Efficient Estimation of Causal Effects Under Two-Phase Sampling with Error-Prone Outcome and Treatment Measurements

Keith Barnatchez [1] Kevin P. Josey [2] Nima S. Hejazi [1] Bryan E. Shepherd [3] Giovanni Parmigiani [1] Rachel C. Nethery [1]

[1]Harvard T.H. Chan School of Public Health [2]Colorado School of Public Health [3]Vanderbilt University Medical Center

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

Motivating Application: Error-Prone EHR data

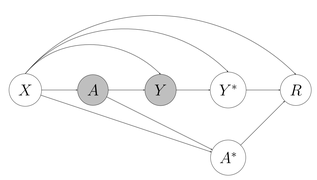
- The Vanderbilt Comprehensive Care Clinic (VCCC) maintains an EHR database with $\approx$ 1300 people living with HIV receiving care from the VCCC
- Substantial error in key variables, including occurrences of AIDS-defining events (ADEs) and dates of antiretroviral therapy (ART) initiation
- **Causal estimand**: Average causal effect of beginning ART within 1 month of first visit on 3-year ADE risk
- **Key problem**: Both the outcome and treatment of interest are measured with error

Two-Phase Sampling to Address Measurement Error

Outcomes and treatments stored in EHR data are often measured with substantial error

In many settings, it is possible to **validate** a random subset of error-prone observations

Y	Y^*	A	A^*	X	R
Y_1	Y_1^*	A_1	A_1^*	X_1	1
	Y_2^*		A_2^*	X_2	0
	Y_3^*		A_3^*	X_3	0
Y_4	Y_4^*	A_4	A_4^*	X_4	1
	Y_5^*		A_5^*	X_5	0



In practice, validated subjects are often selected according to a **sampling rule** which depends on all initially-observed data: $\boldsymbol{X}, Y^*$ and A^*

Goal: Construct semiparametric efficient estimators of counterfactual means $\mathbb{E}[Y(a)]$

Challenge: Validation datasets are typically small in practice – sources of finite-sample instability can play a prominent role in estimation

Assumptions

1. Consistency: $Y = AY(1) + (1-A)Y(0)$
2. Treatment positivity: $\mathbb{P}(A = 1|\boldsymbol{X}) \in (0, 1)$
3. No unmeasured confounding: $(Y(1), Y(0)) \perp\!\!\!\perp A|\boldsymbol{X}$
4. A and Y missing-at-random: $(A, Y) \perp\!\!\!\perp R|\boldsymbol{Z}$, where $\boldsymbol{Z} = (\boldsymbol{X}, A^*, Y^*)$
5. Validation positivity: $\mathbb{P}(R = 1|\boldsymbol{Z}) \in (0, 1)$

Our Contributions

- We present two asymptotically equivalent, semiparametric efficient one-step estimators
- We document unique sources of finite-sample instability faced by each estimator
- We present modifications to improve finite-sample behavior, and construct an ensemble estimator designed to prioritize finite-sample efficiency
- We developed the R package `drcmd`, which implements the Approach 2 estimator in *general* two-phase sampling and missing data settings

References

[1] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

We connect two **asymptotically equivalent** approaches for constructing **semiparametric efficient** one-step estimators, and propose an **ensemble** estimator that optimizes finite sample efficiency.

Approach 1: Observed Data Distribution

High-level idea: Follow the standard semiparametric statistics pipeline

$$\underset{\text{causal estimand}}{\mathbb{E}[Y(a)]} \xrightarrow[\text{Assumptions}]{} \underset{\text{stat. estimand}}{\mathbb{E}[f(\text{data})]} \longrightarrow \underbrace{\hat{\psi}_a^{\text{PI}} = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(\text{data}_i)}_{\text{plug-in estimator}} \xrightarrow[\text{(tedious) work}]{} \underset{\text{debiased est.}}{\hat{\psi}_a^{\text{OS}}}$$

We show under Assumptions 1-5,

$$\mathbb{E}[Y(a)] = \psi_{a,1} = \mathbb{E}\left[\frac{\mathbb{E}\{\lambda_a(\boldsymbol{Z}) \cdot \mu_a(\boldsymbol{Z})|\boldsymbol{X}\}}{\mathbb{E}\{\lambda_a(\boldsymbol{Z})|\boldsymbol{X}\}}\right]$$

where $\lambda_a(\boldsymbol{Z}) = \mathbb{P}(A = a|\boldsymbol{Z}, R = 1)$ and $\mu_a(\boldsymbol{Z}) = \mathbb{E}(Y|A = a, \boldsymbol{Z}, R = 1)$ are imputation models for A and Y. Suggests the **one-step** estimator

$$\hat{\psi}_{a,1}^{\text{OS}} = \hat{\psi}_{a,1}^{\text{PI}} + \frac{1}{n}\sum_{i=1}^{n} \widehat{\text{EIF}}(\psi_{a,1}, O_i).$$

Approach 2: Complete Data Distribution

High-level idea: [1] With complete data, could construct standard AIPW estimator

$$\hat{\psi}_{a,\text{C}}^{\text{PI}} = \frac{1}{n}\sum_{i=1}^{n} \hat{m}_a(\boldsymbol{X}_i) \quad \text{and} \quad \hat{\psi}_{a,\text{C}}^{\text{OS}} = \hat{\psi}_{a,\text{C}}^{\text{PI}} + \frac{1}{n}\sum_{i=1}^{n} \underbrace{\left(\hat{m}_a(\boldsymbol{X}_i) + \frac{I(A=a)}{\hat{g}_a(\boldsymbol{X}_i)}\{Y - \hat{m}_a(\boldsymbol{X}_i)\} - \hat{\psi}_{a,\text{C}}^{\text{PI}}\right)}_{:=\chi_a(\boldsymbol{O};\hat{m}_a,\hat{g}_a)}$$

where $m_a(\boldsymbol{X}) = \mathbb{E}(Y|A = a, \boldsymbol{X})$ and $g_a(\boldsymbol{X}) = \mathbb{P}(A = a|\boldsymbol{X})$ can be fit with weighted regressions that add weights $R/\hat{\mathbb{P}}(R = 1|\boldsymbol{Z})$ to the underlying loss functions.

Treating $\chi_a(O; \hat{m}_a, \hat{g}_a)$ as pseudo-outcomes,

$$\hat{\psi}_{a,2}^{\text{OS}} = \hat{\psi}_{a,\text{C}}^{\text{PI}} + \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\varphi}_a(\boldsymbol{Z}_i) + \frac{R_i}{\hat{\mathbb{P}}(R_i = 1|\boldsymbol{Z}_i)}\{\chi_a(\boldsymbol{O}_i; \hat{m}_a, \hat{g}_a) - \hat{\varphi}_a(\boldsymbol{Z}_i)\}\right\}$$

where $\varphi_a(\boldsymbol{Z}) = \mathbb{E}[\chi_a(\boldsymbol{O}; m_a, g_a)|\boldsymbol{Z}, R = 1]$, is an efficient one-step estimator.

Properties

Under Assumptions 1-5 and standard regularity conditions, we show that $\hat{\psi}_{a,1}^{\text{OS}}$ and $\hat{\psi}_{a,2}^{\text{OS}}$ are asymptotically equivalent and semiparametric efficient

- Approach 2 can be viewed as a reparametrization of Approach 1

Both approaches have numerous unique sources of finite-sample instability

- Approach 1: debiasing term introduces numerous multiplicative, unstable weighting terms and requires estimation of 6 nuisance functions
- Approach 2: estimation of $\varphi_a(\boldsymbol{Z})$ is an inherently difficult task in small samples

We propose estimating $\varphi_a(\boldsymbol{Z})$ to minimize the **empirical variance** of $\hat{\psi}_a^{\text{OS},2}$

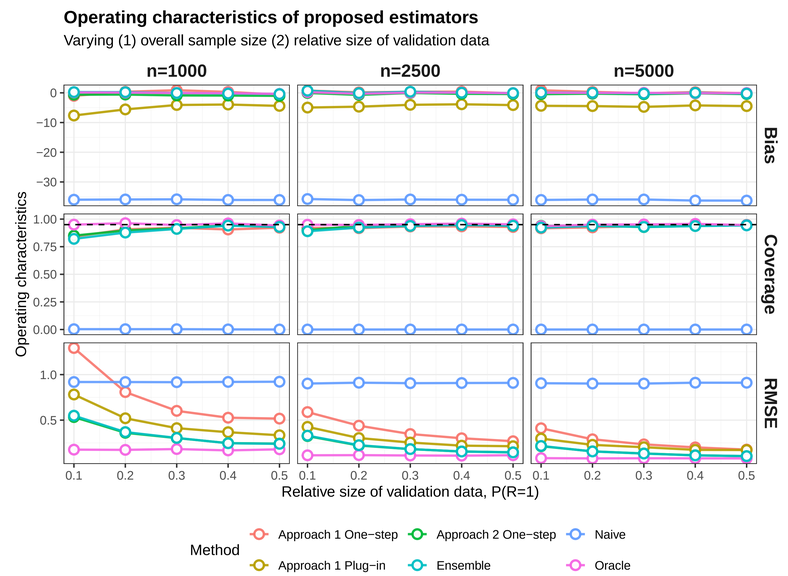
Ensemble Estimator

High-level idea: $\hat{\psi}_{a,1}^{\text{OS}}$ and $\hat{\psi}_{a,2}^{\text{OS}}$ can differ substantially in finite samples

$$\hat{\psi}_{a,\text{E}}^{\text{OS}} = \hat{w} \cdot \hat{\psi}_{a,1}^{\text{OS}} + (1 - \hat{w}) \cdot \hat{\psi}_{a,2}^{\text{OS}},$$

where $\hat{w}$ is chosen in a manner that (i) minimizes finite-sample variance, and (ii) remains **well-defined** asymptotically

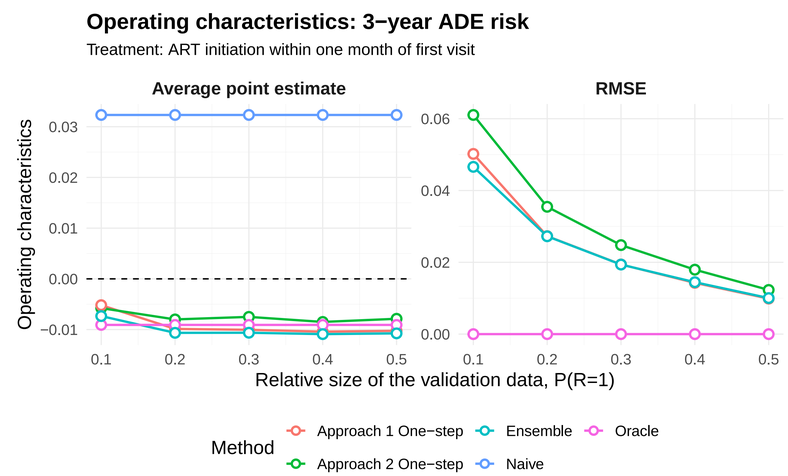
Simulation



Real Data: Vanderbilt Comprehensive Care Clinic (VCCC)

Team of researchers validated **every** observation in VCCC database

- Revealed increasingly larger shares of validated data, in manner which depends on $\boldsymbol{Z}$



- High rates of measurement error in ADEs ($\approx 12.5\%$) and early ART (4.1%) leads to large **bias** in estimate of ATE
- Ensemble estimator achieves **lowest RMSE** for all validation proportions considered

https://www.kbarnatchez.com https://www.github.com/keithbarnatchez/drcmd keithbarnatchez@g.harvard.edu