Combining RCT and Observational Data for Causal Inference

Hill Zhang Advised by: Professor Evan Rosenman

Department of Mathematical Sciences Claremont Mckenna College

Motivation

- In medicine and public health, randomized controlled trials (RCTs) are the gold standard for estimating treatment effects—they provide unbiased estimates but are often small and expensive.
- Observational studies are typically much larger and more affordable, but they are vulnerable to confounding and selection bias.
- The core challenge: RCTs offer **low bias, high variance**, while observational data offer **high bias, low variance**.
- **Goal:** combine both sources to estimate treatment effects more accurately by balancing bias and precision.

We estimate subgroup-specific treatment effects (CATEs) from each source, then blend them with empirical-Bayes shrinkage.

Data & Methods

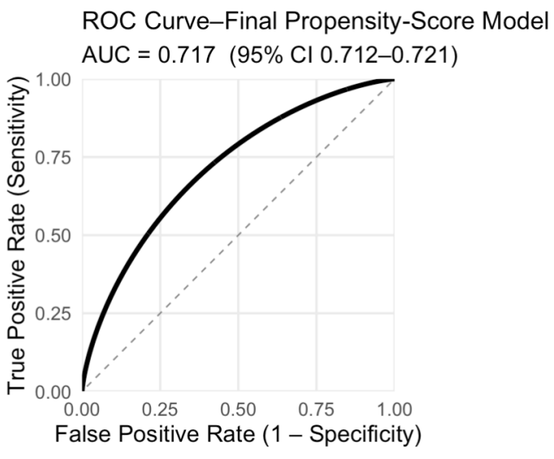
- **Dataset (WHI)** – 1993 study of hormone therapy in postmenopausal women, combining an RCT and a large observational cohort.
- **Propensity Score (PS)** – logistic regression of `Treatment Indicator` on *AGE_CAT*, *BMI*, *SMOKING*, *ETHNIC_GRP*, *WAIST*, *HOTFLASH*, *MENO* using the observational cohort. Diagnostics: AUC = 0.717, max SMD = 0.027.
- **CATE**

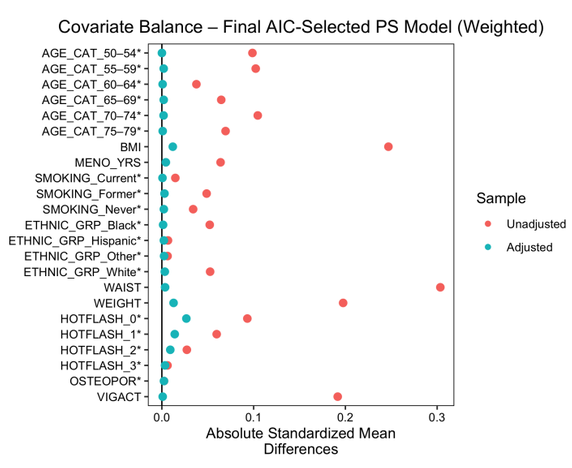
$$\tau(x) = E\big[Y(1) - Y(0) \,\big|\, X = x\big]$$

  (subgroup risk difference).
- **Estimation** – X-learner on both datasets (OBS: fitted PS; RCT: randomized with constant PS = 0.506).
- **Empirical-Bayes shrinkage** – combines RCT (low bias, high variance) and OBS (high bias, low variance) to improve accuracy.

Propensity-Score Model Diagnostics

- **Final PS model (OBS cohort).** AUC = 0.717 ; max weighted SMD = 0.027 ⇒ good discrimination and excellent balance.





Which Variables Drive the PS Model?

Top Propensity-Model Coefficients

VARIABLE	COEF.	SE	Z-STAT
AGE_CAT70–74	-1.901	0.041	-46.574
AGE_CAT75–79	-2.484	0.059	-42.130
AGE_CAT65–69	-1.393	0.036	-39.013
HOTFLASH1	-0.863	0.029	-29.255
AGE_CAT60–64	-0.799	0.034	-23.611
HOTFLASH2	-1.134	0.056	-20.395
ETHNIC_GRPWhite	0.947	0.053	17.753

Uncertainty Estimation

- **Bootstrap resampling** — Resampling procedure to estimate uncertainty for complex statistical estimates when standard errors are unavailable.
- **Implementation for RCT CATEs:**
  - 100 independent bootstrap draws of the RCT cohort
  - Re-fit the X-learner on each draw
  - Compute the variance of the resulting CATE estimates across draws
- **Why it matters** — These variance estimates provide the precision weights used in our empirical-Bayes shrinkage to combine RCT and OBS CATEs.

Empirical-Bayes Shrinkage: Two Estimators

- **Inputs (on RCT rows):** $\hat{f}_r(x)$ = RCT CATE (bootstrap mean), $\mathrm{Var}_r(x)$ = RCT variance (from bootstraps), $\hat{f}_o(x)$ = OBS CATE predicted on RCT rows.

1) Global shrinkage weight (one λ for all rows)

$$\lambda_{\mathrm{URE}} = \frac{\sum_k \mathrm{Var}_r(x_k)}{\sum_k \left(\hat{f}_r(x_k) - \hat{f}_o(x_k)\right)^2}$$

$$\hat{f}_c(x) = (1 - \lambda_{\mathrm{URE}})\,\hat{f}_r(x) + \lambda_{\mathrm{URE}}\,\hat{f}_o(x)$$

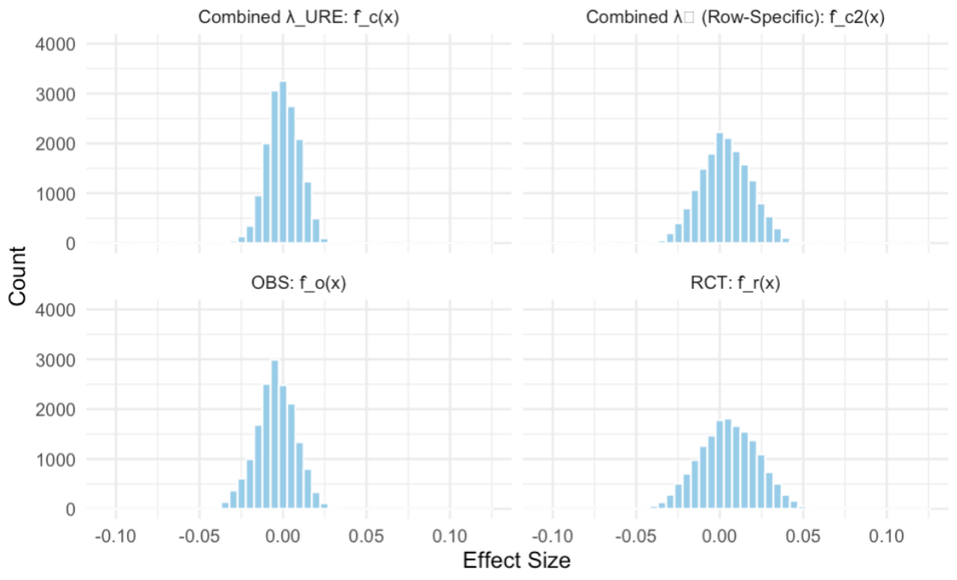
2) Row-specific shrinkage weight (a distinct $\lambda_2(x_k)$ per row)

$$\lambda_2(x_k) = \frac{\left(\sum_j \mathrm{Var}_r(x_j)^2\right)\mathrm{Var}_r(x_k)}{\sum_j \mathrm{Var}_r(x_j)^2\left(\hat{f}_r(x_j) - \hat{f}_o(x_j)\right)^2}$$

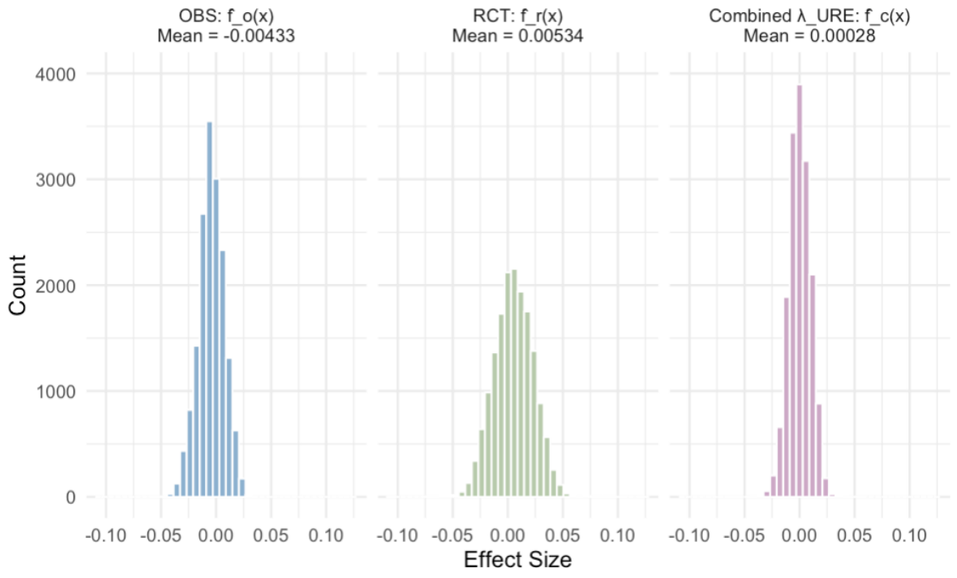
$$\hat{f}_{c2}(x_k) = \big(1 - \lambda_2(x_k)\big)\,\hat{f}_r(x_k) + \lambda_2(x_k)\,\hat{f}_o(x_k)$$

Implementation mapping (R): $\hat{f}_r \leftarrow$ `f_r_hat`, $\mathrm{Var}_r \leftarrow$ `var_r_hat`, $\hat{f}_o \leftarrow$ `f_o_hat`, $\lambda_{\mathrm{URE}} \leftarrow$ `lambda_URE`, $\lambda_2(x) \leftarrow$ `lambda2_vec`, $\hat{f}_c \leftarrow$ `f_combined`, $\hat{f}_{c2} \leftarrow$ `f_combined2`.

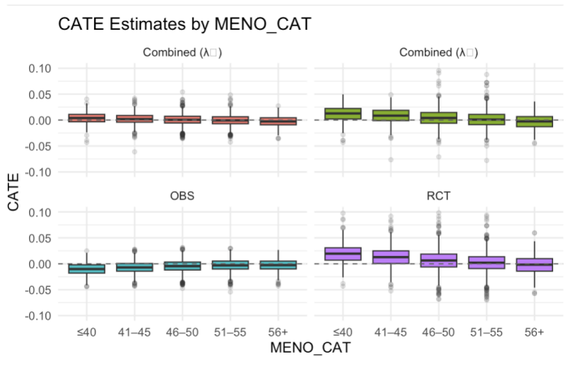
CATE Distributions Across Methods

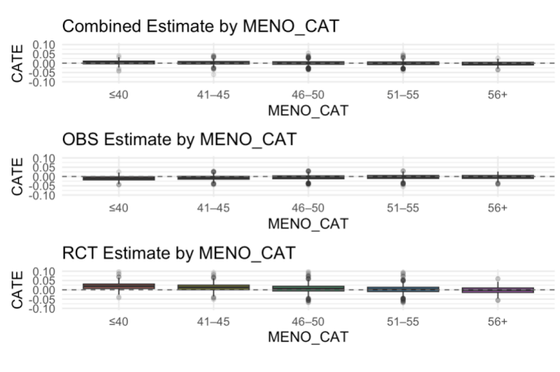


Global Shrinkage Pulls Estimates Together



Subgroup Evaluation: CATE by `MENO_CAT`





References

[1] Rosenman, E. T. R., Basse, G., Owen, A. B., & Baiocchi, M. (2023). *Combining observational and experimental datasets using shrinkage estimators*. *Biometrics*, 79(4), 2961–2973. https://doi.org/10.1111/biom.13827