Enhancing Statistical Validity and Power in Hybrid Controlled Trials

A Randomization Inference Approach with Conformal Selective Borrowing

Ke Zhu

Department of Statistics, North Carolina State University
Department of Biostatistics and Bioinformatics, Duke University

Joint work with Shu Yang (NCSU) and Xiaofei Wang (Duke)

June 13-19, 2025
Forty-Second International Conference on Machine Learning (ICML), Vancouver

Disclaimer

FDA grant U01FD007934: Methods to improve efficiency and robustness of clinical trials using information from real-world data with hidden bias

This project is supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) as part of a financial assistance award U01FD007934 totaling $1,674,013 over two years funded by FDA/HHS. The contents do not necessarily represent the official views of, nor an endorsement by, FDA/HHS, or the U.S. Government.

Motivation

Integrating CALGB 9633 with NCDB External Controls

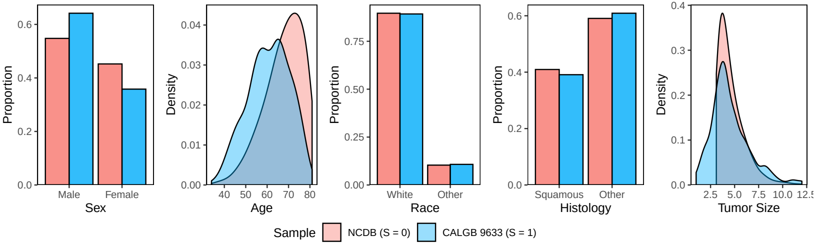
Scientific Objective: Evaluate the efficacy of *adjuvant chemotherapy* vs. *observation* after surgery in Stage IB non–small-cell lung cancer patients (Strauss et al., 2008).

CALGB 9633 trial: Underpowered, took 12 years due to slow accrual.

National Cancer Database (NCDB): Large database including patients under observation (external controls), which may have covariate shift and outcome incomparability.

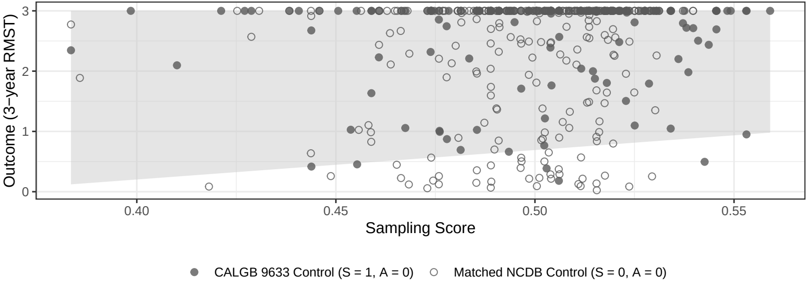
A hybrid controlled trial: CALGB 9633 trial + NCDB external controls (ECs) to improve treatment effect estimation and inference.

Covariate Shift



ECs are older with larger tumors than CALGB 9633 patients.

Outcome Incomparability



After adjusting for covariate shift (by matching and comparing within similar sampling scores)

- some ECs are comparable
- some ECs exhibit lower Y than RCT controls

Challenge and Contribution

- **RCT-only**: underpowered.
- **RCT+EC**: estimation bias and inflated Type I error from covariate shift and outcome incomparability.
- Covariate shift has been addressed by propensity score methods.
- Our contributions:
  - **Conformal selective borrowing** for outcome comparability.
  - **Fisher randomization tests** to control Type I error.
  - Power gain via combining both methods.

Problem Setup & Benchmarks

Causal Inference Framework

Source	Total	Treated ($A = 1$)	Control ($A = 0$)
CALGB 9633 ($S = 1$)	335 ($n_{\mathcal{R}}$)	167 (n_1)	168 (n_0)
NCDB ($S = 0$)	11,446 ($n_{\mathcal{E}}$)	–	11,446

Outcome Y: 3-year Restricted Mean Survival Time $\min(T, 3)$.

Covariates X: Sex, age, race, histology, and tumor size.

Data: $\{Y_i, X_i, A_i, S_i\}_{i=1}^n$, $n = n_{\mathcal{R}} + n_{\mathcal{E}}$.

Potential Outcomes: $Y(1)$, $Y(0)$.

Estimand: Average treatment effect (ATE) in the RCT population,

$$\tau = \mathbb{E}\{Y(1) - Y(0) \mid S = 1\}.$$

Benchmark 1: No Borrow AIPW

Assumption 1: Identification (Held by RCT Design)

1.(Consistency) $Y = AY(1) + (1 - A)Y(0)$.

2.(Positivity) $0 < e(x) < 1$ for all x with $f_{X|S}(x|1) > 0$, where $f_{X|S}(x|s)$ is the conditional density of X.

3.(Randomization) $Y(a) \perp\!\!\!\perp A \mid (X, S = 1)$, for $a = 0, 1$.

Propensity Score: $e(X) = \mathbb{P}(A = 1 \mid X, S = 1)$.

Outcome Model: $\mu_a(X) = \mathbb{E}\{Y(a) \mid X, S = 1\}$.

No Borrow AIPW (RCT-only, covariate-adjusted ATE estimator)

$$\hat{\tau}_{\mathcal{R}} = \frac{1}{n_{\mathcal{R}}} \sum_{i=1}^{n} S_i \Bigg[\hat{\mu}_{1,\mathcal{R}}(X_i) + \frac{A_i}{\hat{e}(X_i)} \{Y_i - \hat{\mu}_{1,\mathcal{R}}(X_i)\} - \hat{\mu}_{0,\mathcal{R}}(X_i) - \frac{1 - A_i}{1 - \hat{e}(X_i)} \{Y_i - \hat{\mu}_{0,\mathcal{R}}(X_i)\} \Bigg].$$

Benchmark 2: Borrow AIPW

Assumption 2: Mean Exchangeability of ECs (Relaxed Later)

$$\mathbb{E}\{Y(0) \mid X, S = 0\} = \mathbb{E}\{Y(0) \mid X, S = 1\}.$$

Sampling Score: $\pi(X) = \mathbb{P}(S = 1 \mid X)$.

Borrow AIPW (RCT + All ECs, address covariate shift)

$$\hat{\tau}_{\mathcal{R}+\mathcal{E}} = \frac{1}{n_{\mathcal{R}}} \sum_{i=1}^{n} \Bigg[S_i \, \hat{\mu}_{1,\mathcal{R}}(X_i) + S_i \frac{A_i}{\hat{e}(X_i)} \{Y_i - \hat{\mu}_{1,\mathcal{R}}(X_i)\} - S_i \, \hat{\mu}_{0,\mathcal{R}+\mathcal{E}}(X_i) - \hat{\pi}_{\mathcal{E}}(X_i) \frac{S_i(1 - A_i) + (1 - S_i)\hat{r}_{\mathcal{E}}(X_i)}{\hat{\pi}_{\mathcal{E}}(X_i)\{1 - \hat{e}(X_i)\} + \{1 - \hat{\pi}_{\mathcal{E}}(X_i)\}\hat{r}_{\mathcal{E}}(X_i)} \{Y_i - \hat{\mu}_{0,\mathcal{R}+\mathcal{E}}(X_i)\} \Bigg].$$

- **Outcome modeling** using both RCT data and ECs.
- **Inverse sampling score weighting** to align ECs's covariate distribution.
- **Inverse variance weighting** by $r(X) = \frac{\mathbb{V}\{Y(0) \mid X, S=1\}}{\mathbb{V}\{Y(0) \mid X, S=0\}}$ for maximal efficiency.
- Doubly robust and locally efficient (Li et al., 2023); biased if Assumption 2 fails.

Conformal Selective Borrowing

Testing Individual Outcome Comparability

For EC $j \in \mathcal{E}$, define individual bias as $b_j \equiv Y_j - \mathbb{E}\{Y(0) \mid X, S = 1\}$.

$H_0^j : b_j = 0$ is testable with RCT controls.

Conformal p-value (Vovk, Gammerman, and Shafer, 2005)

1. **Split** RCT controls into calibration set $\mathcal{C}_1$ and training set $\mathcal{C} \setminus \mathcal{C}_1$.
2. **Train** $\hat{f}_{-\mathcal{C}_1}(x)$ on $\mathcal{C} \setminus \mathcal{C}_1$ to predict comparable EC outcomes.
3. **Measure the comparability** of EC j to $\hat{f}_{-\mathcal{C}_1}(x)$ by conformal score

$$s_j = |Y_j - \hat{f}_{-\mathcal{C}_1}(X_j)|.$$

4. **Calibrate** the conformal score using $s_i = |Y_i - \hat{f}_{-\mathcal{C}_1}(X_i)|$ for $i \in \mathcal{C}_1$,

$$p_j = \frac{\sum_{i \in \mathcal{C}_1} \mathbb{I}(s_i \geq s_j) + 1}{|\mathcal{C}_1| + 1}.$$

Boosting performance: (i) Split $\rightarrow$ CV+ (Barber et al., 2021), (ii) Absolute Residual $\rightarrow$ Conformalized Quantile Regression (Romano, Patterson, and Candès, 2019).

Conformal Selective Borrow AIPW

Full EC set $\mathcal{E} \to$ Selected EC set $\hat{\mathcal{E}}(\gamma) = \{j \in \mathcal{E} : p_j > \gamma\}$.

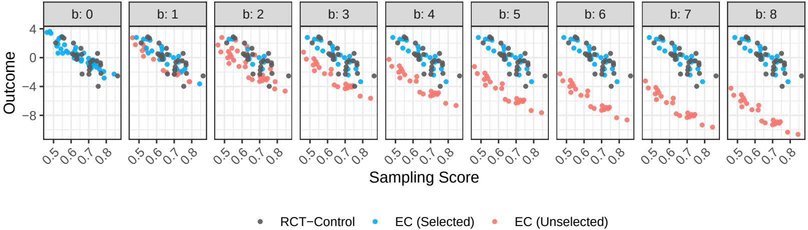
Borrow AIPW $\hat{\tau}_{\mathcal{R}+\mathcal{E}} \to$ a class of estimators indexed by γ:

$$\hat{\tau}_\gamma = \frac{1}{n_{\mathcal{R}}} \sum_{i=1}^{n} \Bigg[S_i\, \hat{\mu}_{1,\mathcal{R}}(X_i) + S_i \frac{A_i}{\hat{e}(X_i)} \{Y_i - \hat{\mu}_{1,\mathcal{R}}(X_i)\} - S_i\, \hat{\mu}_{0,\mathcal{R}+\hat{\mathcal{E}}(\gamma)}(X_i)$$
$$- \hat{\pi}_{\hat{\mathcal{E}}(\gamma)}(X_i) \frac{S_i(1-A_i) + (1-S_i)\mathbb{I}\{i \in \hat{\mathcal{E}}(\gamma)\}\hat{r}_{\hat{\mathcal{E}}(\gamma)}(X_i)}{\hat{\pi}_{\hat{\mathcal{E}}(\gamma)}(X_i)\{1-\hat{e}(X_i)\} + \{1-\hat{\pi}_{\hat{\mathcal{E}}(\gamma)}(X_i)\}\hat{r}_{\hat{\mathcal{E}}(\gamma)}(X_i)} \{Y_i - \hat{\mu}_{0,\mathcal{R}+\hat{\mathcal{E}}(\gamma)}(X_i)\} \Bigg].$$

Examples

1. **No Borrow AIPW** $\hat{\tau}_{\mathcal{R}} = \hat{\tau}_1$ since $\hat{\mathcal{E}}(1) = \varnothing$.
2. **Borrow AIPW** $\hat{\tau}_{\mathcal{R}+\mathcal{E}} = \hat{\tau}_0$ since $\hat{\mathcal{E}}(0) = \mathcal{E}$.
3. **Conformal Selective Borrow AIPW** $\hat{\tau}_{\hat{\gamma}}$ with $\hat{\gamma}$ minimizing $\widehat{\mathrm{MSE}}(\gamma)$.
   - $\mathrm{MSE}(\gamma) = \{\mathbb{E}(\hat{\tau}_\gamma) - \tau\}^2 + \mathbb{V}(\hat{\tau}_\gamma)$.
   - Use $\hat{\tau}_1$ (consistent for τ) to approximate squared bias:
   $$\{\mathbb{E}(\hat{\tau}_\gamma - \tau)\}^2 \approx \{\mathbb{E}(\hat{\tau}_\gamma - \hat{\tau}_1)\}^2 = \mathbb{E}(\hat{\tau}_\gamma - \hat{\tau}_1)^2 - \mathbb{V}(\hat{\tau}_\gamma - \hat{\tau}_1).$$
   - Estimate $\mathbb{E}(\hat{\tau}_\gamma - \hat{\tau}_1)^2$ via $(\hat{\tau}_\gamma - \hat{\tau}_1)^2$. Estimate $\mathbb{V}(\hat{\tau}_\gamma - \hat{\tau}_1)$ and $\mathbb{V}(\hat{\tau}_\gamma)$ via bootstrap.

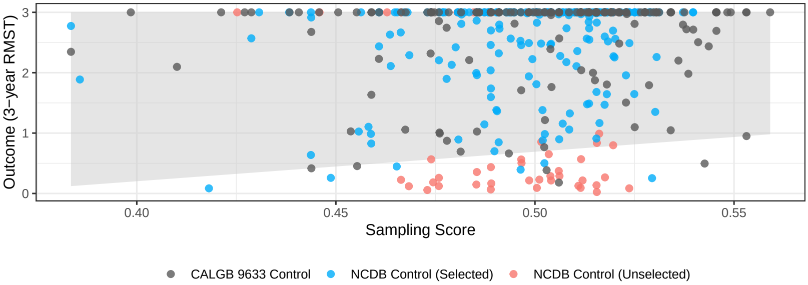
Simulation*: EC Selection



For various levels *b* of hidden bias, CSB AIPW discards biased ECs conditional on all measured covariates.

*See appendix for simulation setup

Real Data: EC Selection



CSB AIPW selects ECs with conditional outcomes closer to RCT controls, reducing hidden bias beyond balancing *X* alone.

Fisher Randomization Test

Randomization Inference in Hybrid Controlled Trials

Fisher Randomization Test (Fisher, 1935)

1. **Sharp Null**: $H_0 : Y_i(0) = Y_i(1), \forall i \in \mathcal{R}$, imputing all $Y_i(a)$.
2. **Test Statistic**: Compute $T(\boldsymbol{A}^{\mathrm{obs}})$ for actual assignment $\boldsymbol{A}^{\mathrm{obs}}$.
3. **Analyze as You Randomize**:
   - Generate A_i^b for RCT patients per the actual randomization procedure.
   - Keep $A_i^b \equiv 0$ for ECs, as they remain fixed during randomization in RCT.
4. **Compute p value**: Repeat for B iterations and compute:

$$\hat{p}^{\mathrm{FRT}} = \frac{\sum_{b=1}^{B} \mathbb{I}\{T(\boldsymbol{A}^b) \geq T(\boldsymbol{A}^{\mathrm{obs}})\} + 1}{B+1}.$$

FRT: A Backup for Strict Type I Error Control

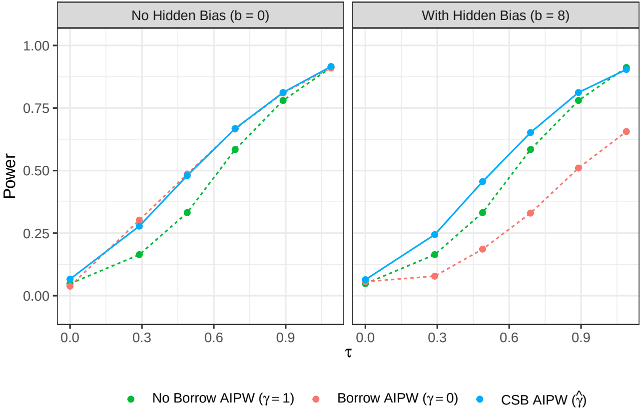
Finite-Sample Exact: Valid for any sample size.

Model-Free: Remains valid even if models are misspecified.

Valid for Any Test Statistic:

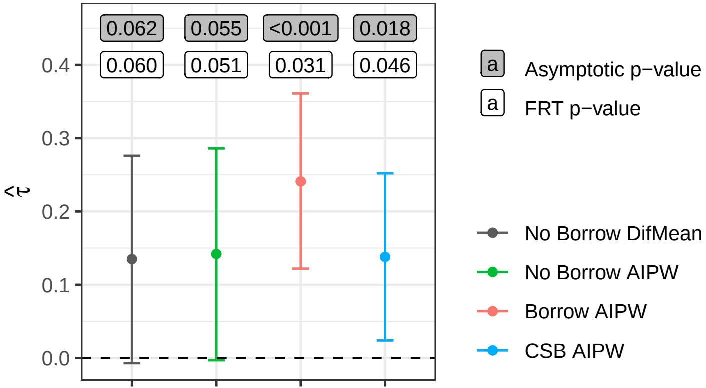
- **Bias-aware**: If T is Borrow AIPW that is biased without Assumption 2, FRT replicates the biased distribution.
- **Post-selection valid**: If T is CSB AIPW, FRT accounts for selection uncertainty by varying $\hat{\mathcal{E}}(\gamma)$ with $\boldsymbol{A}^b$.

Simulation: Power Curves of FRTs



- **FRTs** control type I error at $\tau = 0$ for any test statistic.
- **CSB AIPW** achieves the highest power.

Real Data: Inference Results



- **CSB AIPW** improves borderline non-significant **No Borrow AIPW**.
- **CSB AIPW** mitigates overly large **Borrow AIPW** estimates.

Conclusion

Takeaway Messages

1. **Conformal Selective Borrow AIPW** addresses both covariate and outcome incomparability of external controls.
    - Finite-sample exact, model-free, selective borrowing.

2. **Fisher randomization test** with Conformal Selective Borrow AIPW as a test statistic controls type I error and gains power when EC bias is negligible or detectable.
    - Finite-sample exact, model-free, post-selection valid inference.

3. User-friendly R package `intFRT` available at

    `github.com/ke-zhu/intFRT`

Thank you!

Simulation Setup

Sample Sizes	$(n_1, n_0, n_{\mathcal{E}}) = (50, 25, 50)$
Covariates	$X \sim \text{Unif}(-2, 2)$, $p = 2$
Sampling	$S \sim \text{Bernoulli}(\pi(X))$ $\pi(X) = \{1 + \exp(\eta_0 + X^{\mathrm{T}}\eta)\}^{-1}$, $\eta = (0.1, 0.1)$
Assignment	$A \sim \text{Bernoulli}(n_1/n_{\mathcal{R}})$ for $S = 1$ $A = 0$ for $S = 0$
Potential Outcomes ($S = 1$)	$Y(0) = X^{\mathrm{T}}\beta_0 + \varepsilon$, $Y(1) = 0.4 + X^{\mathrm{T}}\beta_1 + \varepsilon$ $\varepsilon \sim N(0, 1)$, $\beta_0 = (1, 1)$, $\beta_1 = (2, 2)$
Potential Outcomes ($S = 0$)	**(i) No Hidden Bias** $b = 0$ $Y(0) = X^{\mathrm{T}}\beta_0 + 0.5\varepsilon$ **(ii) Half of ECs with Hidden Bias** $b = 1, 2, \ldots, 8$ For 50% of ECs, $Y(0) = -b + X^{\mathrm{T}}\beta_0 + 0.5\varepsilon$
Observed Outcomes	**Under** H_1: $Y = AY(1) + (1 - A)Y(0)$ **Under** H_0: $Y = Y(0)$

References i

- Barber, Rina Foygel et al. (2021). "Predictive inference with the jackknife+". In: *The Annals of Statistics* 49.1, pp. 486–507.
- Fisher, R. A. (1935). *The Design of Experiments*. 1st. Oliver and Boyd, Edinburgh.
- Li, Xinyu et al. (2023). "Improving efficiency of inference in clinical trials with external control data". In: *Biometrics* 79.1, pp. 394–403.
- Romano, Yaniv, Evan Patterson, and Emmanuel J Candès (2019). "Conformalized quantile regression". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Vol. 32, pp. 3543–3553.

Strauss, Gary M et al. (2008). “Adjuvant paclitaxel plus carboplatin compared with observation in stage IB non-small-cell lung cancer: CALGB 9633 with the Cancer and Leukemia Group B, Radiation Therapy Oncology Group, and North Central Cancer Treatment Group Study Groups”. In: *Journal of Clinical Oncology* 26.31, pp. 5043–5051.

Vovk, Vladimir, Alexander Gammerman, and Glenn Shafer (2005). *Algorithmic Learning in a Random World*. Vol. 29. Springer.