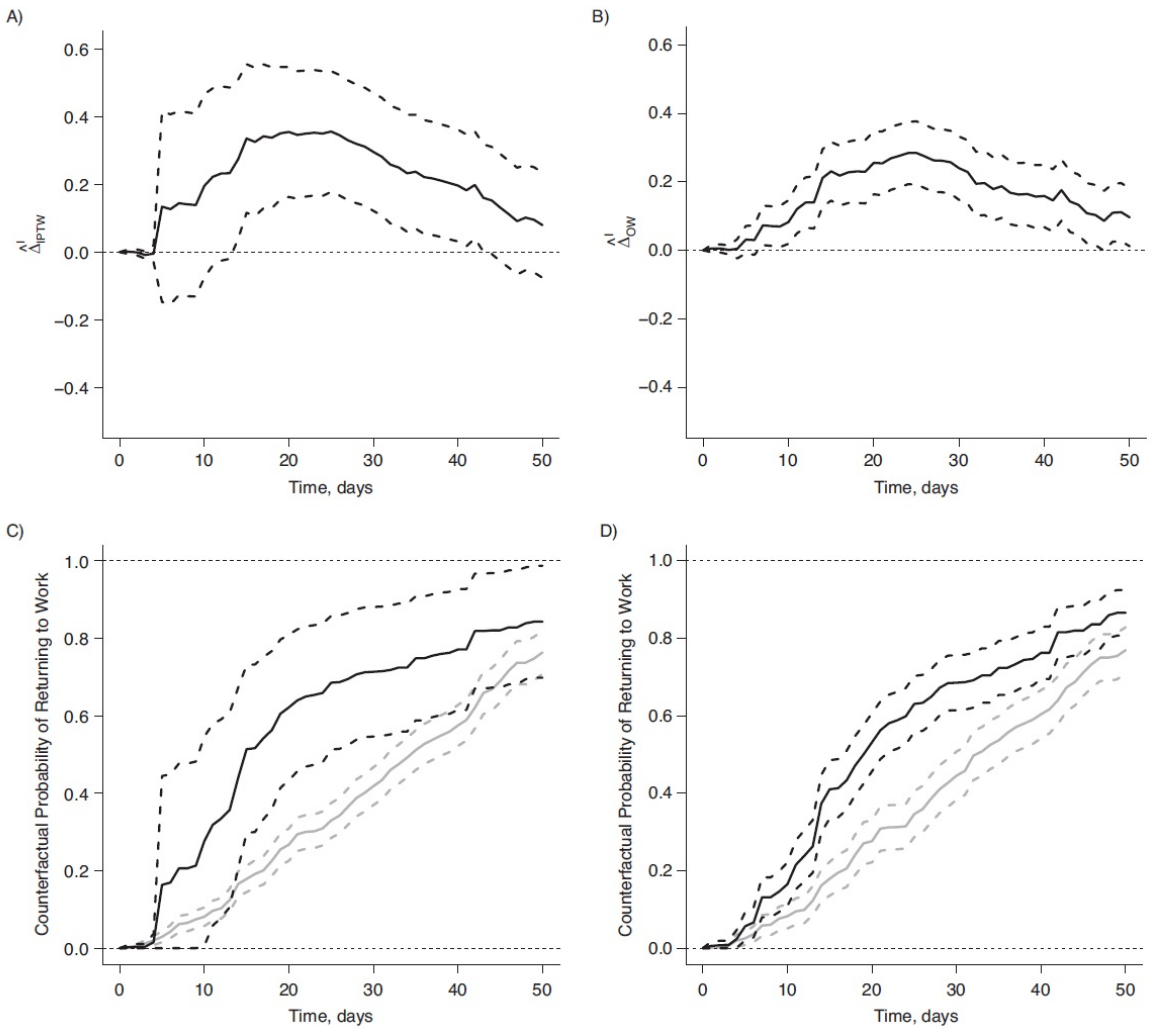
- COMPARE-UF Fibroid Registry (2015-2019)
- Goal: compare the effectiveness of hysterectomy versus myomectomy in the treatment of symptomatic leiomyomas
- Population: women receiving these procedures who were at least 30 years of age, not trying to get pregnant
- Treatments: minimally invasive hysterectomy (506 patients) vs. minimally invasive myomectomy (213 patients)
- Outcome: time from the procedure to returning to work as a time-to-event outcome, with administratively censoring at 60 days
- Approximately 10% of outcomes censored
- Covariates: age, race, ethnicity, type of health insurance, time since first diagnosis of leiomyoma, bleeding symptoms, prior procedures, prior pregnancies, depression/anxiety, etc...

Example: COMPARE-UF

- Propensity score: estimated via a logistic regression with the above covariates
- Limited overlap between the groups: hysterectomy patients are generally older, are more likely to be multiparous
- Censoring scores: estimated via Weibull regression within each group, same covariates as in the PS model

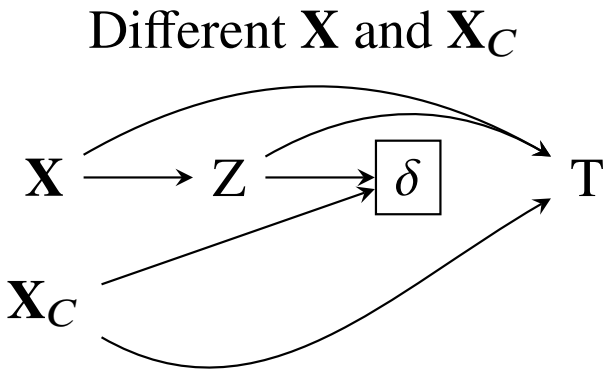


Example: COMPARE-UF



Left: IPTW; Right: OW. Big difference in variance.

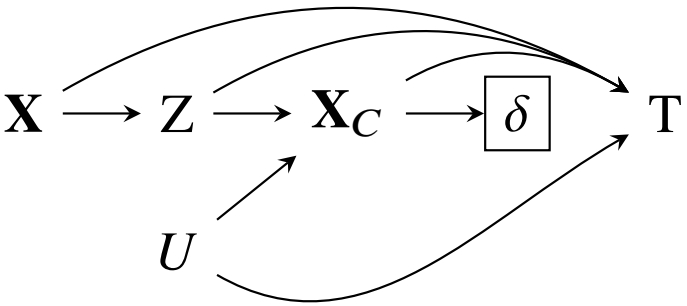
Conditionally independent censoring extensions



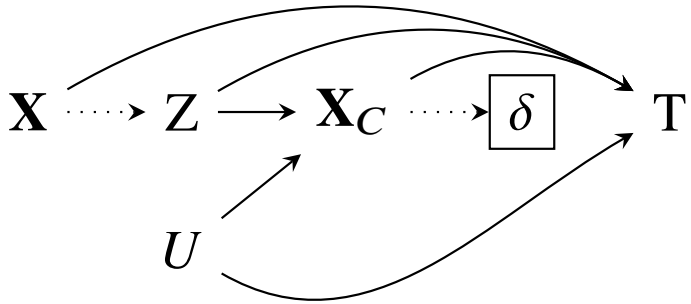
- **X** confounders may be clinical risk factors that indicate treatment (LDL, history of MI) (common causes of treatment and outcome)
- $\mathbf{X}_C$ bias due to post-treatment censoring may be related to health efficacy (age, SES) (common causes of censoring and outcome)
- **X** and $\mathbf{X}_C$ are still pre-treatment variables
- Most likely these are partially overlapping and we can simply include them all into one big set of covariates

Conditionally independent censoring extensions

$\mathbf{X}_C$ arises post-treatment

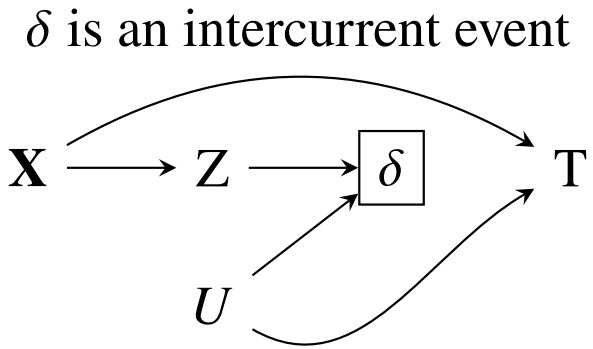


Pseudo-randomization



- Example: $\mathbf{X}_C(t,z)$ are side effects of the drug Z
- Now we need to be careful!
- $\hat{e}(\mathbf{X}) = \widehat{\Pr}(Z = 1|\mathbf{X})$
- $\hat{K}_c^{(z)}(t, \mathbf{X}_C(t,z)) = \Pr(C \geq t|\mathbf{X}_C(t,z), Z = z)$

Dependent censoring



- Example: Patients who experience an intercurrent event are censored δ
- Common causes U of intercurrent events and outcome are often unmeasurable because they include the complete longitudinal disease process (genetics, lifestyle, stress, etc.)
- In this case censoring will be dependent and conditioning on measured covariates will not be sufficient.
- Entirely different methodology is required (Principal Stratification)

Versions of Assumption A3

- Independent censoring:

$$T(z) \perp C(z)$$

- Conditionally independent censoring:

$$T(z) \perp C(z) | \mathbf{X}_C, Z = z$$

*can be extended to allow for dependent censoring conditional on time-varying covariates $\mathbf{X}_C(z,t)$ which are themselves potential outcomes (hence the dependency on both z and t)

- Dependent censoring that is not conditionally independent:

$$T(z) \not\perp C(z) | \mathbf{X}_C, Z = z$$

Key Points

- Confounding is different than the bias that arises due to censoring
- To correctly account for censoring we have to think about the causal structure related to censoring (common causes of censoring and outcome) and identify $\mathbf{X}_C$
- Different methods are applicable depending on the assumptions we are willing to make for A3 (independent, conditionally independent, not conditionally independent)
- Censoring for highly informative reasons (like intercurrent events) will often require alternative methods

References

- Hernán MA. (2010). The hazards of hazard ratios. Epidemiology. 21(1):13.
- Martinussen T, Vansteelandt S, Andersen PK. (2020). Subtleties in the interpretation of hazard contrasts. *Lifetime Data Analysis*. 26:833-55.
- Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in Medicine.* 2013 Jul 20;32(16):2837-49.
- Mao H, Li L, Yang W, Shen Y. (2018) On the propensity score weighting analysis with survival outcome: Estimands, estimation, and inference. *Statistics in Medicine*. 37(26), 3745-3763.
- Cole SR, Hernan MA. Adjusted survival curves with inverse probability weights. *Computer methods and programs in biomedicine.* 2004 Jul 1;75(1):45-9.
- Xie J, Liu C. Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine.* 2005 Oct 30;24(20):3089-110.
- Conner SC, Sullivan LM, Benjamin EJ, LaValley MP, Galea S, Trinquart L. Adjusted restricted mean survival times in observational studies. *Statistics in Medicine.* 2019 Sep 10;38(20):3832-60.
- Satten GA, Datta S. The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. *American Statistician*. 2001;55(3):207-210.
- Cheng C, Li F, Thomas LE, Li F. (2022). Addressing extreme propensity scores in estimating counterfactual survival functions via the overlap weights. *American Journal of Epidemiology*. 191(6), 1140-1151.