# Causal Inference of Truncation-by-Death with Unmeasured Confounding

Xiao-Hua Zhou, Joint Work with Yuhao Deng and Yingjun Chang

Peking University

September 18, 2022

# Motiviting Example

- Stem cell transplantation is a widely adopted approach to treat acute lymphoblastic leukemia, including human leukocyte antigen (HLA)-matched sibling donor transplantation (MSDT) and haploidentical stem cell transplantation (haplo-SCT) from family.
- MSDT remains to be the first choice because the non-relapse mortality (NRM) for MSDT is lower than that for haplo-SCT.
- However, not all patients have HLA-matched MSDT donors, so they can only receive haplo-SCT.
- We want to know whether haplo-SCT can achieve competitive outcomes in terms of leukemia relapse compared with MSDT, by properly adjusting NRM.

# Truncation by Death

- Truncation by death often occurs in longitudinal studies.
- Truncation by death is different from censoring. Censoring means refers to that the outcome exist but was masked by loss of follow-up. Truncation by death renders the outcome undefined.
- The observed survivors in the treated and control groups may possess different underlying features so they are not comparable.
- Analysis based solely on observed survivors may lead to biased conclusions.

# Notations

- $X$: Covariates
- $Z$: Treatment (binary)
- $S(z)$: Survival status
  - $S(z) = 1$, survived
  - $S(z) = 0$, dead (truncated)
- $Y(z)$: Outcome
  - $Y(z)$ is only defined when $S(z) = 1$.
  - Write $Y(z) = *$ when $S(z) = 0$.
- SUTVA and consistency
  - $S = ZS(1) + (1 - Z)S(0),\ Y = ZY(1) + (1 - Z)Y(0)$.
- $V$: Substitutional variable, a proxy of $(S(0), S(1))$

# Principal Stratification

- By principal stratification, we divide the whole population into four strata, indicated by $G$.

| $S(1)$ | $S(0)$ | $G$ | Description | $Y(1)$ | $Y(0)$ |
|---|---|---|---|---|---|
| 1 | 1 | LL | Always-survivors | ✓ | ✓ |
| 1 | 0 | LD | Protected | ✓ | $*$ |
| 0 | 1 | DL | Harmed | $*$ | ✓ |
| 0 | 0 | DD | Doomed | $*$ | $*$ |

- Usually, we assume the DL stratum does not exist (monotonicity).

# Causal Parameters of Interest

- The survivor average causal effect on the control (SACEC)

$$\Delta_C = E[Y(1) - Y(0) \mid G = \mathrm{LL}, Z = 0].$$

(SACE is not identiable)

# Existing Methods

- A biased selection model between $S$ and $Y$.
- Bounds Zhang et al, 2003; Shan et al, 2015; Chiba et al, 2011).
- Sensitivity analysis Egleston et al, 2007; Chiba et al, 2011).
- Invoking additional information (e.g., post-treatment correlates, substitutional variables) (Tchetgen et al, 2014; Ding et al, 2011; Wang et al, 2017)

# Existing Methods with additional information

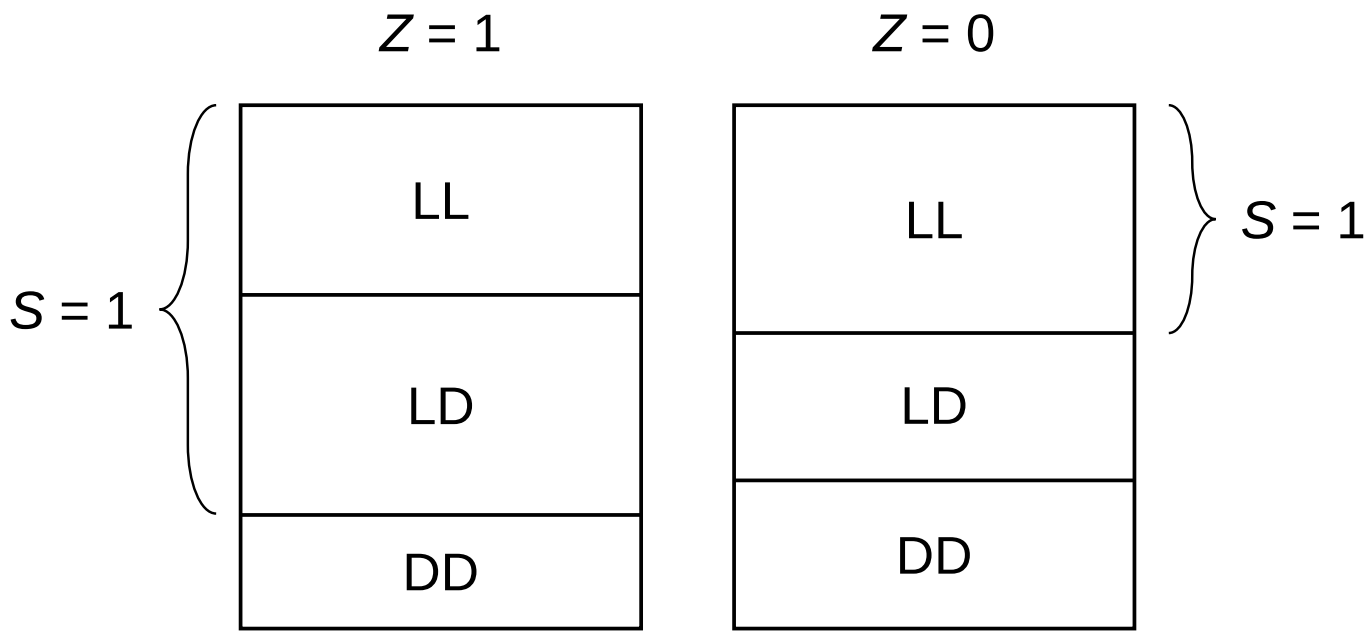
- Randomized experiments with no common causes of the survival and outcome process.
- Randomized experiments with some common causes of the survival and outcome process.

# Observational Studies without Ignorability

- Randomized experiments may not be available in practice.
- Even if randomized experiments are available, they have strict enrollment criteria, so the volunteers cannot reflect the target population.
- In retrospective studies, the treated and controlled units may come from different sources. Observed covariates cannot exactly capture the mechanism of treatment assignments.
- The sample size of randomized experiments are usually small.

# Challenge: Strata Proportions without Ignorability

- The conditional proportions of each principal stratum, $P(G \mid V, X)$, are not identifiable any more.



# Assumptions

- A1. (Latent ignorability) $Y(1), Y(0) \perp\!\!\!\perp Z \mid G = \text{LL}, V, X$. The latent $Y$-ignoarbility assumption states that $X$ and $G$ can adjust the confounding between $Z$ and $Y$.
- A2. (Monotonicity) $S(1) \geq S(0)$.
- A3. (Positivity) $0 < P(Z \mid V, X) < 1$, $0 < P(S(0) \mid Z, V, X) < 1$. The positivity assumption guarantees that always-survivors exist in the treated group and the control group.

- We do not need $S$-ignorability: $(S(1), S(0)) \perp\!\!\!\perp Z \mid X$.
- To disentangle the mixture of LL and LD in the treated group, a substitutional variable for $S(0)$ is needed with following assumptions.

# Assumptions on the Substitutional Variable

- A4. (Substitution relevance) $V \not\perp\!\!\!\perp S(0) \mid Z = 1, S(1) = 1, X$
- A5. (Nondifferential substitution) $V \perp\!\!\!\perp S(1) \mid Z = 1, S(0) = 0, X$.
- A6. (Non-interaction) $E\{Y(z) \mid G, V, X\} = E(Y(z) \mid G, X) + f(V, X)$

# Assumptions on the Substitutional Variable

- Substitution relevance (Assumption **??**) helps distinguish LL and LD strata among observed survivors using the substitutional variable.
- Nondifferential substitution (Assumption **??**) means that for patients who would die if controlled, their chance of survival under the active treatment should only depend on covariates $X$ but not the substitutional variable $V$.
- The substitutional variable $V$ can be understood as a baseline proxy of $S(0)$ with measurement error, and the measurement error is irrelevant to $S(1)$.
- Non-interaction (Assumption **??**) means that the effect of $V$ on $Y$ does not modify the effects of $G$ or $Z$ on $Y$. A special example of non-interaction is exclusion restriction, where $V$ does not have direct effect on $Y$ so that $f(v, x) \equiv 0$.

# Models of Observed Data

- Define the propensity score, survival score and outcome regression model as

$$
\begin{aligned}
e(v,x) &= P(Z=1 \mid V=v, X=x), \\
\pi_z(v,x) &= P(S=1 \mid Z=z, V=v, X=x), \\
m_z(v,x) &= E(Y \mid Z=z, S=1, V=v, X=x),
\end{aligned}
$$

  respectively.

- We denote the principal proportions in the treated group as

$$
\pi_g(v,x) = P(G=g \mid Z=1, V=v, X=x)
$$

  for $g \in \{\mathrm{LL}, \mathrm{LD}, \mathrm{DD}\}$.

# Identification of SACEC

- **Theorem:** Under A1-A6, SACEC is identifiable.

$$\begin{aligned}\Delta_C &= E[R_1(V,X)\{m_1(V,X) - m_0(V,X)\}] \\ &= E\left[R_1(V,X)\left\{\frac{ZS(Y - m_1(V,X))}{e(V,X)\pi_1(V,X)} + m_1(V,X) - m_0(V,X)\right\}\right. \\ &\qquad \left. - R_0(V,X)\left\{\frac{(1-Z)S(Y - m_0(V,X))}{(1-e(V,X))\pi_0(V,X)}\right\}\right].\end{aligned}$$

- The first equation motivates a regression estimator, and the second equation motivates an augmented inverse probability weighting (AIPW) type estimator.

# Estimation

- If $V$ is not binary, then there are many possible ways to construct $h(v, x)$ because only two values of $V$ are required for identification.
- Different choices of $h(v, x)$ can lead to a unique solution of $\Delta_C$. A similar issue for the bridge function in proximal causal inference was discussed in Cui et al (2020). For example, we can take $h(v, x) = v - E(V|X = x)$.
- Since the observed survivors consist solely of always-survivors in the control group according to monotonicity, the target population is individually identified in the expression of SACEC.
- In contrast, the proportion of always-survivors cannot be determined in the treated group, so the SACE in the overall population is not identifiable.

# AIPW Type Estimators

- We focus on the property of the AIPW type estimator

$$\widehat{\Delta}_C = \mathbb{E}_n \left[ \widehat{R}_1(V,X) \left\{ \frac{ZS(Y-\widehat{m}_1(V,X))}{\widehat{e}(V,X)\widehat{\pi}_1(V,X)} + \widehat{m}_1(V,X) - \widehat{m}_0(V,X) \right\} \right.$$
$$\left. - \widehat{R}_0(V,X) \left\{ \frac{(1-Z)S(Y-\widehat{m}_0(V,X))}{(1-\widehat{e}(V,X))\widehat{\pi}_0(V,X)} \right\} \right],$$

  where $\mathbb{E}_n$ means empirical average on the full sample.

- Define an oracle estimator in which all nuisance models are known,

$$\widehat{\Delta}_C^* = \mathbb{E}_n \left[ R_1(V,X) \left\{ \frac{ZS(Y-m_1(V,X))}{e(V,X)\pi_1(V,X)} + m_1(V,X) - m_0(V,X) \right\} \right.$$
$$\left. - R_0(V,X) \left\{ \frac{(1-Z)S(Y-m_0(V,X))}{(1-e(V,X))\pi_0(V,X)} \right\} \right] =: \mathbb{E}_n\{\phi(O)\}.$$

# Large Sample Properties of the Oracle Estimator

- The oracle estimator $\widehat{\Delta}_C^*$ enjoys asymptotic law

$$n^{1/2}\left(\widehat{\Delta}_C^* - \Delta_C\right) \xrightarrow{d} N(0, C),$$

where

$$C = E\left\{\frac{R_1(V,X)^2\sigma_1(V,X)^2}{e(V,X)\pi_1(V,X)} + \frac{R_0(V,X)^2\sigma_0(V,X)^2}{(1-e(V,X))\pi_0(V,X)}\right\} + \text{var}\{\Delta(X)\}.$$

# Large Sample Properties of the AIPW Type Estimator

- Suppose all models are sup-norm consistent upon $(v, x) \in \mathcal{V} \times \mathcal{X}$, and converge at rates faster than $o_p(n^{-1/4})$.
- Define a drift term

$$D_n = \mathbb{E}_n\Big\{(\widehat{R}_1(V,X) - R_1(V,X))(m_1(V,X) - m_0(V,X)) \\ - (R_1(V,X) - R_0(V,X))(\widehat{m}_0(V,X) - m_0(V,X))\Big\},$$

then $\widehat{\Delta}_C - \widehat{\Delta}_C^* \xrightarrow{p} 0$ and $n^{1/2}\left(\widehat{\Delta}_C - \widehat{\Delta}_C^* - D_n\right) \xrightarrow{p} 0$. Thus,

$$n^{1/2}\left(\widehat{\Delta}_C - \Delta_C - D_n\right) \xrightarrow{d} N(0, C).$$

# Large Sample Properties of the AIPW Type Estimator

- Further, if the estimators of the parameters in $R_1(v, x; \theta_1)$ and $m_0(v, x; \theta_0)$ are regular and asymptotic linear (RAL) with influence functions (vectors) $\psi_1(O)$ and $\psi_0(O)$, then $\widehat{\Delta}_C$ is regular and asymptotic linear with influence function

$$\psi(O) = \phi(O) - \Delta_C + \Gamma_1 \psi_1(O) - \Gamma_0 \psi_0(O)$$

  under some regularity conditions, where

$$\Gamma_1 = E\left[\frac{\partial R_1(V, X; \theta_1)}{\partial \theta_1}\{m_1(V, X) - m_0(V, X)\}\right],$$
$$\Gamma_0 = E\left[\frac{\partial m_0(V, X; \theta_0)}{\partial \theta_0}\{R_1(V, X) - R_0(V, X)\}\right].$$

- Thus, $n^{1/2}\left(\widehat{\Delta}_C - \Delta_C\right) \xrightarrow{d} N(0, E\{\psi(O)^2\})$.

# Double Robustness

- The AIPW type estimator has double robustness.
- If $R_z(v,x)$ and $m_0(v,x)$ are correctly specified, then $\widehat{\Delta}_C^*$ and $\widehat{\Delta}_C$ are consistent if either $m_1(v,x)$ or $\{e(v,x), \pi_z(v,x), z=0,1\}$ is correctly specified.
- In practice, the observed survivors in the treated group come from two principal strata, so $m_1(v,x)$ is likely to be misspecified.

# Nondifferential Substitution Revisited

- To relax nondifferential substitution, we require additional knowledge regarding the behaviors of $V$ in the LD and DD strata.
- For an arbitrary $v_0 \in \mathcal{V}$, let

$$\rho(v,x) = \frac{\pi_{\mathrm{DD}}(v,x)}{\pi_{\mathrm{LD}}(v,x)} \Big/ \frac{\pi_{\mathrm{DD}}(v_0,x)}{\pi_{\mathrm{LD}}(v_0,x)},$$

  which measures the odds ratio of DD over LD in the treated group.
- Nondifferential substitution implies that $\rho(v,x) \equiv 1$.

# To Relax Nondifferential Substitution

- Let

$$R^*(v,x) = \frac{\frac{\pi_1(v,x)}{1-\pi_1(v,x)}\rho(v,x)h(v,x)}{\int_{\mathcal{V}} \frac{\pi_1(v,x)}{1-\pi_1(v,x)}\rho(v,x)h(v,x)p(v|x)dv}.$$

- Without nondifferential substitution, if $\rho(v,x)$ is known, then CSACE can be identified as

$$\Delta(x) = E_{V|X=x}\left[R^*(V,X)\{m_1(V,X) - m_0(V,X)\}\right].$$

- Analogously, SACEC is identified by replacing $R(v,x)$ with $R^*(v,x)$.

# Remark: Explainable Nonrandom Survival

- Although the substitutional variable $V$ plays an important role in observational studies for identifying the target causal estimand, interestingly, invoking the substitutional variable does not help identify $\Delta(x)$ under explainable nonrandom survival (also known as principal ignorability):

$$E\{Y(1) \mid G = \mathrm{LL}, V, X\} = E\{Y(1) \mid G = \mathrm{LD}, V, X\}.$$

- For any misspecified $\rho(v, x)$,

$$\Delta(x) = E_{V|X=x}[R^*(V, X)\{m_1(V, X) - m_0(V, X)\}]$$

always holds.

# Simulation Settings

- Baseline covariates: $(X_1, X_2) \sim N(1, 1, 1^2, 1^2, 0.5)$, $X_3 \sim U(0, 2)$. Write $X = (1, X_1, X_2, X_3)$.
- Substitutional variable: $V \sim N(X\zeta, 2^2)$.
- Treatment assignment:

$$P(Z = 1|V, X) = expit\{(X, V)\alpha\}.$$

- Survival process:

$$P(S(0) = 1|V, X, Z) = expit\{(X, V)\beta_Z\},$$
$$P(S(1) = 1|V, X, Z, S(0) = 0) = expit\{(X, V)\gamma_Z\}.$$

- Outcome process:

$$Y(z) \sim N((X, V, (0), S(1))\delta_Z, 1^2).$$

# Simulation Settings

- Set $\zeta = (-1, 1, 0, 0)'$, $\alpha = (0, -1, 0, 1, 1)'$, $\beta_0 = (-2, 2, 2, 2, 4)'$, $\gamma_0 = (-1, -1, 1, -1, 0)$.
- Setting 1: Constant treatment effects.
- Setting 2: Ignorability (stratified randomized experiment) and exclusion restriction (no unmeasured confounding between $Z$ and $S$).
- Setting 3: Explainable nonrandom survival and exclusion restriction (unmeausred confounding between $S$ and $Y$)
- Setting 4: Heterogeneous treatment effects with exclusion restriction.

# Simulation results for estimating the SACEC

| Setting | $n$ | Average bias | | | | Root mean squared error | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | SC | WZR | AIPW | REG | SC | WZR | AIPW | REG |
| 1 | 200 | -4.536 | -3.412 | -0.052 | 0.227 | 4.587 | 3.457 | 0.891 | 0.937 |
| | 500 | -4.520 | -3.374 | -0.061 | 0.376 | 4.539 | 3.391 | 0.587 | 0.742 |
| | 2000 | -4.527 | -3.388 | -0.048 | 0.617 | 4.532 | 3.392 | 0.332 | 0.724 |
| 2 | 200 | -0.782 | 0.053 | -0.061 | 0.099 | 0.941 | 0.301 | 1.212 | 1.231 |
| | 500 | -0.773 | 0.089 | -0.022 | 0.184 | 0.836 | 0.213 | 0.746 | 0.762 |
| | 2000 | -0.783 | 0.095 | -0.014 | 0.264 | 0.800 | 0.134 | 0.526 | 0.579 |
| 3 | 200 | -0.722 | 0.165 | -0.010 | -0.004 | 0.898 | 0.377 | 0.615 | 0.583 |
| | 500 | -0.714 | 0.211 | 0.016 | 0.011 | 0.782 | 0.300 | 0.416 | 0.392 |
| | 2000 | -0.730 | 0.203 | -0.003 | -0.004 | 0.750 | 0.230 | 0.234 | 0.205 |
| 4 | 200 | -3.047 | -1.991 | -0.039 | 0.122 | 3.158 | 2.087 | 1.557 | 1.580 |
| | 500 | -3.019 | -1.944 | -0.007 | 0.199 | 3.061 | 1.980 | 0.920 | 0.942 |
| | 2000 | -3.037 | -1.955 | 0.007 | 0.285 | 3.048 | 1.964 | 0.655 | 0.712 |

# Simulation Conclusion

- The proposed AIPW type estimator is asymptotically unbiased when Assumptions 1 to 6 hold and is robust to model misspecification.
- Even if nondifferential substitution is violated, sensitivity analysis can help assess the influence of such violation.

# Clinical Background

- Stem cell transplantation is a widely adopted approach to treat acute lymphoblastic leukemia, including human leukocyte antigen (HLA)-matched sibling donor transplantation (MSDT) and haploidentical stem cell transplantation (haplo-SCT) from family.
- There are fewer mismatched HLA loci between the donor and patient undergoing MSDT than underdoing haplo-SCT.
- The non-relapse mortality (NRM) for MSDT is lower than that for haplo-SCT (monotonicity), so doctors prefer MSDT.
- However, not all patients have HLA-matched MSDT donors, so they can only receive haplo-SCT.
- We want to know whether haplo-SCT can achieve competitive outcomes in terms of leukemia relapse compared with MSDT, by properly adjusting NRM.

# Data Notations

- Data from a retrospective study were collected at Peking University People's Hospital.
- The data include 1161 patients undergoing allogeneic stem cell transplantation from 2009 to 2017.
- The transplantat policy was "to treat with what you have" (MSDT preferred).
- $Z$: Transplantation type.
  - $Z = 1$: MSDT; $Z = 0$: Haplo-SCT.
- $S$: Absence of non-relapse mortality (NRM).
- $Y$: Leukemia relapse in two years.

# Baseline Covariates and Latent Ignorability

- Previous studies found risk factors for relapse:

1. $X_1$: Presence of minimum residual disease (MRD) before transplantation (Positive/Negative);
2. $X_2$: Disease status (CR1/CR>1);
3. $X_3$: Diagnosis (T-ALL/B-ALL);
4. Post-treatment chronic graft-versus-host disease (GVHD).

- To avoid conditioning on post-treatment variables, we condition on $(S(0), S(1))$ rather than post-treatment GVHD, since acute GVHD is a risk factor for mortality.
- So it is unnatural to assume explainable nonrandom survival.
- By conditioning on $X = (X_1, X_2, X_3)$ and $G = (S(0), S(1))$, we accept latent ignorability.

# Finding A Substitutional Variable

- Association studies have indicated that older people have a higher probability of non-relapse mortality because older people are more vulnerable to infection after surgery (substitution relevance).
- Meanwhile, they are more likely to have siblings and access MSDT as a result of the one-child policy in China.
- However, age is not considered as a risk factor of relapse based on clinical evidence. This result is also confirmed using our data.
- Even if age could influence relapse, age should have a similar effect on $Y(1)$ and $Y(0)$ in all survivors, because there is no biological mechanism indicating pleiotropy for age between different transplant modalities (non-interaction).
- We will do sensitivity analysis on nondifferential substitution.

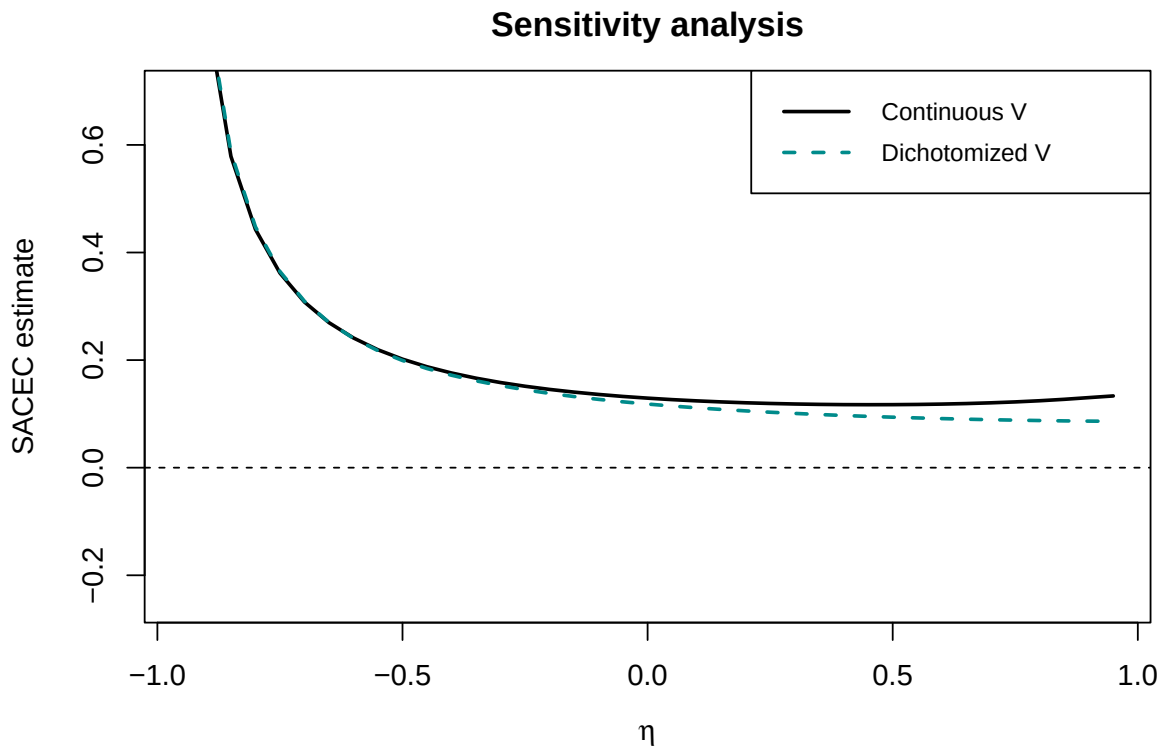
# Data Analysis Results

- All methods yield positive point estimates of SACEC, indicating that haplo-SCT has stronger graft-versus-leukemia effect.
- The confidence interval is wide probably because the substitutional variable $V$ is weak.
- We conclude that halpo-SCT is a non-inferior alternative to MSDT in terms of relapse.

| Method | SACEC estimate (se) | 95% Confidence interval |
|---|---|---|
| Survivor-case | 0.0703 (0.0298) | (0.0118, 0.1288) |
| WZR | 0.0999 (0.3020) | (-0.2022, 0.4020) |
| Regression | 0.1522 (0.0995) | (-0.0428, 0.3472) |
| AIPW type | 0.1292 (0.1291) | (-0.1240, 0.3824) |

# Sensitivity Analysis

- The conclusion is robust to violation of nondifferential substitution.
- The dashed line represents dichotomized $V$ cut at its mean ($V > 27$ or not).



# Concluding Remarks

- We proposed an identification method for the survival average causal effect on the control, when both (a) the treatment assignment and survival process and (b) the survival and outcome process are confounded.
- Under monotonicity, the target population consists of survivors in the control group. This target population is well defined.
- The AIPW type estimator is robust to model misspecification. When all models are correctly specified, it has good asymptotic properties.
- The proposed method provides convenience for post-marketing safety or efficiency evaluation, after a drug has been proven beneficial on survival (or a surrogate, or repsonse).

# Relation with Proximal Causal Inference

- The idea of introducing a substitutional variable has some connections with proximal causal inference.
- We are interested in local effects, defined on a single principal stratum.
- We identify SACEC using a substitutional variable for $G$, rather than finding proxies of unmeasured confounders.
- Proximal causal inference requires the proxies of unmeasured confounders be rich enough to cover the information in the unmeasured confounders (referred to as necessity). It may be easy to find a risk factor for survival, but it could be difficult to find such proxies.

# Acknowledgements


- We thank Shanshan Luo at Peking University for discussion.
- We thank Leqing Cao at Peking University People's Hospital for cleaning the data.
- Methods are implemented using R 4.1.3. Wang et al. (2017)'s method WZR is available on CRAN R package "tbd".
- Funding information: National Natural Science Foundation of China, Grant No. 81773546, 12026606; National Key Research and Development Program of China, Grant No. 2021YFF0901400.

Thanks!