



Principled Approaches To Handle Partially Observed Confounder Data From Electronic Health Records: A Plasmode Simulation Study






Janick Weberpals, Sudha R. Raman, Pamela A. Shaw, Hana Lee, Bradley G. Hammill, Sengwee Toh, John G. Connolly, Kimberly J. Dandreo, Fang Tian, Wei Liu, Jie Li, José J. Hernández-Muñoz, Robert J. Glynn, Rishi J. Desai

Disclosures

- Janick Weberpals reports prior employment by Hoffmann-La Roche and previously held shares in Hoffmann-La Roche
- This project was supported by Task Order 75F40119F19002 under Master Agreement 75F40119D10037 from the U.S. Food and Drug Administration (FDA)



Background

- Administrative claims databases are increasingly linked to electronic health records (EHR) to **improve confounding adjustment** for variables which cannot be measured in claims.
- Examples of variables which cannot be measured in claims:
 - Labs (HbA1c, LDL, etc.) 
 - Vitals (Blood pressure, BMI, etc.) 
 - Disease-specific data (cancer stage, biomarkers, etc.) 
 - Physician assessments (ECOG, etc.) 
 - Lifestyle factors (smoking, alcohol, etc.) 
- These variables are often just partially observed in EHR for various reasons
 - Physician did not perform/order a certain test
 - Certain measurements are just collected for particularly sick patients
 - Information is ‘hiding’ in unstructured records (e.g., clinical notes)



Knowledge Gaps and Objectives

Missing data in confounding factors are frequent

- **Mechanisms:** Missing completely at random (**MCAR**), at random (**MAR**) and not at random (**MNAR**)
- **Patterns:** Monotone, Non-monotone

Unresolved challenges for causal inference

- In **an empirical study**, it is usually unclear **which of the missing data mechanisms and patterns (most likely) hold**.
- How do any of these mechanisms and patterns relate to **bias in a given real-world data (RWD) study**, given the strength of correlations between exposure, covariates and outcomes in high-dimensional covariate spaces (e.g., database linkages)?

Objectives:

- **Develop a framework and tools to assess the mechanisms and patterns of missing data processes in EHR studies**
- **Connect this with the most appropriate analytical approach, followed by sensitivity analyses**

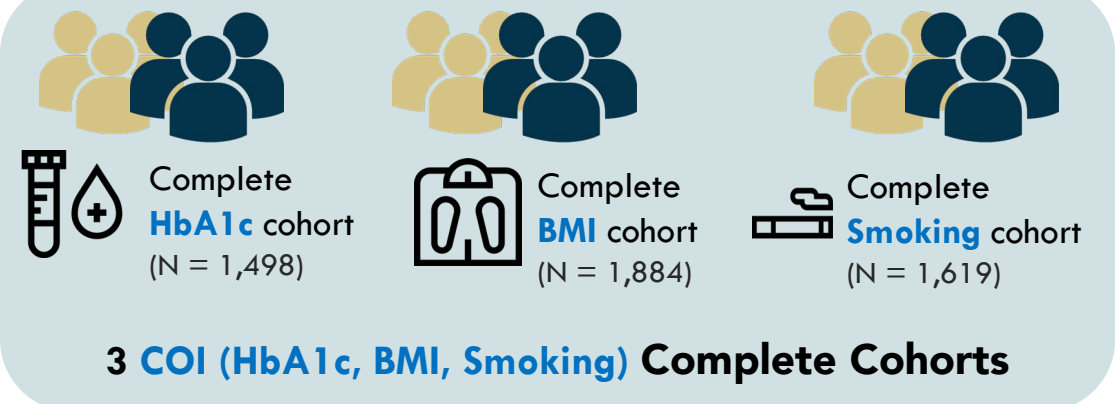
- Rubin DB. Inference and Missing Data. *Biometrika*. 1976;63(3):581-592. doi:10.2307/2335739
- Mitra, R., McGough, S.F., Chakraborti, T. et al. Learning from data with structured missingness. *Nat Mach Intell* 5, 13–23 (2023)
- Mohan K, Pearl J, Tian J. Graphical models for inference with Missing data. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'13. Curran Associates Inc.; 2013:1277-1285.

SGLT-2 inhibitor 

DPP-4 inhibitor 

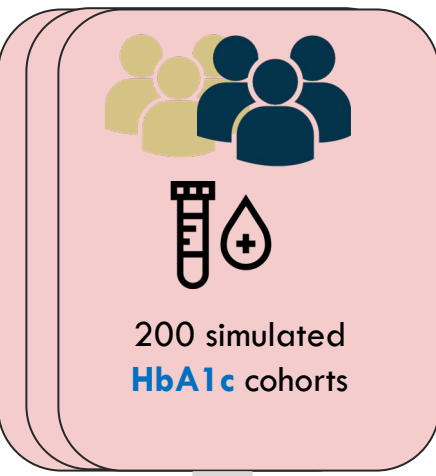
Empirical Cohort (N=9,339)

Restriction to sub-cohorts with complete information on **confounder of interest (COI)**

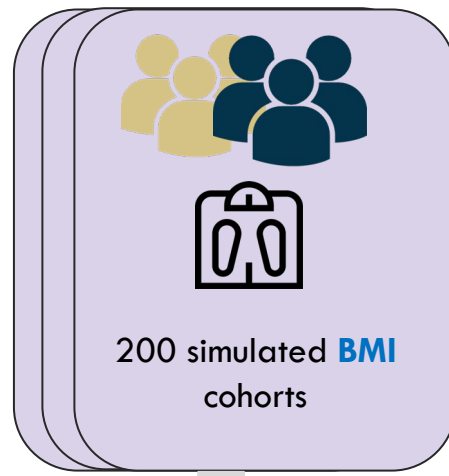


3 COI (HbA1c, BMI, Smoking) Complete Cohorts

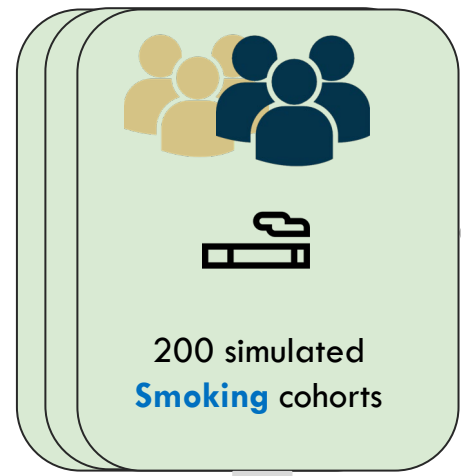
- Complete HbA1c cohort** (N = 1,498)
- Complete BMI cohort** (N = 1,884)
- Complete Smoking cohort** (N = 1,619)



200 simulated **HbA1c** cohorts



200 simulated **BMI** cohorts



200 simulated **Smoking** cohorts

Parametric Bootstrap (Plasmode) for each COI cohort

- Select of 25 prognostic covariates (C1 covariates)
- Model empirical associations of outcome and censoring as function of Exposure + **COI** + C1
- Use modeled parameter estimates to estimate new survival functions and simulate **true null association for the exposure:**

log hazard ratio [HR]_{SGLT-2i vs. DPP-4i} = 0

- Simulate outcome and resample 100 cohorts of each 1,000 patients
- Simulate additional 100 cohorts including an interaction term for Exposure and COI to simulate heterogeneous treatment effects

For all simulated cohorts:

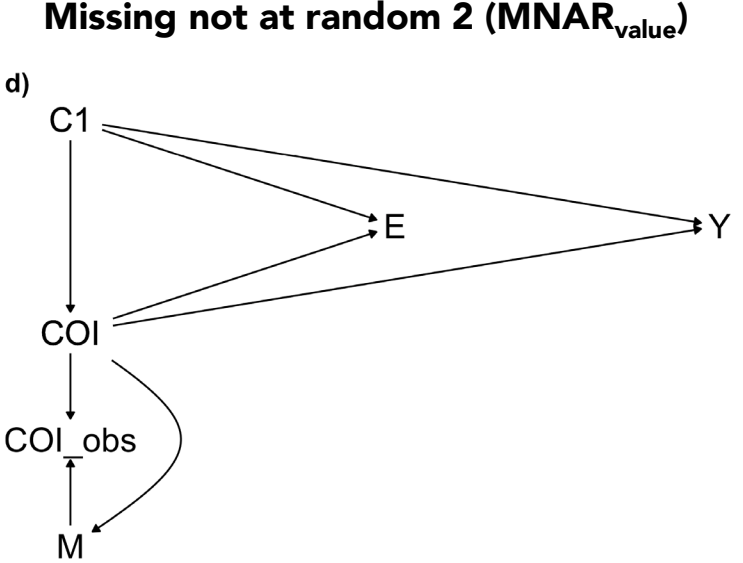
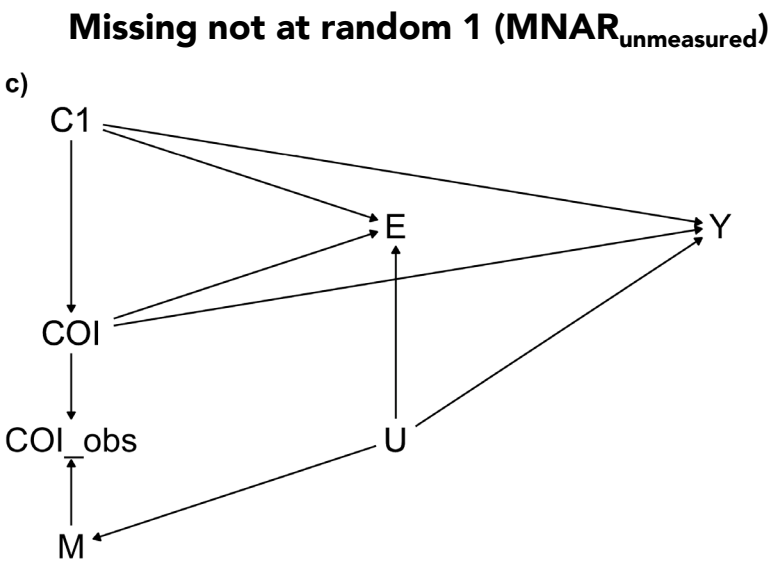
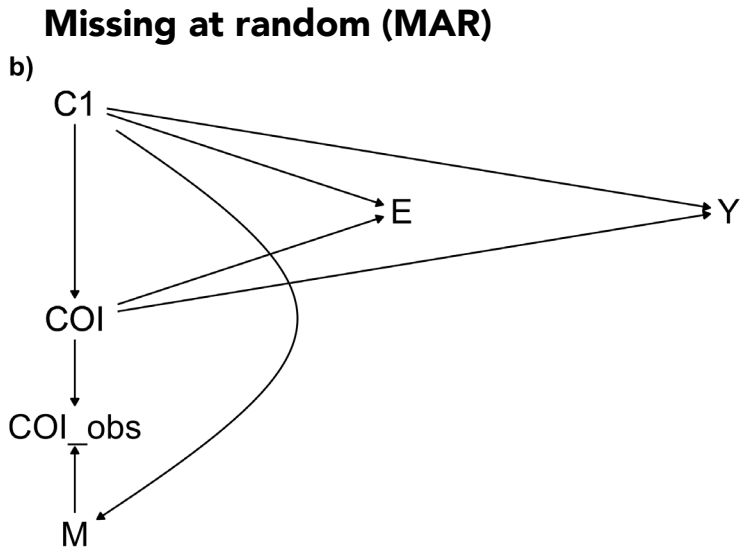
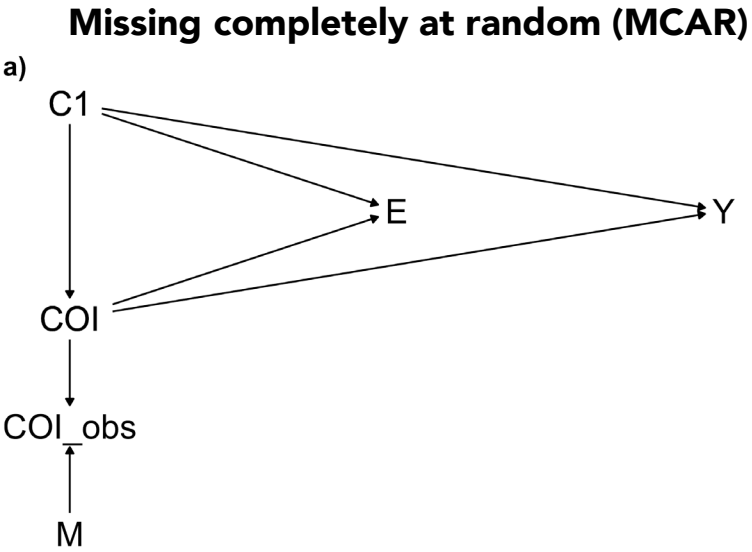
- Introduce varying proportions (10-50%) of missingness to COI** following MCAR, MAR, MNAR(unmeasured), MNAR(value) mechanisms
- Compute **Diagnostics** to investigate how well the **missingness mechanisms of the COI** can be characterized
- Examine how well different **Analytic Approaches** (i.e., complete case analysis, inverse probability weighting, missing indicator method, single and multiple Imputation) are able to recover the true log HR

Simulated Causal Missingness Mechanisms

Causal diagrams/M-graphs

E	Exposure (SGLT2i vs. DPP4i)
Y	Outcome
COI	Confounder of interest (HbA1c, BMI, Smoking)
COI_obs	Observed portion of COI
M	Missingness of COI
C1	Fully observed confounders
U	Unmeasured confounder

- Choi J, Dekkers OM, le Cessie S. A comparison of different methods to handle missing data in the context of propensity score analysis. *Eur J Epidemiol.* 2019 Jan;34(1):23-36.
- Mohan K, Pearl J, Tian J. Graphical models for inference with Missing data. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1.* NIPS'13. Curran Associates Inc.; 2013:1277-1285.



Empirical Diagnostics to Characterize Missingness Mechanisms

Group 1 Diagnostics

	Median Absolute standardized mean difference (ASMD)	P-value Hotelling/Little
Purpose	Comparison of distributions of observed covariates between patients with vs w/o observed value of the partially observed confounder	
Example value	ASMD = 0.1	p-value <0.001
Interpretation	<p><0.1*: no imbalances in observed patient characteristics; missingness may be likely completely at random or not at random (~MCAR, ~MNAR).</p> <p>>0.1*: imbalances in observed patient characteristics; missingness may be likely at random (~MAR).</p> <p>* Equivalent to propensity score-based balance measures (Austin PC, Multivariate Behavioral Research, 46:3, 399-424 [2011])</p>	<p>High test statistics and low p-values indicate differences in baseline covariate distributions and null hypothesis would be rejected (~MAR).</p> <p>Hotelling H. Ann Math Stat. 2(3):360-378. (1931) & Little RJA. J Am Stat Assoc. 83(404):1198-1202. doi:10.2307/2290157 (1988)</p>

Empirical Diagnostics to Characterize Missingness Mechanisms

	Group 1 Diagnostics		Group 2 Diagnostics
	Median Absolute standardized mean difference (ASMD)	P-value Hotelling/Little	AUC (area under the receiver operating characteristic curve)
Purpose	Comparison of distributions of observed covariates between patients with vs w/o observed value of the partially observed confounder		Assessing the ability to predict confounder missingness based on observed covariates
Example value	ASMD = 0.1	p-value <0.001	AUC = 0.5
Interpretation	<p><0.1*: no imbalances in observed patient characteristics; missingness may be likely completely at random or not at random (~MCAR, ~MNAR).</p> <p>>0.1*: imbalances in observed patient characteristics; missingness may be likely at random (~MAR).</p> <p>* Equivalent to propensity score-based balance measures (Austin PC, Multivariate Behavioral Research, 46:3, 399-424 [2011])</p>	<p>High test statistics and low p-values indicate differences in baseline covariate distributions and null hypothesis would be rejected (~MAR).</p> <p>Hotelling H. Ann Math Stat. 2(3):360-378. (1931) & Little RJA. J Am Stat Assoc. 83(404):1198-1202. doi:10.2307/2290157 (1988)</p>	<p>AUC values ~ 0.5 indicate completely random or not at random prediction (~MCAR, ~MNAR).</p> <p>Values meaningfully above 0.5 indicate stronger relationships between covariates and missingness (~MAR).</p>

Empirical Diagnostics to Characterize Missingness Mechanisms

	Group 1 Diagnostics	Group 2 Diagnostics	Group 3 Diagnostics	
	Median Absolute standardized mean difference (ASMD)	P-value Hotelling/Little	AUC (area under the receiver operating characteristic curve)	Log HR (missingness indicator)
Purpose	Comparison of distributions of observed covariates between patients with vs w/o observed value of the partially observed confounder		Assessing the ability to predict confounder missingness based on observed covariates	Check whether confounder missingness is associated with the outcome (differential missingness)
Example value	ASMD = 0.1	p-value <0.001	AUC = 0.5	log HR = 0.1 (0.05 to 0.2)
Interpretation	<p><0.1*: no imbalances in observed patient characteristics; missingness may be likely completely at random or not at random (~MCAR, ~MNAR).</p> <p>>0.1*: imbalances in observed patient characteristics; missingness may be likely at random (~MAR).</p> <p>* Equivalent to propensity score-based balance measures (Austin PC, Multivariate Behavioral Research, 46:3, 399-424 [2011])</p>	<p>High test statistics and low p-values indicate differences in baseline covariate distributions and null hypothesis would be rejected (~MAR).</p> <p>Hotelling H. Ann Math Stat. 2(3):360-378. (1931) & Little RJA. J Am Stat Assoc. 83(404):1198-1202. doi:10.2307/2290157 (1988)</p>	<p>AUC values ~ 0.5 indicate completely random or not at random prediction (~MCAR, ~MNAR).</p> <p>Values meaningfully above 0.5 indicate stronger relationships between covariates and missingness (~MAR).</p>	<p>No association in either univariate or adjusted model and no meaningful difference in the log HR after full adjustment (~MCAR).</p> <p>Association in univariate but not fully adjusted model (~MAR).</p> <p>Meaningful difference in the log HR also after full adjustment (~MNAR).</p>

Diagnosics Results (Averaged Over All Simulation Parameters)

- Plasmode simulation revealed [characteristic patterns in the diagnostic parameters that matched each missing data structure](#)
- Patterns were consistent across simulation parameters (COI/variable type, addition of auxiliary variables, % missing)
 - Higher AUC values under simulated MAR with higher proportions of missing

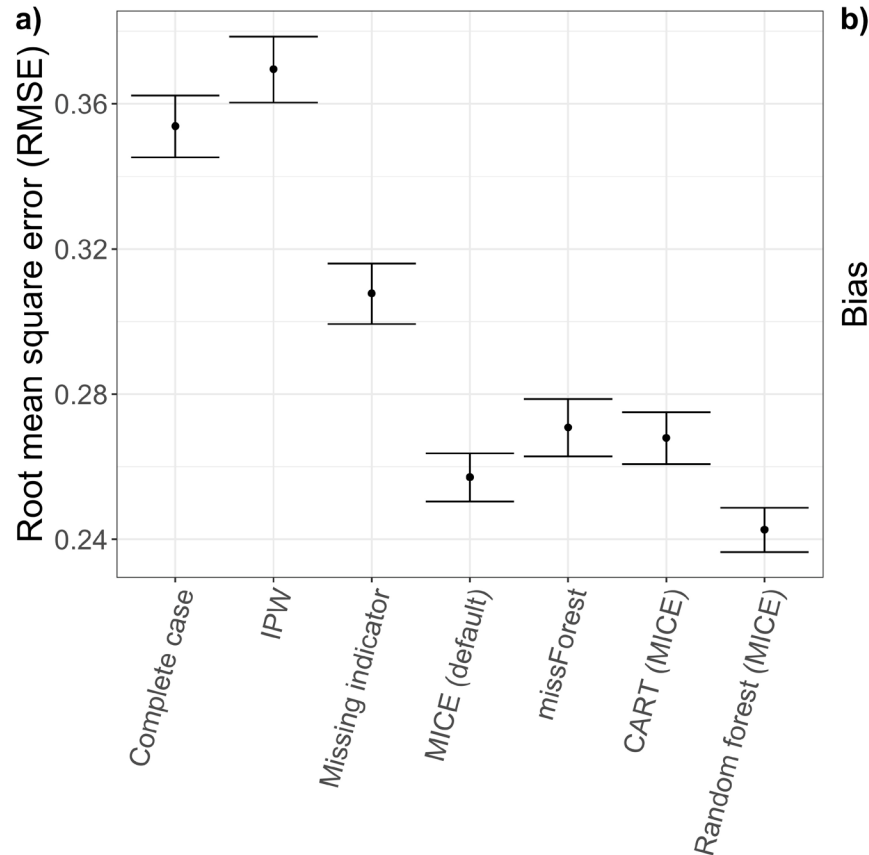
Expected parameter constellations	Group 1 Diagnostics		Group 2 Diagnostics	Group 3 Diagnostics	
	ASMD (Absolute standardized mean difference)	P-value Hoteling/Little	AUC (are under the receiver operating characteristic curve)	Log HR (univariate)	Log HR (adjusted)
MCAR	0.05 (0.05, 0.05)	0.5	0.50 (0.50, 0.50)	0.00 (-0.01, 0.00)	0.00 (-0.01, 0.00)
MAR	0.20 (0.20, 0.20)	<.001	0.58 (0.58, 0.59)	0.53 (0.53, 0.54)	0.00 (-0.01, 0.00)
MNAR_{unmeasured}	0.09 (0.09, 0.09)	0.02	0.54 (0.54, 0.54)	0.43 (0.43, 0.44)	0.31 (0.31, 0.32)
MNAR_{value}	0.06 (0.06, 0.06)	0.10	0.53 (0.53, 0.53)	0.05 (0.04, 0.05)	0.10 (0.09, 0.10)

Confidence intervals were derived using the Monte Carlo simulation standard error as described by White and Crowther, *Statistics in Medicine*, 38: 2074-2102 (2019) and Gasparini, *Journal of Open Source Software*, 3, 739 (2018)

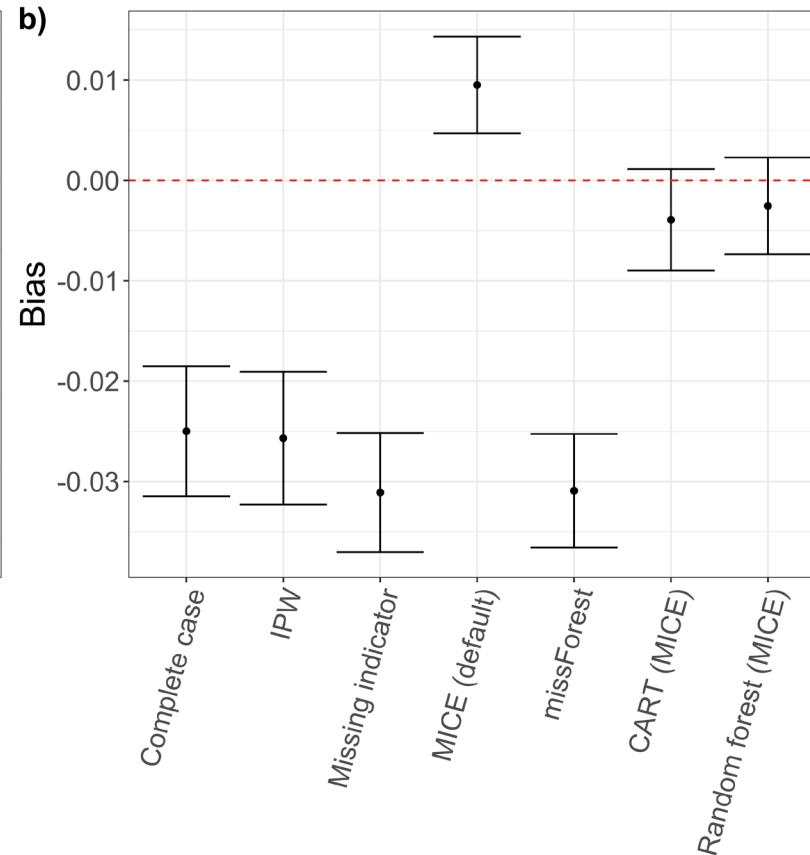
Analytics Results (Averaged Over All Simulation Parameters)

- Overall, multiple imputation (MI) approaches resulted in the lowest RMSE, bias and were most efficient models
- Non-parametric MI (missForest, CART, Random Forest) showed marginal benefits in situations with heterogeneous treatments effects
- Naïve IPW models showed similar bias compared to CC analysis but were consistently less efficient (positivity violation?)

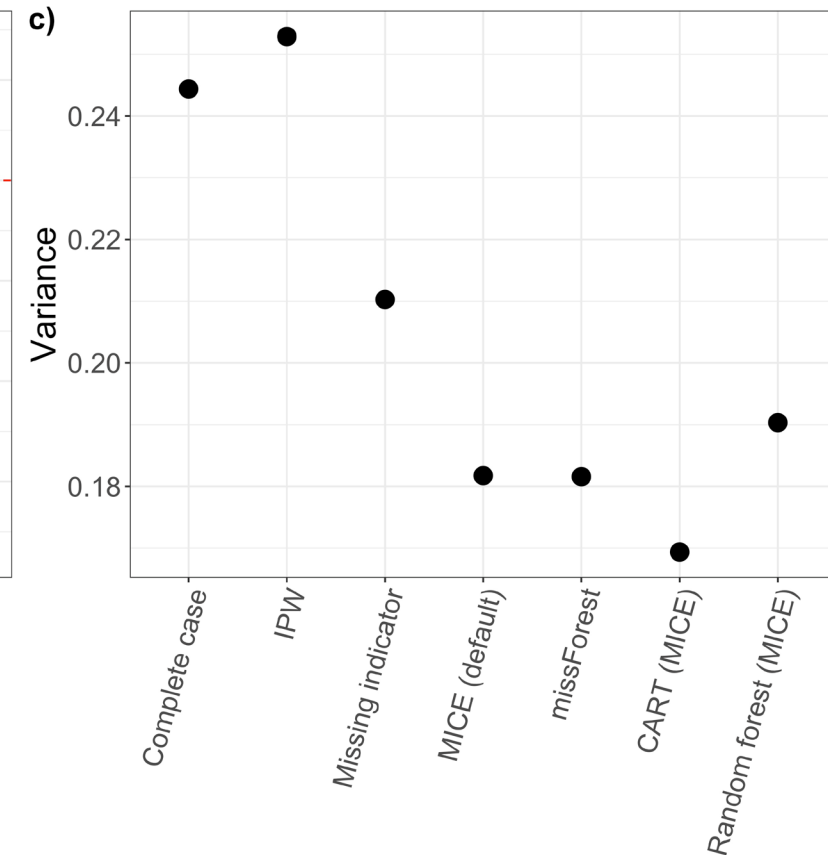
Root mean squared error (RMSE)



Bias

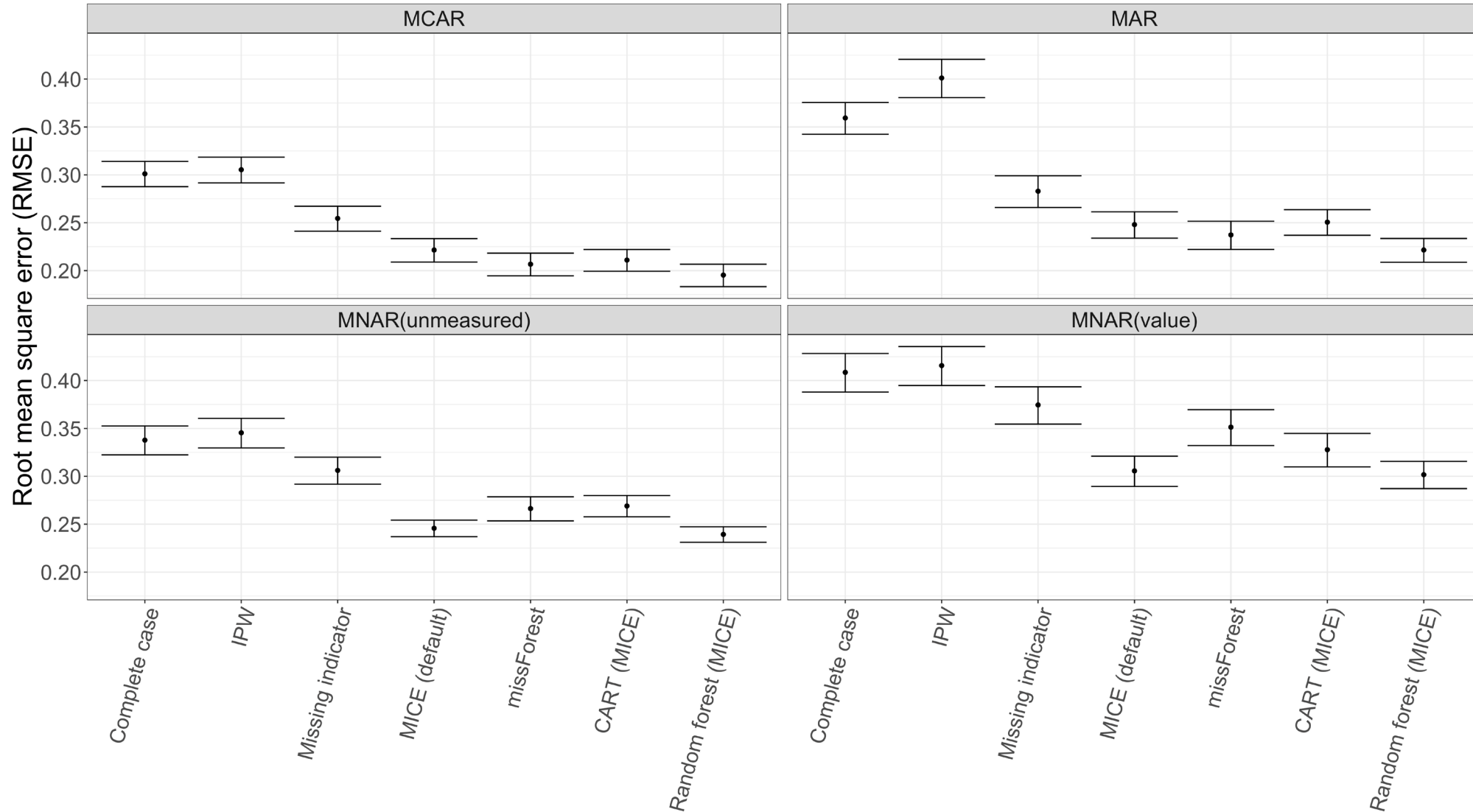


Variance



Analytics Results (by Missingness Mechanism)

- Similar relative performance and patterns across missingness mechanisms
- In comparison, RMSE was highest for all analytic approaches for MNAR(value) mechanisms



smdi R Package for **Diagnostics** implementation

Easy implementation of **routine structural missing data investigations (smdi)**

Package is currently in beta testing, open source and is currently tested at the **Duke University** partner site

```
smdi_diagnose(  
  data = smdi_data,  
  covar = NULL, # NULL includes all covariates with at least one NA  
  model = "cox",  
  form_lhs = "Surv(eventtime, status)"  
) %>%  
smdi_style_gt()
```

Covariate	ASMD (min/max) ¹	p Hotelling ¹	AUC ²	beta univariate (95% CI) ³	beta (95% CI) ³
ecog_cat	0.029 (0.003, 0.071)	0.783	0.510	-0.06 (95% CI -0.16, 0.03)	-0.06 (95% CI -0.16, 0.03)
egfr_cat	0.243 (0.010, 0.485)	<.001	0.629	0.06 (95% CI -0.03, 0.15)	-0.01 (95% CI -0.10, 0.09)
pd1_num	0.062 (0.019, 0.338)	<.001	0.516	0.12 (95% CI 0.01, 0.23)	0.11 (95% CI -0.00, 0.22)

p little: <.001, Abbreviations: ASMD = Median absolute standardized mean difference across all covariates, AUC = Area under the curve, beta = beta coefficient, CI = Confidence interval, max = Maximum, min = Minimum

¹ Group 1 diagnostic: Differences in patient characteristics between patients with and without covariate

² Group 2 diagnostic: Ability to predict missingness

³ Group 3 diagnostic: Assessment if missingness is associated with the outcome (univariate, adjusted)



janickweberpals.gitlab-pages.partners.org/smdi

Thank You

Janick Weberpals

Contact: janick.weberpals@bwh.harvard.edu

R code for this study can be accessed under

<https://gitlab-scm.partners.org/drugapi/missingehr>

Backup slides

Diagnostic results boxplots (distributions across all iterations)

