Assessing treatment effects in observational data with missing confounders: a comparative study of practical doubly-robust and traditional missing-data methods

Brian D. Williamson, PhD

Kaiser Permanente Washington Health Research Institute
brian.d.williamson@kp.org

WNAR 2025: Whistler, BC, Canada
June 2025

Disclaimer

This work was supported by Task Order **75F40122F19010** under Master Agreement **75F40119D10037** from the U.S. Food and Drug Administration (FDA).

The views expressed in this presentation represent those of the presenter and do not necessarily represent the official views of the U.S. FDA.

Acknowledgments

This presentation features work from the Sentinel CI4 Subset Calibration Working Group:

Kaiser Permanente Washington Health Research Institute

Eric Johnson
Chloe Krakauer
Jen Nelson
Pamela Shaw (PI)
Susan Shortreed
Gregory Simon
Chris Stewart
Brian Williamson

FDA

Sarah Dutcher
Noah Argual
Jose Hernandez
Hana Lee
Mingfeng Zhang
Fengyu Zhao
Jummai Apata
Lucia Menegussi
Yan Li

Sentinel Operations Center

John Connolly
Christine Halbig
Meighan Rogers Driscoll
Katherine Farineau

Brigham and Women's Hospital/Harvard Medical School

Rishi Desai
Richard Wyss

TL Revolution/ UC Berkeley

Susan Gruber
Mark van der Laan

University of Auckland

Thomas Lumley

Vanderbilt University

Bryan Shepherd

University of Washington

James Floyd

Motivation

Important confounders or covariates are often missing in pharmacoepidemiologic studies.

- In FDA's Sentinel Innovation Center:
  - sources link data from administrative claims and electronic health records (EHR)
  - certain clinical data (e.g., vital signs) are not available on individuals with claims data only
- More generally:
  - Invasiveness (e.g., biopsy)
  - Expense (e.g., novel biomarkers)

Prior work within Sentinel Innovation Center:

- Investigated inverse probability weighting (IPW), complete case, missingness indicator, single imputation (missForest), MICE (default), MICE (CART), MI-RF (random forest)
- Considered 10–50% missing data, under MCAR, MAR, MNAR
- Found MICE (default) and MI-RF had best overall performance

Motivation

Building on prior work, we sought to:

- Compare MICE and MI-RF to two more robust methods
    - generalized raking, commonly used in surveys or two-phase samples
    - targeted maximum likelihood estimation, allowing machine learning
- Consider performance of methods under higher levels of missingness: 40–80%
- Create synthetic data scenarios that challenge each method
- Investigate performance in a realistic plasmode scenario

KPWA Antidepressant Initiator (ADI) cohort

Kaiser Permanente Washington (KPWA) is an integrated health care system in the Pacific Northwest that provides care and health insurance to over 700,000 members.

Our sample:

- 112,770 adults at KPWA aged 13+ years, initiating antidepressant medication or psychotherapy from January 1, 2008 to December 31 2018
- Key confounders:
  - 9-item Patient Health Questionnaire (PHQ-9)
  - First 8 items (PHQ-8) measure depressive symptoms
  - Ninth item (PHQi9) measures suicidal ideation
- 55% missing the PHQ-9 (50,337 individuals with complete data)
- Outcome: composite outcome of self-harm (fatal or non-fatal) or psychiatric hospitalization within 5 years following treatment initiation (5193 events, 10.3%)

Missing data as a two-phase design

Phase 1 (always observed): Measure (X_i, Z_i, Y_i) for $i = 1, \ldots, N$ subjects

- X_i: binary treatment
- Z_i: confounders
- Y_i: binary outcome

Phase 2: Variables only available on a subset

- W_i: additional confounders (e.g., PHQ-8, PHQi9)

Missing data as a two-phase design

Goal: estimate a treatment effect, e.g.,

- average treatment effect (ATE): $E\{E(Y \mid X = 1, Z, W)\} - E\{E(Y \mid X = 0, Z, W)\}$
- conditional log odds ratio: parameter β_1 from model

$$\text{logit}P(Y = 1 \mid X = x, Z = z, W = w) = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 w$$

Traditional two-phase study design: individuals selected into phase 2 with sampling probability π_i

Missing data setting: π_i are not known and need to be estimated

Key functionals:

- Outcome regression: $Q \equiv Q(x, z, w) = E(Y \mid X = x, Z = z, W = w)$
- Missing-data model: setting $\Delta = 1$ if observed, $\pi \equiv \pi(x, y, z) = P(\Delta = 1 \mid X = x, Y = y, Z = z)$
- Propensity score model: $g \equiv g(z, w) = P(X = 1 \mid Z = z, W = w)$

Choosing an estimand

A key step in any data analysis is choosing a target of estimation (estimand).

Treatment-specific mean outcome values:

$$\begin{aligned}\mu_1 &= E\{E(Y \mid X = 1, W, Z)\}\\ \mu_0 &= E\{E(Y \mid X = 0, W, Z)\}\end{aligned}$$

Estimands we consider:

- marginal risk difference (mRD): $\mu_1 - \mu_0$
- marginal relative risk (mRR): μ_1/μ_0
- marginal odds ratio (mOR): $\frac{\mu_1/(1-\mu_1)}{\mu_0/(1-\mu_0)}$
- conditional odds ratio (cOR): regression parameter from model

$$\operatorname{logit} P(Y = 1 \mid X = x, W = w, Z = z) = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 w$$

Choosing an estimand

Viewing the cOR as resulting from a working model allows us to define two levels of estimand.

Oracle estimand:

- a function of only the data-generating distribution
- cOR is only an oracle estimand if the working model is identical to the true data-generating model
- natural target of many causal inference methods

Census estimand:

- a function of the data-generating distribution and working regression model
- cOR always defined at this level
- natural target of many parametric regression methods
- missing-data methods often benchmarked against this estimand

Importantly, oracle and census estimands are only equal if the working model corresponds exactly to the data-generating model.

Common approaches to handle missing confounders

Complete case analysis (CC):

- drop individuals with missing information for some variables.
- fit working regression model (targeting census estimand)
- But this can lead to bias and inefficiency!

Confounded approach (CNFD):

- Drop variables prone to missingness
- fit regression model
- But this can lead to confounder bias! (for the census estimand)

Horvitz-Thompson inverse probability weighted estimator (IPW):

- Estimate π
- fit weighted working regression model among complete cases (targeting census estimand)
- Typically unbiased, but can be inefficient

Multiple imputation (MI):

- Use partial information to impute missing data
- fit working regression model (targeting census estimand)
- Under missing at random (MAR), can avoid bias and inefficiency

Generalized raking: improving IPW

Procedure: (Deville and Särndal, 1992)

1. calculate IPW weights
2. calibrate IPW weights using phase 1 variables
   - optimal calibration: based on expected value of influence function given observed data (Breslow et al., 2009)
   - in practice, can be based on working model
3. fit weighted outcome regression using calibrated weights

Generalized raking (GR):

- is equivalent to optimal augmented IPW estimator if using optimal calibration variable (Lumley et al., 2011)
- can be easily implemented in R package `survey`
- is doubly-robust to misspecification of outcome model or missing-data model

Targeted maximum likelihood estimation (TMLE)

Developed to target oracle marginal quantities (e.g., mRD) under less-restrictive assumptions. (van der Laan and Rubin, 2006)

Requirements:

1. Estimating Q, g, π (Rose and van der Laan, 2011)
2. Use maximum likelihood to solve score equation involving efficient influence function

Appealing properties:

1. Can use machine learning for nuisance parameters (e.g., super learner (van der Laan et al., 2007))
2. Doubly-robust for outcome regression and propensity score

Implemented in the R package `twoStageDesignTMLE`.

Robustness

Suppose working outcome model Q, missing-data model π, imputation model f, treatment assignment model g

Analysis approach	Required models for estimation	Correct specification for consistent estimation of oracle parameters	Correct specification for consistent estimation of census parameters
IPW	Q and π	Q and π	π or (Q and CD-MCAR)
MI	Q and f	Q and f	f
GR	Q and π	Q or π	π or Q
TMLE	Q, g and π	(Q or g) and π	π or [(Q or g) and CD-MCAR]

CD-MCAR: covariate-dependent missing completely at random, where missing data can depend only on always-observed covariates, not outcomes (Seaman et al., 2013)

Numerical experiments

We consider both synthetic and plasmode experiments.

Synthetic:

- specify all data-generating models
- investigate performance of estimators in range of settings

Plasmode:

- data-generating models based on KPWA cohort
- bootstrap sampling of covariates
- investigate performance of estimators in a more realistic setting

Synthetic data generation

Common parameters:

- Cohort size $n = 10000$
- 2500 Monte-Carlo replications

Base case scenario:

- simple data-generating models for outcome, treatment, missingness
- roughly 40% missingness in confounders
- roughly 12% outcome incidence
- in this case, oracle estimand = census estimand

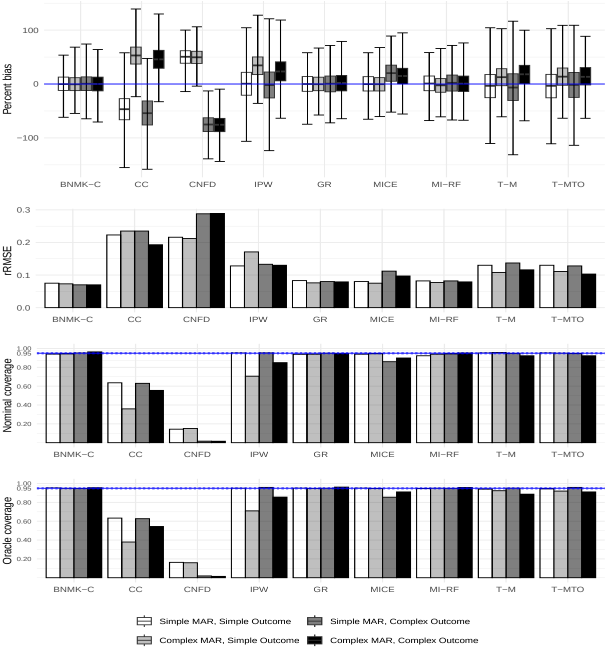
Variations:

- Complex outcome or missingness model (interactions, nonlinear terms)
- Increased missingness (80%)
- Lower outcome incidence (5%)

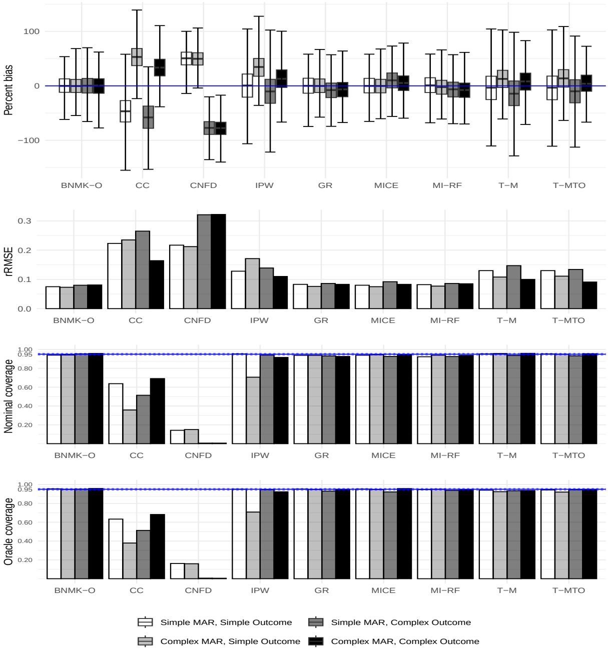
Estimators

1. Oracle model: true outcome model fit with complete data for entire cohort (benchmark for oracle estimand)
2. Census model: working outcome model fit with complete data for entire cohort (benchmark for census estimand)
3. Confounded model (CNFD): fit working model dropping variables with missing data
4. Complete-case (CC): fit working model dropping observations with missing data
5. Inverse probability weighting (IPW): fit working model with weights obtained using logistic regression
6. Generalized raking (GR): fit working model with calibrated weights
7. MICE: fit working model after multiple imputation via chained equations (MICE)
8. MI-RF: fit working model after MI using random forests
9. TMLE:
   - TMLE-M: use super learner to estimate π, but working models for Q and g
   - TMLE-MTO: use super learner to estimate π, Q, g

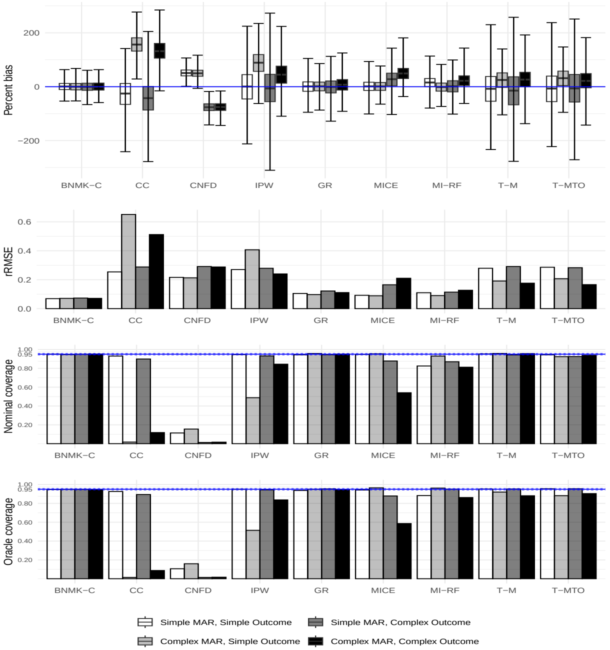
Synthetic base case simulation results: census cOR



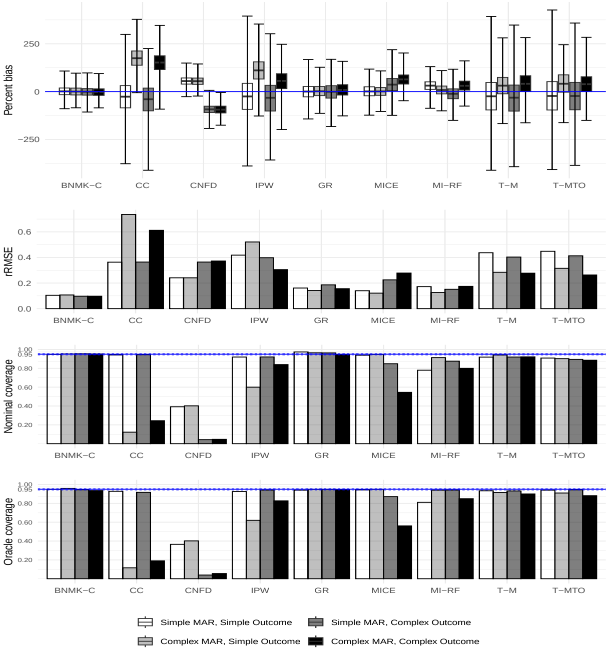
Synthetic base case simulation results: oracle cOR



80% missing-data results: census cOR



80% missing-data, 5% outcome results: census cOR



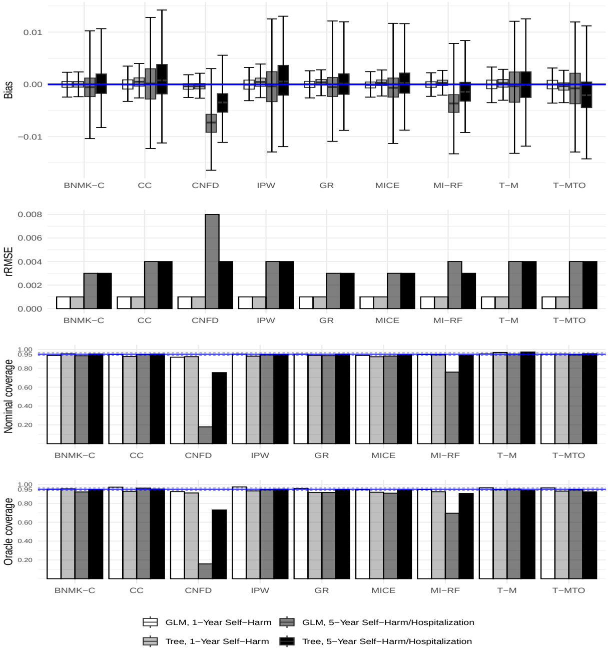
Plasmode data generation

Data-generating models:

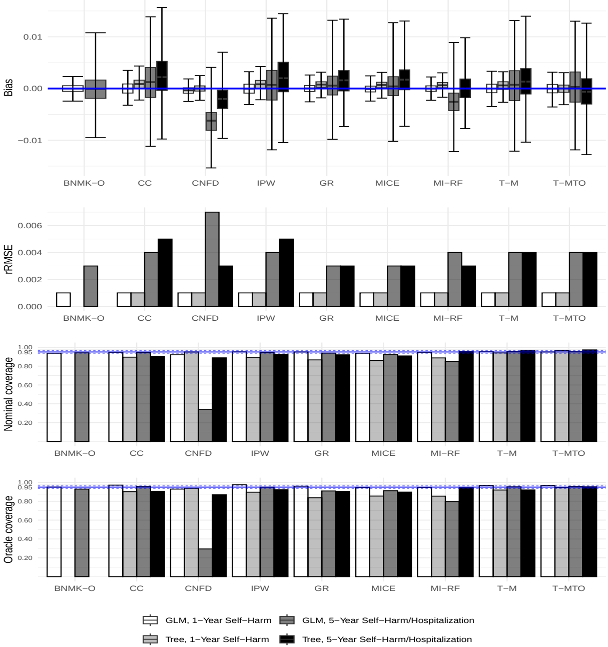
- π: tree or glm fit to entire cohort ($n = 112770$) to predict observation of PHQ-9
- g: tree or glm fit to complete cohort ($n = 50337$) to predict assignment to psychotherapy vs antidepressant medication
- Q: tree or glm fit to complete cohort

Other confounders: sex, age, Charlson comorbidity index, anxiety diagnosis in past year, self-harm in prior 6 months, psychiatric hospitalization in prior 5 years, alcohol use disorder in past year

Plasmode simulation results: census mRD



Plasmode simulation results: oracle mRD



Conclusions

In our simulations, we observed:

- the importance of first choosing a target estimand, then an estimation procedure
- confounded and complete-case methods performed poorly
- at least one variant of MI often performed well, but no one variant performed well uniformly
- generalized raking among the best in the majority of settings
- TMLE often had small bias but larger variance

For more, check out:

- the paper, `https://arxiv.org/abs/2412.15012`
- the repo, `https://github.com/PamelaShaw/Missing-Confounders-Methods`

`https://github.com/bdwilliamson`
`https://bdwilliamson.github.io`
✉ `brian.d.williamson@kp.org`