Assessing Treatment Effects in Observational Data with Missing Confounders: A Comparative Study of Practical Doubly-Robust and Traditional Missing Data Methods

Pamela Shaw
Division of Biostatistics
Kaiser Permanente Washington Health Research Institute
Pamela.A.Shaw@kp.org

VICBiostat Seminar

June 26, 2025

KAISER PERMANENTE®

Disclaimer

- This work was supported by Task Order **75F40122F19010** under Master Agreement **75F40119D10037** from the U.S. Food and Drug Administration (FDA) and National Institutes of Health (NIH) grant R01-AI131771.

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

Acknowledgements

This work is being done by 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 (Lead)
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

Outline

- Introduction
- Methods
- Numerical Study: Synthetic data
- Numerical Study: Plasmode data
- Software and Vignettes
- Discussion

Sentinel Initiative - https://www.Sentinelinitiative.org/

"The U.S. Food and Drug Administration (FDA) leads the Sentinel Initiative. FDA created the Sentinel Initiative to meet a mandate by Congress in the FDA Amendments Act of 2007. Through the Sentinel Initiative, FDA aims to develop new ways to assess the safety of approved medical products including drugs, vaccines, and medical devices."

Subset Calibration Methods Project

The aim of this project was to evaluate and compare methods to address missing data in settings with high levels of missing confounder data

- Estimators for two-phase data were of interest
  - Survey Calibration methods
  - Two stage Targeted Maximum Likelihood Estimation (TMLE)
- Comparison with more traditional methods to evaluate missing data was also of interest
  - Inverse-probability weighting (IPW)
  - Multiple imputation
  - Naïve approaches
- Practical approaches currently available in statistical software were a focus
- Project built on the learnings of previous Sentinel Project that compared imputation methods to handle missing confounder data

Weberpals J, Raman SR, Shaw PA, Lee H, Russo M, Hammill BG, Toh D, Connolly JG, Dandreo KJ, Tian F, Liu W, Li J, Hernández-Muñoz JJ, Glynn RJ, Desai RJ. A Principled Approach to Characterize and Analyze Partially Observed Confounder Data From Electronic Health Records: A Plasmode Simulation Study. *Clinical Epidemiology*, 2024; 16 329–343.

Previous Work of Weberpals et al 2024

Approaches to Handling Partially Observed Confounder Data from Electronic Health Records (EHR) in Non-Randomized Studies of Medication Outcomes

Overall Goal: Develop standardized "toolkits" that can be readily implemented in EHRs to diagnose and, when assumptions permit, address missingness in confounding variables in pharmacoepidemiologic analyses

Key component of this work: Methods comparison

- Investigated: IPW, complete case, missingness indicator, single imputation (missForest), MICE (default), MICE (CART), MICE-RF (random forest)
- Considered 10-50% missingness imputation procedures
- Considered different missingness mechanisms: MCAR, MAR, MNAR
- Found MICE (default) and MICE-RF (random forest) had best overall performance
- Developed SDMI R package to help diagnose type of missing data (Weberpals et al JAMIA open 2024)

Gaps in previous work: lack of doubly-robust methods

Building on this work, we sought to:

- Evaluate how the best performing imputation approaches from Weberpals et al 2024, MICE and MICE-RF, compare to survey calibration and TMLE-based super learner
    - Hypothesis 1: there would be advantages in some settings to survey calibration because it does not need to correctly model outcome in subjects with incomplete data
    - Hypothesis 2: there would be advantages to TMLE-super learner approaches because of the flexible (e.g. non-parametric) approach for estimation
- Consider performance of methods under higher levels of missingness: 40-80%
- Investigate the practicality of implementing survey calibration and TMLE

Project Goals

1. Perform a numerical simulation study to assess relative performance of methods for handling high percentage of missingness in confounders
    - Use synthetic data to create different scenarios that challenge the methods
    - Use plasmode simulation to create a realistic simulation based on real data
    - Ultimate aim: provide guidance on choice of analytical approach

2. Disseminate knowledge of the survey calibration and TMLE super learner as methods for handling missing data

3. Provide vignettes and software to allow easy adoption of methods by members of Sentinel and the broader research community

Setting of Interest

- When evaluating treatments for safety and rare outcomes, it is often necessary to merge multiple, large databases
  - Improves statistical power
  - Create a more generalizable cohort

- A commonly encountered problem in pharmacoepidemiologic settings: important confounders may be missing in a high % of individuals
  - In Sentinel, where data from administrative claims and EHR data are combined, certain clinical data (e.g. vital signs) are not available on individuals with claims data only

- More generally, there are many settings in observational and randomized cohorts where certain covariates may only be obtained on a subset either due to invasiveness (e.g. biopsy) or expense (novel biomarkers) of assays

KPWA antidepressant Initiator (ADI) Cohort

- Kaiser Permanente Washington (KPWA) is an integrated health care system in Pacific Northwest that provides care and health insurance to over 700,000 members
- 112,770 KPWA adults aged 13+ years, initiating antidepressant medication or psychotherapy from January 1, 2008 to December 31 2018 (n=112,770)
  - No antidepressant fills or psychotherapy in the prior year
- **Plasmode data set: 50,337 individuals** with complete data on the Patient Health Questionnaire (PHQ-9)
- **Outcome:** Composite outcome of self-harm (fatal or non-fatal) or psychiatric hospitalization within 5 years following treatment initiation n=5193, (10.3%)
- Missingness: 55% missing the PHQ-9
  - Missing data includes key confounders: depression severity and history of prior self-harm

Missing data: A two-phase design

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

X_i – binary treatment

Z_i – Confounders

Y_i – binary outcome

Phase 2: Variables only available on a subset (W_i)

W_i – Additional confounders available for some individuals

Phase 2 study design: Individual selected into phase 2 with sampling probability π_i

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

Common approaches to handling missing confounders

Complete Case Analysis (CC)- Drop individuals with missing information for some variables

Problem #1: dropping observations with partial information can lead to inefficiency

Problem #2: dropping observations with partial information can lead to bias

Confounded Approach (CNFD) – Drop variables prone to missingness

Problem #3: dropping variables can lead to confounder bias

IPW – Horwitz Thompson inverse probability weighted estimator

Problem #1: dropping observations with partial information can lead to inefficiency

Multiple Imputation (MI) – Use partial information to impute missing data and analyze all individuals with any data

Can avoid problems 1,2,3 under MAR (Little et al 2022)

Improving IPW

- IPW is only using information from the complete case data
  - Ignoring the partially observed data can result in inefficient estimators (wide CI)
- Multiple imputation (MI) uses the partially observed data to impute missing information and incorporate all individuals into final analysis
  - Weberpals et al. *Clin Epi* 2024 demonstrated MI estimators more efficient (narrower CI)
- Han et al SIM 2021 demonstrated that the efficiency of the MI comes at a cost
  - Efficiency that you gain is robustness that you lose (Lumley 2017)
- **Survey calibration, or generalized raking (GR)** is another way to bring in information from individuals with partially observed data
  - GR estimators will be more efficient than IPW if there is at least some "linear correlation" between observed data and unobserved data
  - Unlike MI, GR does not rely on getting outcome model correct in unobserved individuals

Survey Calibration, aka Generalized raking (GR)

GR is a weighted complete case estimator using calibrated IPW weights

- First step: Calculate the IPW weights ($1/\pi_i$)
- Second step: weights are adjusted (calibrated) using "phase 1" variables observed on everyone
  - Calibration variables can be imputed variables using a working model
  - If working model is wrong, raking won't gain efficiency over IPW, but won't introduce bias
- Third step: Fit weighted outcome regression model using calibrated weights

Some observations

- Ideal raking estimator is asymptotically equivalent to optimal AIPW estimator (Lumley et al 2011)
  - Ideal raking leverages maximal information from the ''phase 1" data
- GR targets estimand/estimate that would have been fit if there was no missing data
- Raking can be readily implemented using the survey package in R (Lumley 2011)

  See example code: https://github.com/PamelaShaw/Missing-Confounders-Methods/

GR: Technical Bits (Deville and Särndal 1992)

Let A_i be a vector of auxiliary variables available on everyone at phase one. Then g_i is constructed to

$$\text{minimize} \sum_{i=1}^{N} R_i d\left(\frac{g_i}{\pi_i}, \frac{1}{\pi_i}\right) \text{ subject to } \sum_{i=1}^{N} A_i = \sum_{i=1}^{N} R_i \frac{g_i}{\pi_i} A_i$$

Under mild regularity conditions, know

$$\text{Var}\left(\hat{\beta}_{GR}\right) \approx (1-\rho^2)\text{Var}\left(\hat{\beta}_{IPW}\right), \text{ where } \rho = \text{cor}(h_i, A_i) \text{ and}$$

h_i is the efficient influence function of the target parameter in the population model

Constructing Efficient Raking (Calibration) Variables for Regression Coefficients

- We know the ideal raking variable is the expected value of the influence function given the observed data (Breslow et al 2009)
- Han et al 2021 showed a practical way to estimate the optimal raking variable is to
  - Multiply impute the missing data *for the whole cohort*
  - Fit the target working model on the population for each imputed dataset and obtain the influence functions for each model parameter
  - Average the influence functions across imputed datasets
- **GR Approach:** calibrate IPW weights using the average imputed influence functions for each of the regression coefficients

TMLE Superlearner

- Targeted Maximum likelihood estimation (TMLE) is a semiparametric estimation technique that targets a parameter of interest
- Contrary to maximum likelihood estimation (MLE) which solves an optimization problem based on all parameters, TMLE aims to reduce bias and variance for parameter of interest at expense of other parameters (nuisance parameters)
- Typically implemented with a Super Learner (data adaptive ensemble learner) for key quantities
  - Learning by fitting multiple models to the data
  - Incorporates the flexibility of machine learning, while allowing for statistical inference
  - Will be maximally efficient when the model and nuisance parameters are correctly specified

- Rose and van der Laan (2011) developed a TMLE approach to handle two-phase designs
- Machine learning principles can be used to flexibly model missing data

Key references: van der Laan and Rubin 2006; van der Laan et al 2007; Rose and van der Laan 2011, Gruber and van der Laan 2009

TMLE targets the causal estimand

Define μ_x = P(Y=1 | X=x)

In our point treatment setting, interest is in one of the following estimands:

Marginal risk difference (mRD): $\mu_1 - \mu_0$

Marginal relative risk (mRR): μ_1/μ_0

Marginal odds ratio (mOR): $\frac{\mu_1/(1-\mu_1)}{\mu_0/(1-\mu_0)}$

Our TMLEs will estimate the treatment effect by estimating: E[P(Y=1 | X=x, Z, W)]

TMLE for Missingness Mechanism (TMLE-M)

Table 4: Algorithm used for each nuisance function (missing-data model, propensity score, and outcome regression) in the TMLE-M procedure, along with the candidate algorithms in any Super Learners. Unspecified tuning parameters are set to their default values. Sample size is denoted by n and number of covariates by p.
*sequence contains 10 possible values.

Nuisance function	Algorithm	Tuning parameters
Outcome regression	glm	—
Propensity score	glm	—
Missing-data model	Super Learner	
	glm	—
	gradient boosted trees (`xgboost`)	maximum depth $\in \{1, 3\}$ shrinkage $\in \{.01, .1\}$ number of trees $= 500$
	random forests (`ranger`)	minimum node size $\in \{n/100, \ldots, n/10\}^*$ number of trees $= 500$ mtry $= \sqrt{p}$

TMLE for Missing/Treatment/Outcome (TMLE-MTO)

Table 5: Algorithm used for each nuisance function (missing-data model, propensity score, and outcome regression) in the TMLE-MTO procedure, along with the candidate algorithms in any Super Learners. Sample size is denoted by n and number of covariates by p.
*sequence contains 10 possible values.

Nuisance function	Algorithm	Tuning parameters
Outcome regression	Super Learner	
	glm	—
	gradient boosted trees (`xgboost`)	maximum depth $\in \{1, 3\}$ shrinkage $\in \{.01, .1\}$ number of trees $= 500$
	random forests (`ranger`)	minimum node size $\in \{n/100, \ldots, n/10\}^*$ number of trees $= 500$ mtry $= \sqrt{p}$
Propensity score	Super Learner	
	glm	—
	gradient boosted trees (`xgboost`)	maximum depth $\in \{1, 3\}$ shrinkage $\in \{.01, .1\}$ number of trees $= 500$
	random forests (`ranger`)	minimum node size $\in \{n/100, \ldots, n/10\}^*$ number of trees $= 500$ mtry $= \sqrt{p}$
Missing-data model	Super Learner	
	glm	—
	gradient boosted trees (`xgboost`)	maximum depth $\in \{1, 3\}$ shrinkage $\in \{.01, .1\}$ number of trees $= 500$
	random forests (`ranger`)	minimum node size $\in \{n/100, \ldots, n/10\}^*$ number of trees $= 500$ mtry $= \sqrt{p}$

Numerical Study Part 1: Synthetic Data

Base Case Simulation Set up

- Cohort of N= 10,000 generated
- 2500 Monte Carlo simulation iterations
- Roughly 40% missingness in confounders
- Outcome ~12% incidence
- Compare performance of estimation methods for 4 estimands
  - Conditional odds ratio (cOR)- i.e., canonical parameter from logistic regression model
  - Marginal risk difference (mRD)
  - Marginal risk ratio (mRR)
  - Marginal odds ratio (mOR)
- Consider the mean bias, median bias, ESE, MAD, RMSE, Oracle coverage, Nominal coverage
- **Base Scenario**: Simple data generating models (DGMs) for outcome, propensity and missingness
- **Variations:** Complex outcome and missingness models, increased missingness (80%), lower outcome incidence (5%)

Estimators Under Study

1. **Oracle model** – true outcome model fit with complete data for whole cohort (unobserved, Benchmark model for Oracle estimand)
2. **Population model** –working outcome model fit with complete data for whole cohort (unobserved- Benchmark model for Census estimand)
3. **Confounded model (CFND)**- working model dropping confounders with missing data
4. **Complete case (CC)** - working model dropping observations with missing data
5. **Inverse Probability Weighting (IPW)** – working model using CC data and *glm*-derived IP weights
6. **Generalized Raking (GR)** – working model using CC data and calibrated weights
7. **MICE** – working model using Multiple Imputation (MI) with chained equations (MICE)
8. **MI-RF** – working model using MI with random forest
9. **MI-XGB** – working model with MI with gradient boosted trees (XGBoost)
10. **TMLE-M** – working model using CC data and super learner-derived IPW weights
11. **TMLE -MTO** – super learner-derived IP weights, propensity score (PS) weights and outcome model

Leveling the Playing Ground (1)

- Different methods target different estimands, so important to study both
- Conditional vs marginal estimands:
  - Conditional estimands (e.g. cOR) are commonly a target of inference, typically coefficient from a regression model
  - Marginal estimands (e.g. mRD) are commonly the focus of causal inference
- Census vs Oracle estimands:
  - Census estimand is a traditional target of missing data methods. Benchmark model – the working model fit to the unobserved full cohort data
  - Oracle estimand is a scientific ideal, capturing “true” DGM. A fully non-parametric model generally necessary if targeting the oracle estimand
- Common implementations:
  - CC, IPW, MI, GR- targeting census, conditional estimand
  - TMLE-MTO- targeting oracle, marginal estimand

Leveling the Playing Ground (2)

- **Census estimand** specified by working model, so simpler DGM will make it easier to get this right
- **Oracle estimand-** specified by unknown true DGM, need flexibility to get this right-so more complex DGM will favor more flexible methods (e.g. super learners)

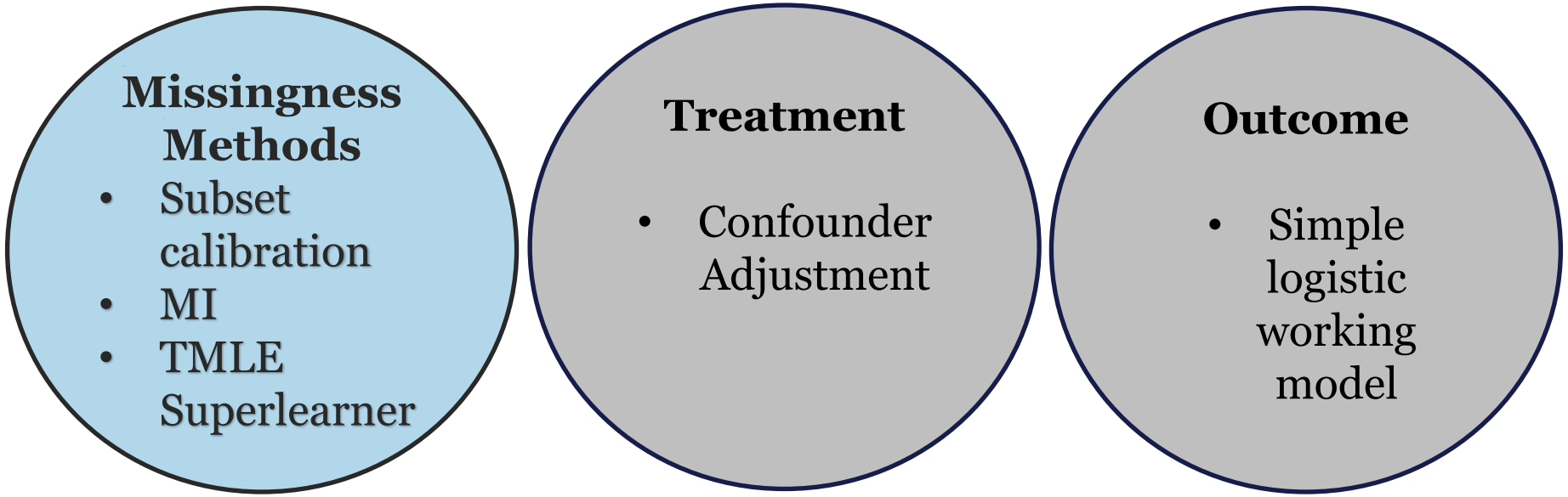
Features of Numerical Simulation Studies

- In Synthetic simulations, we considered both complex and simple DGM
- In Plasmode simulations, we fit very complex DGM to the real data
- Plasmode simulation had complex relationships between covariates, but synthetic simulations had stronger treatment effects and stronger confounding mechanisms, and perhaps stronger patterns of non-linearity than fitted plasmode models

Data Generation Mechanism (DGM) Scenarios

Scenarios	Missingness Model	Treatment Model	Outcome Model	Implications for analysis model
Simple Parametric models No unmeasured confounders in treatment model	1.0 Logistic, simple MAR-XZ **1.1 Logistic, simple MAR-XZY**	linear logistic model • no unobserved confounding	Linear logistic model 1.0- null effect **1.1- cOR=1.5**	Will allow for correct parametric model specification
Complex	MAR: Interactions and non-linear terms MNAR- value MNAR- unobserved variable	N/A	Interactions and non-linear terms MNAR Scenario: unobserved variable	Analysis model simpler than DGM
Plasmode	Complex Tree-based and glm-based implementations for DGM.	Complex Tree-based and glm-based implementations for DGM.	Complex Tree-based and glm-based implementations for DGM.	Analysis model simpler glm than DGM

Analysis Choices



Simple Scenario: Outcome and Treatment DGM

Scenario name	Description	Detailed specification
Missing Scenario 1	Correctly-specified MAR	$\text{logit}\, P(R = 1 \mid X = x, Z = z, Y = y) = \alpha_0 + \alpha_X x + \alpha_Z' z + \alpha_Y y,\ \alpha_Y = 0$
Missing Scenario 1.1	Correctly-specified MAR, depends on Y	$\text{logit}\, P(R = 1 \mid X = x, Z = z, Y = y) = \alpha_0 + \alpha_X x + \alpha_Z' z + \alpha_Y y,\ \alpha_X = \log(2.5),\ \alpha_Z = [\log(1.5), \log(1.5)],\ \alpha_Y = \log(2.5)$

Table 2: Missing-data model scenarios. Unless otherwise specified, $\alpha_0 = -2/3$, $\alpha_X = \log(1.5)$, $\alpha_Z = [\log(2.5), \log(2)]$.

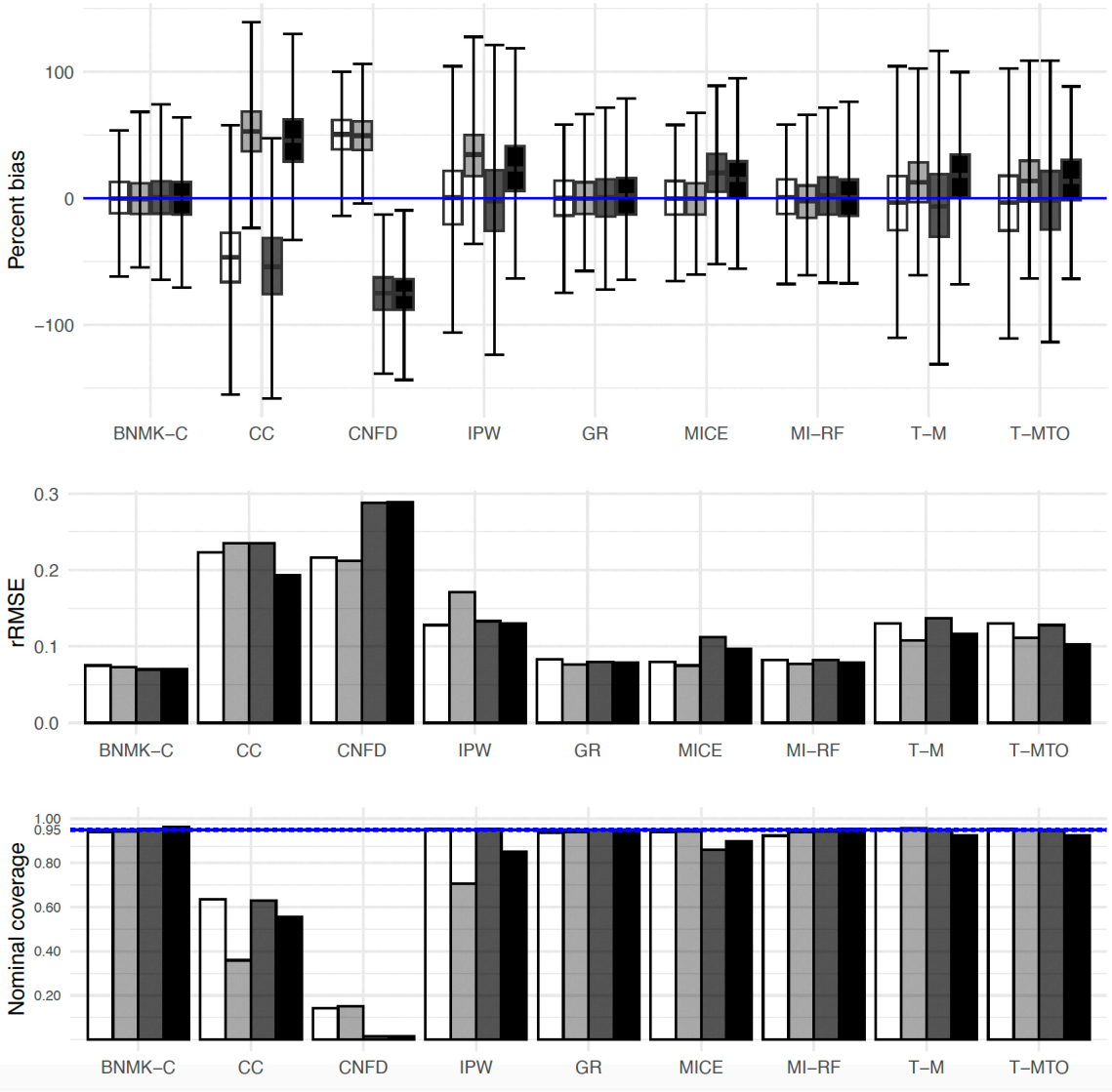
Scenario name	Description	Detailed specification
Y Scenario 1	Simple outcome specification, no treatment effect	$\text{logit}\, P(Y = 1 \mid X = x, W = w, Z = z) = \beta_0 + \beta_X x + \beta_W' w + \beta_Z' z,\ \beta_X = 0$
Y Scenario 1.1	Simple outcome specification, treatment effect	$\text{logit}\, P(Y = 1 \mid X = x, W = w, Z = z) = \beta_0 + \beta_X x + \beta_W' w + \beta_Z' z,\ \beta_X = \log(1.5)$

Table 3: Outcome regression model scenarios. Unless otherwise specified, $\beta_0 = -2.4$, $\beta_W = [\log(1.5), -\log(1.75)]$, $\beta_Z = [\log(1.5), -\log(1.3)]$.

Complex Outcome Model : Nonlinearity + Interactions

Y Scenario 4	Nonlinear dependence and interactions, null X effect	$\text{logit}\, P(Y = 1 \mid X = x, W = w, Z = z) = \beta_0 + \beta_W^\top w + \beta_{W,\text{int}} w_w w_s + \beta_{Z_w,1} I(z_w < -0.5) + \beta_{Z_w,2} I(z_w > 2) + \beta_{Z_s} I(z_s < -1) + \beta_{WZs,\text{inter}} w_s z_s + \beta_{WZw,\text{inter}} w_s I(z_w > 2)$, $\beta_X = 0$, $\beta_{W,\text{int}} = 1$, $\beta_{Z_w,1} = 0.1$, $\beta_{Z_w,2} = 0.8$, $\beta_{WZs,\text{inter}} = 3$, $\beta_{WZw,\text{inter}} = 1$
Y Scenario 4.1	Nonlinear dependence and interactions, non-null X	$\text{logit}\, P(Y = 1 \mid X = x, W = w, Z = z) = \beta_0 + \beta_W^\top w + \beta_{W,\text{int}} w_w w_s + \beta_{Z_w,1} I(z_w < -0.5) + \beta_{Z_w,2} I(z_w > 2) + \beta_{Z_s} I(z_s < -1) + \beta_{WZs,\text{inter}} w_s z_s + \beta_{WZw,\text{inter}} w_s I(z_w > 2)$, $\beta_X = \log(1.5)$, other regression parameters the same as scenario 4

Base Case – MAR, Census clogOR



Simple MAR, Simple Outcome

Complex MAR, Simple Outcome

Simple MAR, Complex Outcome

Complex MAR, Complex Outcome

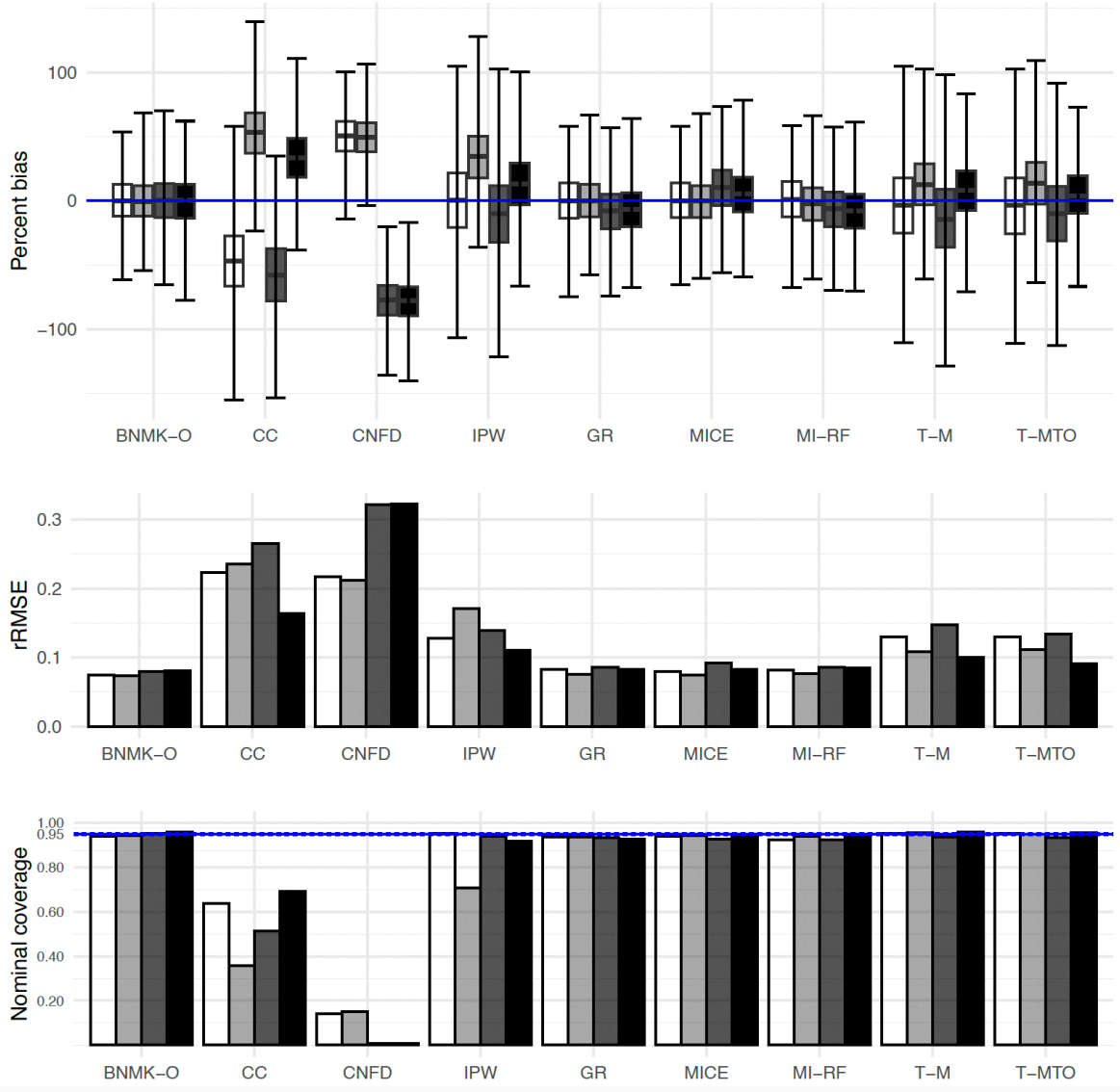
Remarks:

Estimators achieving close to 95% coverage in all 4 scenarios: GR, MI-RF, T-M, T-MTO

Most efficient among these: MI-RF

Mice doing well when outcome model is simple

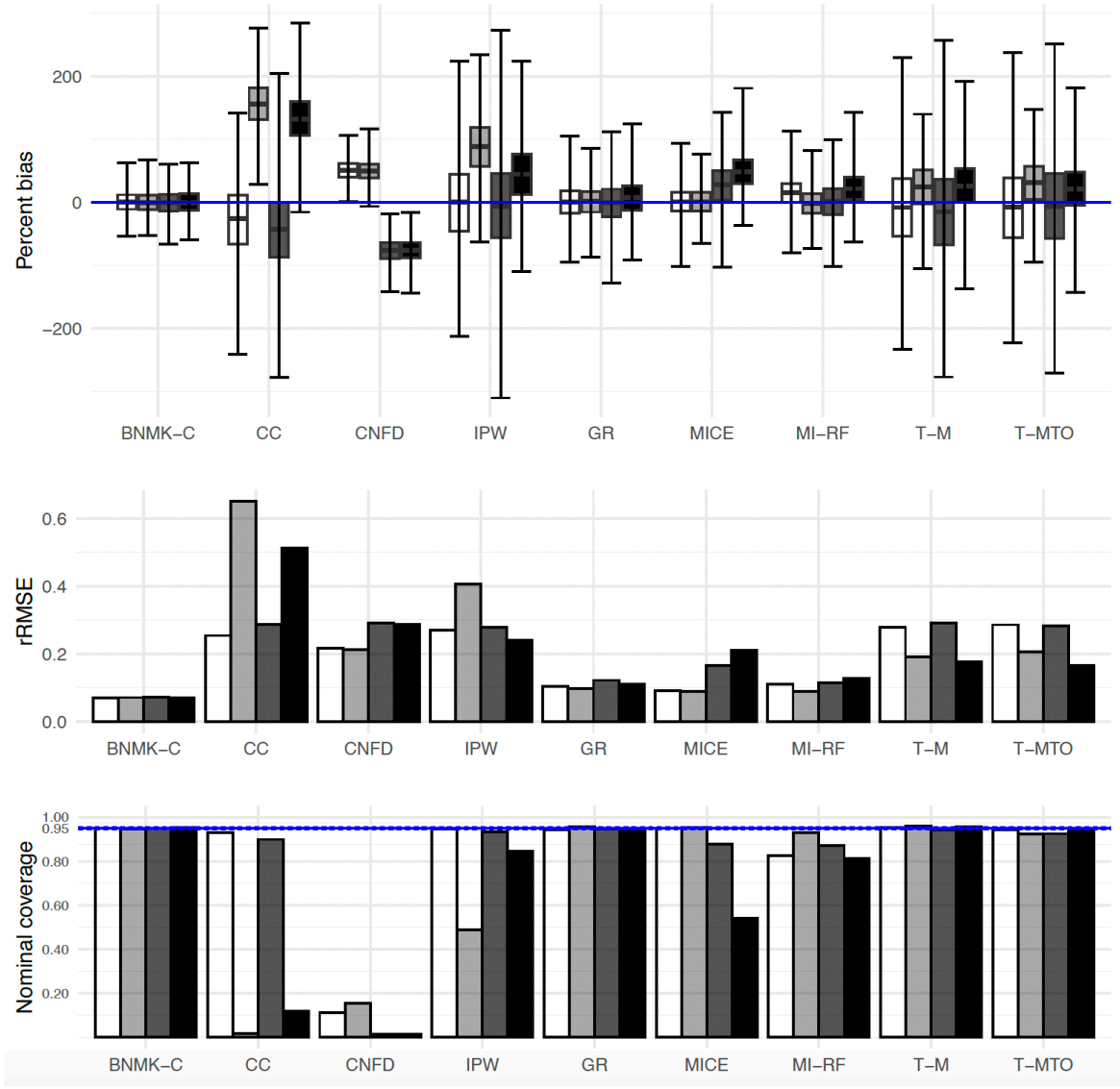
Base Case – MAR, Oracle clogOR



- Simple MAR, Simple Outcome
- Complex MAR, Simple Outcome
- Simple MAR, Complex Outcome
- Complex MAR, Complex Outcome

Remarks:
GR still doing well. MICE now doing better

Base Case with 80% missingness– MAR, Census clogOR



Simple MAR, Simple Outcome

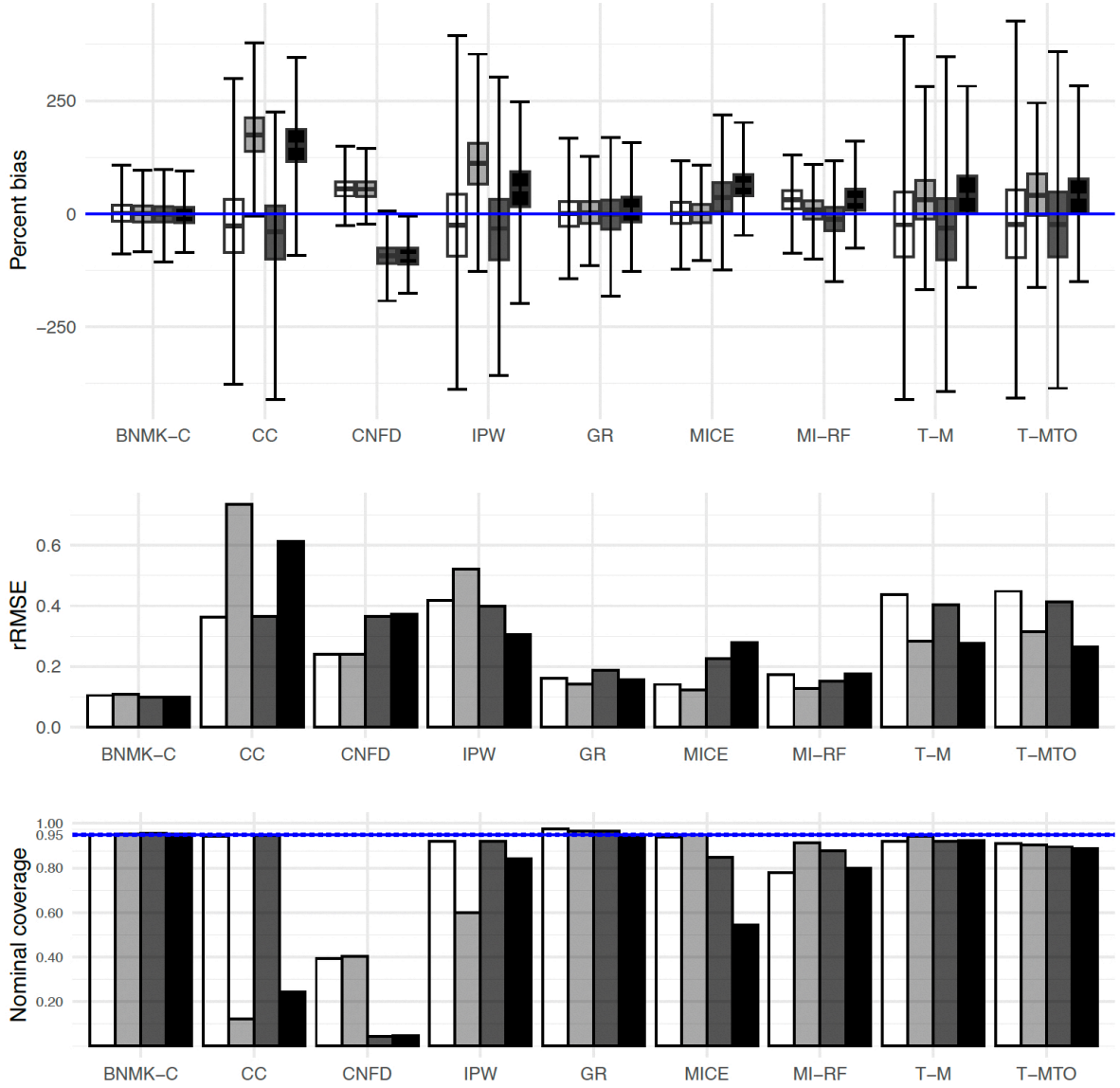
Complex MAR, Simple Outcome

Simple MAR, Complex Outcome

Complex MAR, Complex Outcome

Remarks:
Similar to case with 40% missingness, except MI-RF not doing well any longer

80% missingness + rare outcome: MAR, Census clogOR



Simple MAR, Simple Outcome

Complex MAR, Simple Outcome

Simple MAR, Complex Outcome

Complex MAR, Complex Outcome

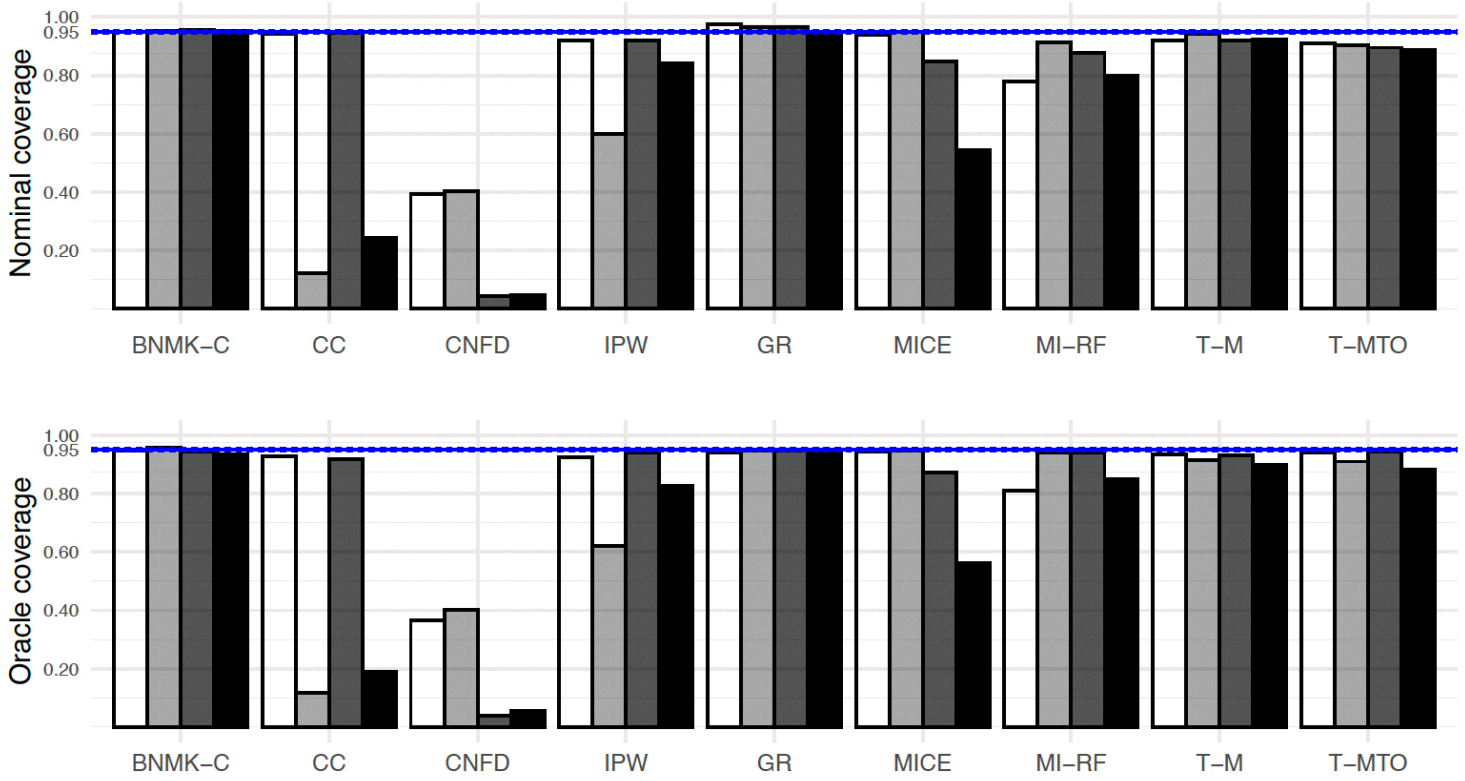
Remarks:

GR still does well, some SE inflation

Both MI and MI-RF fail to achieve 95% coverage with complex outcome. MI-RF doing worse

TMLE having some coverage issues, won't do well under complex MAR

Oracle vs Nominal Coverage



Simple MAR, Simple Outcome

Complex MAR, Simple Outcome

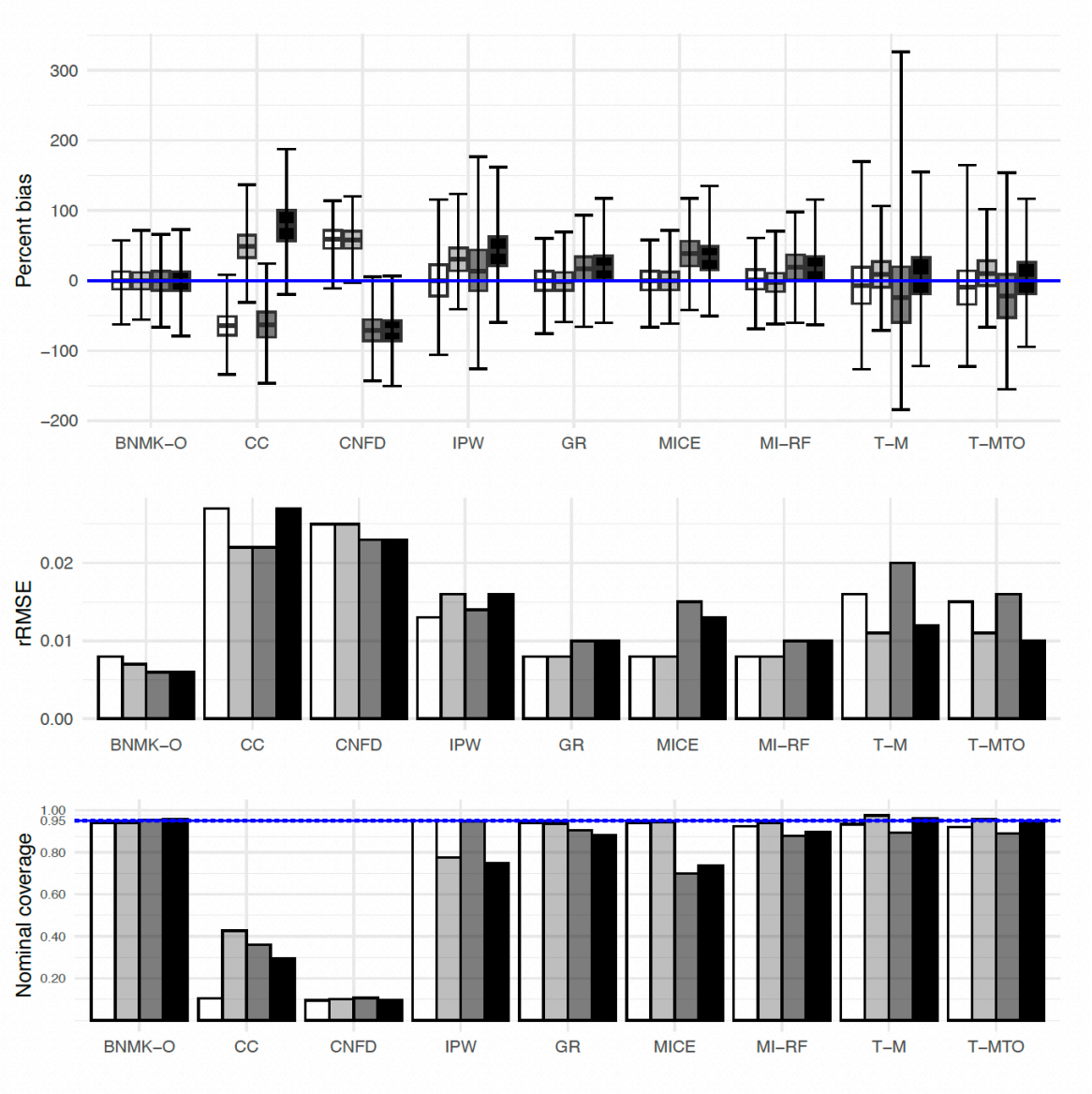
Simple MAR, Complex Outcome

Complex MAR, Complex Outcome

Remarks:
Can see some SE estimation problems for MICE, MI-RF, T-M, T-MTO

Can see double-robustness of TMLE when missingness is correct

Base Case – MAR, Oracle mRD



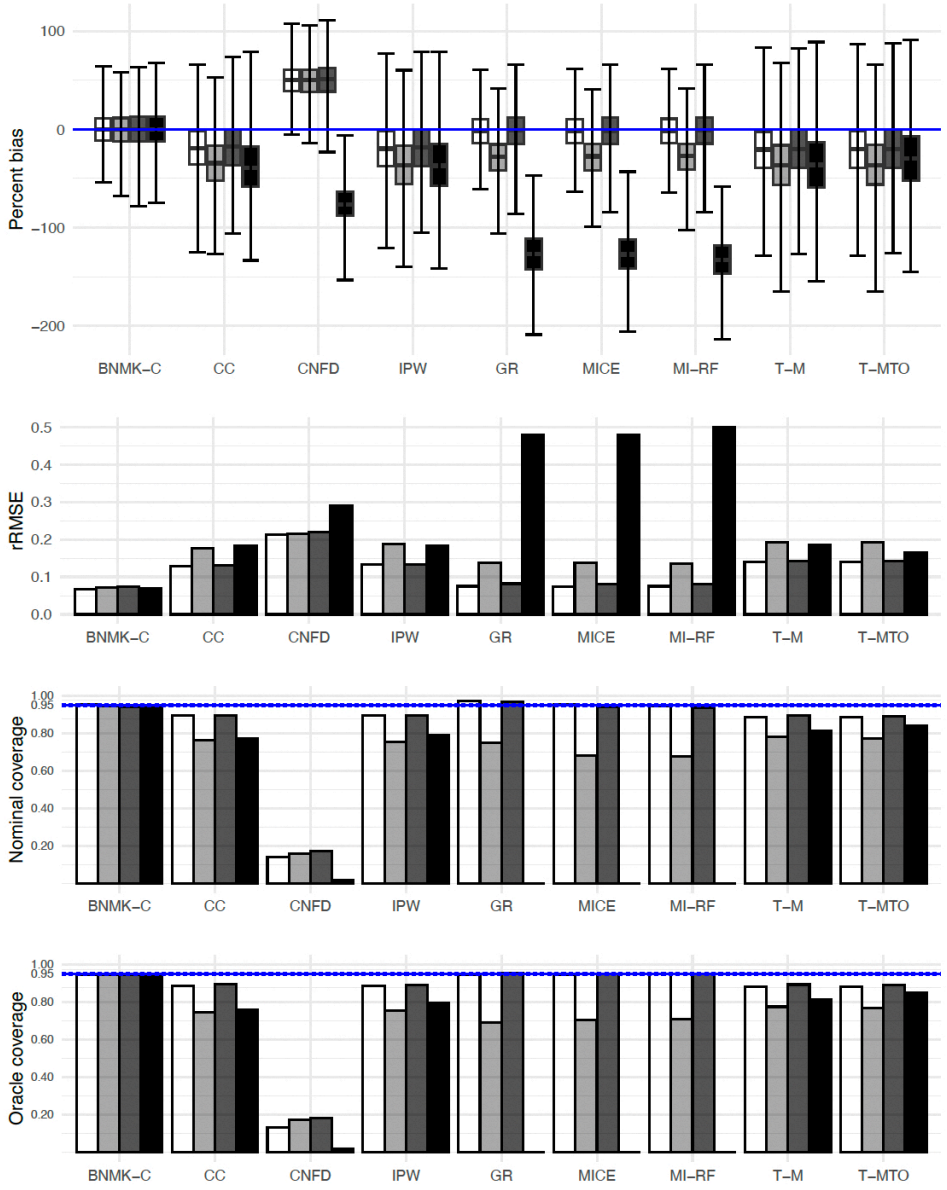
Simple MAR, Simple Outcome

Complex MAR, Simple Outcome

Simple MAR, Complex Outcome

Complex MAR, Complex Outcome

Remark:
GR achieving close to 95% coverage in all 4 scenarios:



MNAR – 12% incidence, 40% missing, Census clogOR

- MNAR unobserved, simple Outcome
- MNAR value, Correct simple Outcome
- MNAR unobserved, simple Outcome +unobserved
- MNAR value, Complex Outcome

Remark:
TMLE had the lowest bias and best overall coverage
GR/MI doing well for MNAR unobserved

For oracle estimators TMLE also showing good efficiency (data not shown)

Takeaways for Raking vs MI vs TMLE from Synthetic Simulations -MAR

Simple MAR, all estimands

- Raking and MI did well and similarly for simple (correctly specified) outcome
- Raking outperformed MI for complex (incorrectly specified) outcome
- TMLE-M, TMLE-MTO did well with respect to bias, but not efficiency

Complex MAR, all estimands

- Raking did well for census estimands and all estimands with simple (correct) outcome model
- Raking still does well for census estimands and cOR, even with complex (incorrect) outcome
- Bias and SE estimation issues for MI with complex (incorrectly specified) outcome model
- TMLE outperformed other estimators for Oracle marginal estimands in terms of coverage and bias. Efficiency good for larger sample size and less missing data.
- A few instances of under coverage for TMLE

Takeaways for Raking vs MI vs TMLE from Synthetic Simulations -MNAR

- Raking still does well for several settings for census estimand (in terms of efficiency and coverage)
- TMLE does well, with notable resilience , for case of the oracle marginal estimands
- MI performance was not consistent in terms of which imputation algorithm did well/sufficiently

A Few Qualifiers

In complex settings studied, conclusions are specific to the flavor of the complex model implemented.

- e.g., raking did well even for Oracle estimand under misspecification. But for more extreme misspecification, might not expect that

"Happy families are all alike; every unhappy family is unhappy in its own way”

-Tolstoy, *Anna Karenina*

Numerical Study Part 2: Plasmode Data

Plasmode Simulation

- Plasmode simulation is a type of numerical simulation that generates the covariate distribution by resampling real data (Franklin et al 2014)
- The specific association of interest is then injected into the data using statistical models
- In our setting we will create 1000 bootstrap samples of the KPWA ADI cohort data
- Still need 3 DGM
  - Missingness model
  - Treatment model
  - Outcome model
- We will compare the same estimation methods as done for the synthetic data
  - XGBoost did not perform that well so was dropped
- In a separate paper, we show for typical causal estimands need to generate treatment from a model and not sample with covariates (Shaw et al. 202X, https://arxiv.org/abs/2504.11740)

Plasmode Data Generating Mechanisms (DGMs)

1. **Treatment data generating model**
   - Antidepressant medication or psychotherapy
2. **Outcome data generating model**
   - Self-harm/Psychiatric hospitalization within 5 years of treatment initiation
3. **Missing data generating model**
   - PHQ-9 measurement (yes/no)

Data generating models estimated from KPWA ADI Cohort

- Treatment and outcome model fit to 50,337 with complete data
- Missing data model considers full population of 112,770 individuals with 50,337 (45%) having complete data

For each type of generating model use ADI cohort to estimate:

1. Parametric model (e.g. logistic regression) with interactions
2. Tree-based model (allowing for complex interactions)

DGM more complex and had more variables than analysis models

Confounders

1. Sex (mostly represents sex assigned at birth)
2. Age in years at time of index visit (18-24, 25-34, 35-44, 45-54, 55-64, 65+ years)
3. Charlson comorbidity index (0, 1, 2, 3+)
4. Anxiety diagnosis in the past year
5. Self-harm in the prior 6 months
6. Psychiatric hospitalization in the prior 5 years
7. Alcohol use disorder in the past year
8. 9 item Patient Health Questionnaire (PHQ-9)
    - Sum of first 8 items to summarize depressive symptoms (PHQ-8)
    - 9th item measure of suicidal ideation

Covariates Included in the DGMs for Outcome, Treatment, and Missingness

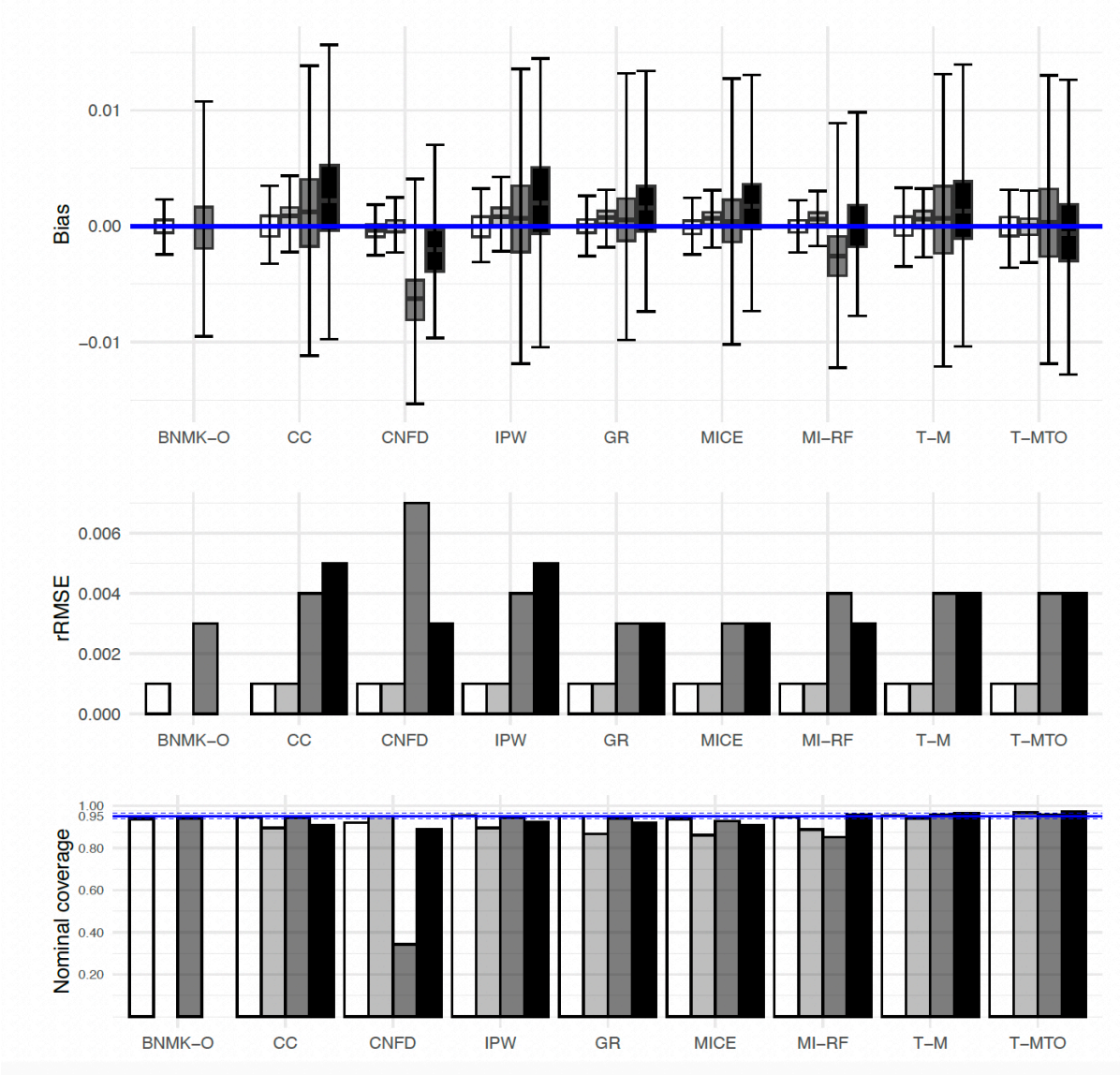
- Treatment exposure + 8 confounders
- Interactions:
  - Categorical age and sex
  - Prior self-harm and sex
  - Prior self-harm and categorical age
  - PHQ 9th item and sex
  - PHQ 9th item and prior self-harm

KPWA ADI Cohort (N=50,337)

Characteristic % (n)	Antidepressant medication	Psychotherapy	Overall
Female	67.1 (18369)	62.0 (14245)	64.8 (32618)
Age in years*	44.4 (19.0)	38.6 (18.4)	41.8 (18.9)
13 to 17	5.0 (1369)	13.5 (3179)	9.0 (4548)
18 to 29	21.7 (5936)	24.6 (5659)	23.0 (11595)
30 to 44	25.2 (6903)	25.1 (5752)	25.1 (12655)
45 to 64	32.5 (8906)	26.0 (5963)	29.5 (14869)
65 or older	15.6 (4263)	10.5 (2407)	13.3 (6670)
Charlson†			
0	75.4 (20654)	79.1 (18168)	77.1 (38822)
1	14.1 (3852)	13.1 (3018)	13.6 (6810)
2	5.0 (1364)	4.0 (919)	4.5(2283)
3 or more	5.5 (1507)	3.7 (855)	4.7 (2364)
Anxiety disorder†	13.2 (3611)	18.7 (4284)	15.7 (7895)
Alcohol use disorder†	2.0 (544)	2.4(558)	2.2 (1102)
Prior Self-harm‡	0.3 (90)	0.7 (172)	0.5 (262)
Prior Psychiatric hospitalization‡	6.8 (1858)	4.9 (1116)	5.9 (2974)
PHQ8 total score*	14.7 (5.0)	11.7 (5.8)	13.3 (5.9)
0-5			9.7 (4868)
5-10 (Mild symptoms)	4.4 (1200)	16.0 (3668)	21.7 (10933)
11-15 (Moderate symptoms)	16.6 (4540)	27.8 (6393)	31.1 (15634)
16-20 (Moderate/Severe symptoms)	33.4 (8934)	19.9(4575)	26.8 (13509)
21-24 (Severe symptoms)	13.0 (3569)	7.9 (1824)	10.7 (5393)
PHQ item 9*			
0 (none of the days)	66.5 (18198)	67.2(15422)	5.1 (2566)
1 (several days)	20.1(5507)	20.4(4687)	7.9 (3957)
2 (more than half the days)	8.4(2296)	7.2(1661)	20.3 (10194)
3 (nearly every day)	5.0(1376)	5.2(1190)	66.8 (33620)

* At time of initiation
† In prior 6 months
‡ In prior 12 months

Plasmode results, Oracle mRD



GLM 1-year self-harm

Tree 1-year self-harm

GLM 5-year self-harm/hospitalization

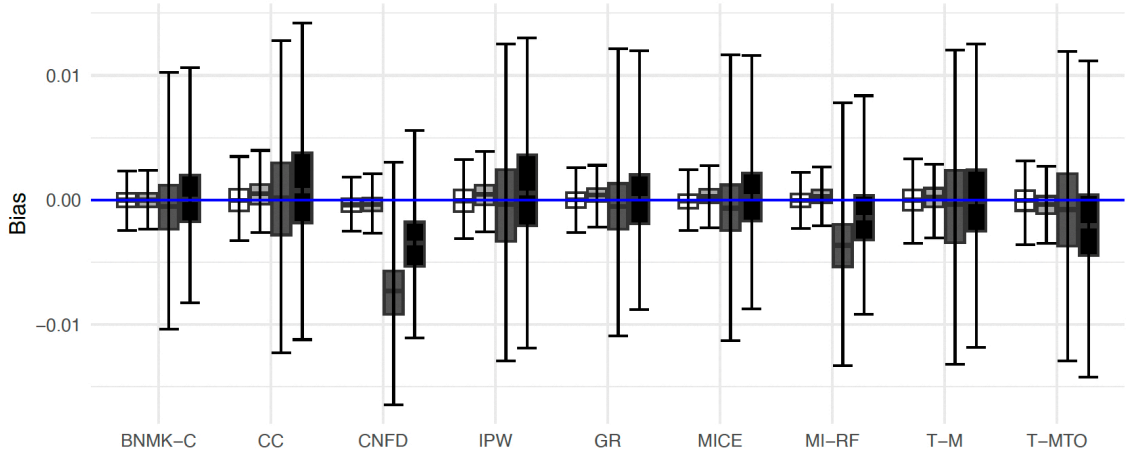
Tree 5-year self-harm/hospitalization

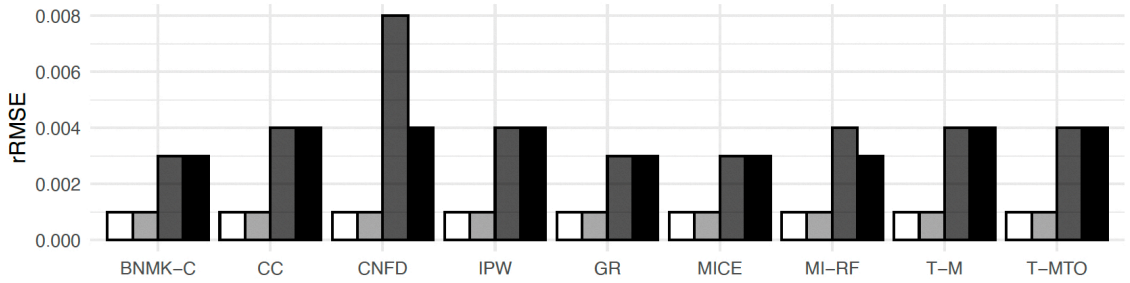
Remarks:

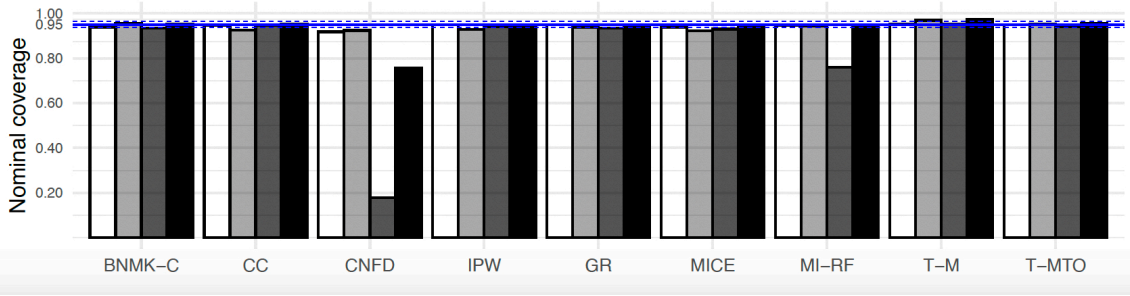
For 5-year outcome: MICE and GR behaved similarly and maintained best performance across all estimands

For 1-year outcome: TMLE-MTO maintained best overall coverage. Most efficient estimator that maintained nominal coverage for Oracle mRD and tree DGM

Plasmode results, Census mRD







Conclusions

- Survey calibration/Generalized raking are rarely performed in biomedical settings, but were observed to have the best overall performance in the majority of settings
  - These estimators are practical to implement in standard software
- TMLE did well with respect to bias, but only did well with respect to efficiency in certain settings
  - Was efficient for oracle estimand mRD when there were complex DGMs and larger sample size/event rates
- One strategy may be to perform TMLE and raking, gaining confidence when estimates agree
  - When estimates don't agree, it may be missingness model is misspecified, or oracle and census estimands are different
- MI often did well with respect to efficiency but no one algorithm did well across all the settings
- Confounded and complete case approaches were generally poor performing methods
- There were a few edge cases (MNAR) where complete case and IPW did well (discussed by Little et al 2022, Lee et al 2023, and others)
- Our numerical experiments highlight the importance of first choosing a target estimand and then determining an estimation procedure.
- Plasmode simulation studies are useful in guiding methods selection, but need to consider DGMs that don't uniformly favor one method
- Scenarios studied are comprehensive, but not exhaustive

Resources

https://github.com/PamelaShaw/Missing-Confounders-Methods/

- Provide R code that implemented these methods
- Provide vignettes that explain in detail the general principles and steps to applying these methods
- Contact pamela.a.shaw@kp.org for further information

Manuscript on Arxiv

https://arxiv.org/abs/2412.15012

Thank you!

Useful recent papers

Little RJ, Carpenter JR, Lee KJ. A comparison of three popular methods for handling missing data: complete-case analysis, inverse probability weighting, and multiple imputation. Sociological Methods & Research. 2022 Aug 5:00491241221113873.

- Discusses what to expect under MAR and MNAR, highlighting when each method may be expected to do well
- Some interesting special (edge?) cases where IPW or CC can beat MI

Weberpals J et al A Principled Approach to Characterize and Analyze Partially Observed Confounder Data from Electronic Health Records. Clinical Epidemiology. 2024 Dec 31:329-43.

- CI3 paper that looked at methods performance under MAR and MNAR
- In separate paper, Weberpals et al saw high dimensional auxiliary data approach offered marginal improvements

Weberpals J, et al. smdi: an R package to perform structural missing data investigations on partially observed confounders in real-world evidence studies. JAMIA open. 2024 Apr 1;7(1):ooae008.

- Software paper describing how smdi package can be used to investigate patterns in missing data

Lee, Katherine J., et al. "Assumptions and analysis planning in studies with missing data in multiple variables: moving beyond the MCAR/MAR/MNAR classification." International Journal of Epidemiology 52.4 (2023): 1268-1275.

- Outlines that MAR and MNAR can be unhelpful distinctions and need to think about causal diagrams

References: Missing Data

- Getz K, Hubbard RA, Linn KA. Performance of Multiple Imputation Using Modern Machine Learning Methods in Electronic Health Records Data. Epidemiology. 2023 Mar:10-97.
- Little RJ, Carpenter JR, Lee KJ. A comparison of three popular methods for handling missing data: complete-case analysis, inverse probability weighting, and multiple imputation. Sociological Methods & Research. 2022 Aug 5:00491241221113873.
- Bartlett J, Keogh R., and Bonneville EF. Smcfcs: Multiple Imputation of Covariates by Substantive Model Compatible Fully Conditional Specification. R package version 1.7.1; 2022. https://CRAN.R-project.org/package=smcfcs.
- Lee KJ, Tilling KM, Cornish RP, Little RJ, Bell ML, Goetghebeur E, Hogan JW, Carpenter JR. Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. Journal of clinical epidemiology. 2021 Jun 1;134:79-88.
- Aleryani A, Wang W, De La Iglesia B. Multiple imputation ensembles (MIE) for dealing with missing data. SN Computer Science. 2020 May;1:1-20.
- Van Buuren, S. (2018). Flexible Imputation of Missing Data. Second Edition. Chapman & Hall/CRC. Boca Raton, FL.
- Bartlett JW, Seaman SR, White IR, Carpenter JR, Alzheimer's Disease Neuroimaging Initiative*. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. Statistical methods in medical research. 2015 Aug;24(4):462-87.
- Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Statistical methods in medical research. 2013 Jun;22(3):278-95.
- Carpenter JR, Kenward MG. Missing data in randomised controlled trials: a practical guide. 2007

References: 2-Phase Design

- Lumley T, Shaw PA, Dai JY. Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review*. 2011; 79 (2): 200–220.
- Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*. 1994 Sep 1;89(427):846-66.
- Han K, Shaw PA, Lumley T. Combining multiple imputation with raking of weights: An efficient and robust approach in the setting of nearly true models. *Statistics in Medicine*. 2021;;40(30):6777-91.
- Särndal CE, Swensson B, Wretman J. *Model assisted survey sampling*. Springer Science & Business Media; 2003.
- Deville JC, Särndal CE. Calibration estimators in survey sampling. *Journal of the American Statistical Association*. 1992 Jun 1;87(418):376-82.
- Deville JC, Särndal CE, Sautory O. Generalized raking procedures in survey sampling. Journal of the American statistical Association. 1993 Sep 1;88(423):1013-20.
- Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Improved Horvitz–Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. *Statistics in Biosciences*. 2009 May;1(1):32-49.
- Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Using the whole cohort in the analysis of case-cohort data. *American Journal of Epidemiology*. 2009 Jun 1;169(11):1398-405.
- Rose S, van der Laan MJ. A targeted maximum likelihood estimator for two-stage designs. *The International Journal of Biostatistics*. 2011 Mar 11;7(1).
- Lumley T. Robustness of semiparametric efficiency in nearly-true models for two-phase samples; 2017. ArXiv e-prints arXiv: 1707.05924.

References: Propensity / Probability of Treatment

- Stephens DA, Nobre WS, Moodie EE, Schmidt AM. Causal inference under mis-specification: adjustment based on the propensity score. Bayesian Analysis. 2022 Jan;1(1):1-24.
- Hernan MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.
- Goetghebeur E, le Cessie S, De Stavola B, Moodie EE, Waernbaum I, ” the topic group Causal Inference (TG7) of the STRATOS initiative. Formulating causal questions and principled statistical answers. Statistics in medicine. 2020 Dec 30;39(30):4922-48.
- Shortreed SM, Moodie EE. Automated analyses: Because we can, does it mean we should?. Statistical science: a review journal of the Institute of Mathematical Statistics. 2020 Aug;35(3):499.
- Alam S, Moodie EE, Stephens DA. Should a propensity score model be super? The utility of ensemble procedures for causal adjustment. Statistics in Medicine. 2019 Apr 30;38(9):1690-702.Karim ME, Pang M, Platt RW. Can we train machine learning methods to outperform the high-dimensional propensity score algorithm?. Epidemiology. 2018 Mar 1;29(2):191-8.
- Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research. 2011 May 31;46(3):399-424.
- D'Agostino Jr RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Statistics in medicine. 1998 Oct 15;17(19):2265-81.

References: TMLE

- Benkeser D, Hejazi N (2023). survtmle: Targeted minimum loss-based estimation for survival analysis. doi: 10.5281/zenodo.835868, R package version 1.1.3.9000; https://github.com/benkeser/survtmle.
- Lendle SD, Schwab J, Petersen ML, van der Laan MJ (2017). "ltmle: An R Package Implementing Targeted Minimum Loss-Based Estimation for Longitudinal Data." *Journal of Statistical Software*, 81(1), 1–21. doi:10.18637/jss.v081.i01.
- Gruber S, van der Laan MJ. (2012). tmle: An R Package for Targeted Maximum Likelihood Estimation. *Journal of Statistical Software*, 51(13), 1-35. http://www.jstatsoft.org/v51/i13.
- Rose S, van der Laan MJ. A targeted maximum likelihood estimator for two-stage designs. *The International Journal of Biostatistics*. 2011 Mar 11;7(1).
- Gruber S and van der Laan, MJ. Targeted Maximum Likelihood Estimation: A Gentle Introduction. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 252. Berkeley Electronic Press; 2009. https://biostats.bepress.com/ucbbiostat/paper252
- van der Laan MJ and Rubin D. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.
- van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2007). Super learner. Statistical Applications in Genetics and Molecular Biology 6 (1), 1–20.

Extra slides

Estimating Marginal Estimands

TMLE-based methods provide marginal estimates by default

Other approaches:

1. Fit logistic regression model, obtain predicted probabilities $\widehat{p_{i,x}} = \hat{P}(Y_i = 1 \,|X_i = x, W_i, Z_i)$ using observed (W_i, Z_i) and $X_i = x$
2. Obtain weights w_i to generalize to the full population
   - For complete-case methods and MI-based methods, $w_i = 1$ for $i = 1, \ldots, n$
   - For IPW, $w_i = \frac{1}{\widehat{h_i}}$, where $\widehat{h_i} = \hat{P}(\Delta_i = 1 \,|X_i = x, Y_i = y, Z_i = z)$ (Δ is missing indicator)
   - For raking, $w_i =$ calibrated IPW weights (from raking procedure)
3. Obtain estimators $\widehat{\mu_x} = n^{-1} \sum_{i=1}^{n} w_i \hat{p}_{i,x}$
4. Estimate marginal parameters:

$$\widehat{mRR} = \frac{\widehat{\mu_1}}{\widehat{\mu_0}}$$

$$\widehat{mRD} = \widehat{\mu_1} - \widehat{\mu_0}$$

$$\widehat{mOR} = \frac{\widehat{\mu_1}/(1-\widehat{\mu_1})}{\widehat{\mu_0}/(1-\widehat{\mu_0})}$$

Model assumptions required for consistency

Table 1: Model specifications for missing-data estimation procedures and requirements for consistency for census and oracle estimands, under missing-at-random and assuming that there are no unmeasured confounders. The working outcome model Q is assumed to be the same as the census model. The missing-data model is π, the imputation model is f, and the treatment/exposure propensity score model is 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
IPCW-TMLE	Q, g and π	(Q or g) and π	π or [(Q or g) and CD-MCAR]

IPW: inverse probability weighted outcome regression where weights account for missing data; CD-MCAR: covariate-dependent missing completely at random, where missing data can depend only on always-observed covariates, not outcomes (Seaman et al., 2013); MI: multiple imputation; GR: generalized raking; IPCW: inverse probability of coarsening weighted; TMLE: targeted minimum loss-based estimation