Sentinel

High-dimensional Multiple Imputation (HDMI) for Partially Observed Confounders Including Natural Language Processing-Derived Auxiliary Covariates

Janick Weberpals, Pamela A. Shaw, Kueiyu Joshua Lin, Richard Wyss, Joseph M Plasek, Li Zhou, Kerry Ngan, Thomas DeRamus, Sudha R. Raman, Bradley G. Hammill, Hana Lee, Darren Toh, John G. Connolly, Kimberly J. Dandreo, Fang Tian, Wei Liu, Jie Li, José J. Hernández-Muñoz, Sebastian Schneeweiss, Rishi J. Desai

Presented at the 2024 ISPE Annual Meeting by: Janick Weberpals
Brigham and Women's Hospital, Harvard Medical School

08/24/24

Disclosures

- This project was supported by Task Order 75F40119F19002 under Master Agreement 75F40119D10037 from the US Food and Drug Administration (FDA).
- Additional funding was provided by NIH RO1LM013204.
- FDA Disclaimer: The contents are those of the authors and do not necessarily represent the official views of, nor and endorsement, by FDA/HHS, or the U.S. Government.
- Some co-authors on this abstract are employed at organizations which conduct work for government and private organizations, including pharmaceutical companies.

Background

- Missing confounder data is a pervasive problem in electronic healthcare databases (+ linkages) when estimating causal treatment effects
- Assumptions on potential missingness mechanism may be empirically checked (smdi)[1,2] along with domain knowledge
- Multiple imputation (MI) has several beneficial characteristics to mitigate bias
  - All patients are retained
  - Flexible modeling (parametric, non-parametric)
  - Can incorporate additional information
  - Realistic variance estimation (Rubin's rule)
- **Assumption: missing at random (MAR)**

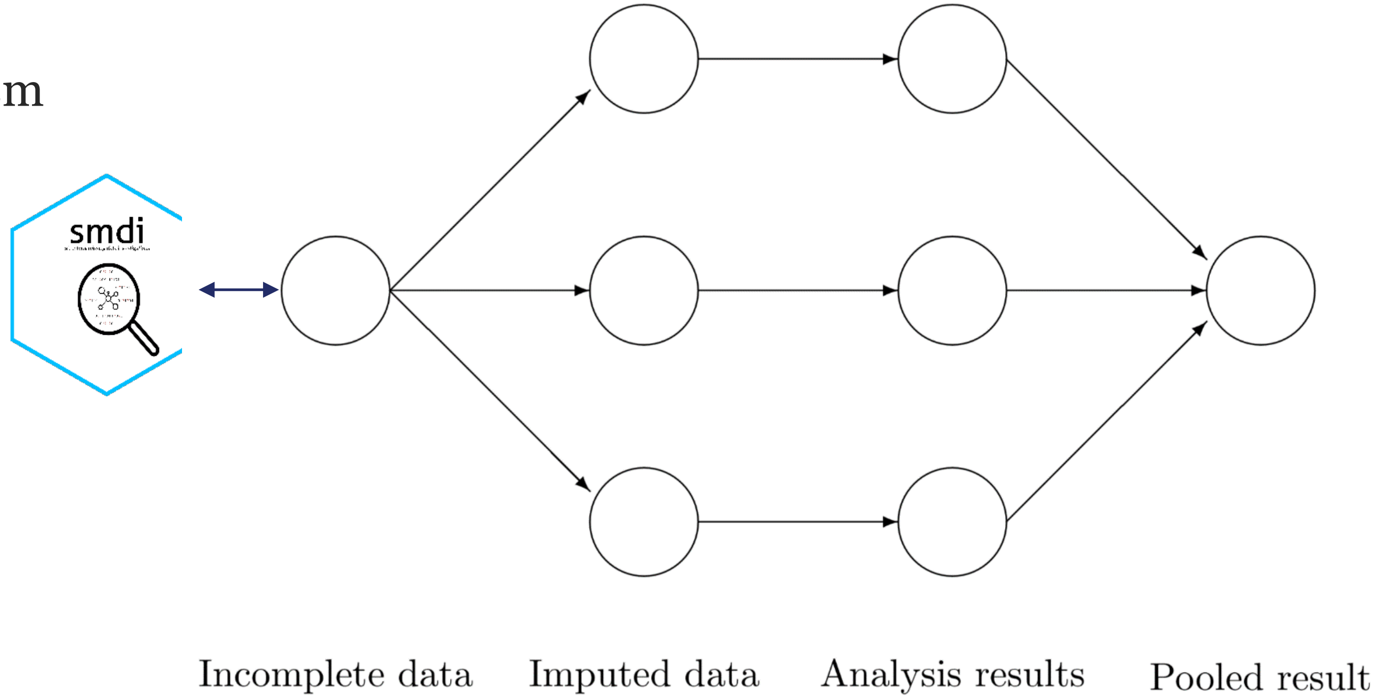
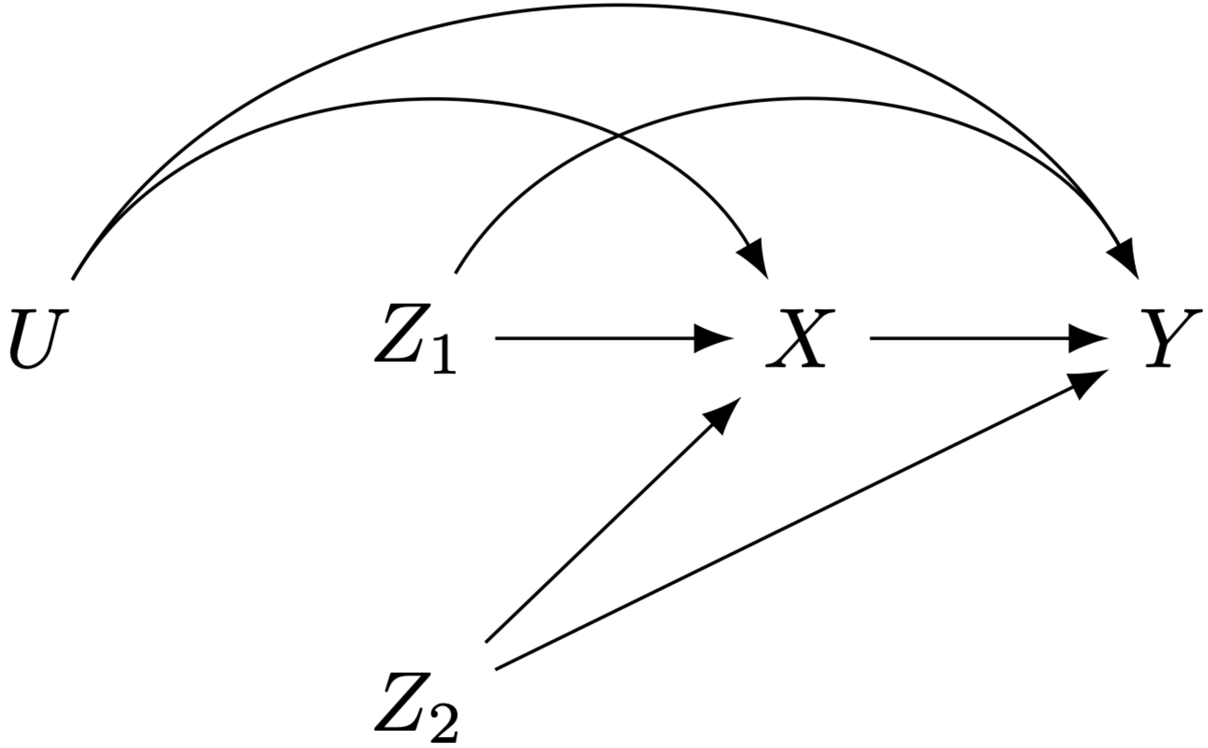


Figure modified from Van Buuren, Stef. Flexible imputation of missing data. CRC press, 2018. https://stefvanbuuren.name/fimd/sec-nutshell.html

[1] Weberpals et al. Clin Epidemiol. 2024 May 21;16:329-343.
[2] Weberpals et al. JAMIA Open. 2024 Jan 31;7(1):ooae008.

Auxiliary Covariates (AC)

- = Covariates that are correlated with the partially observed confounder and possibly related to the missingness of the partially observed confounder, but are not part of the main analysis that estimates the treatment effect
- Inclusion of AC in MI model
  - ➢ Increases statistical efficiency
  - ➢ **Reduces Bias by making the MAR assumption more likely**
- **Problem: Data-adaptive approaches to identify AC for MI models are not well understood**



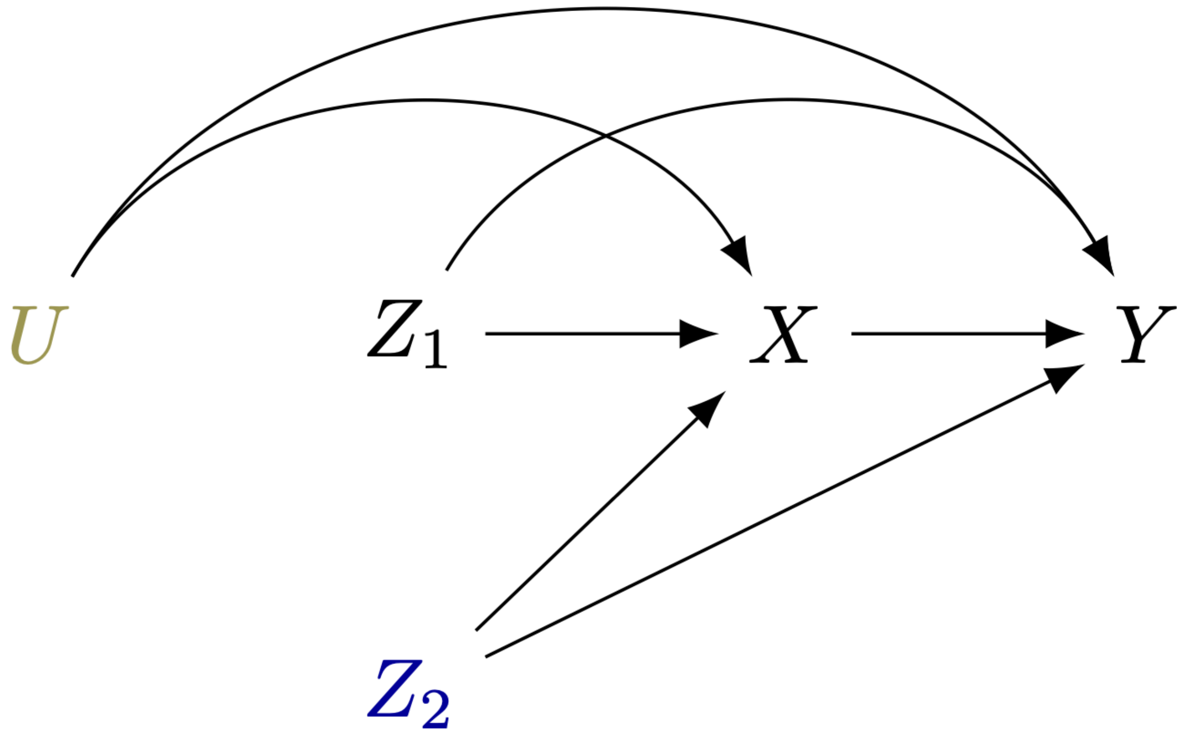
U = Unmeasured confounder, X = Exposure, Y = Outcome, Z1 = Completely observed confounders, Z2 = Partially observed confounder(s)

Canonical DAGs adapted from *Moreno-Betancur et al. American Journal of Epidemiology, 187(12), 2705–2715. https://doi.org/10.1093/aje/kwy173*

High-dimensional Multiple Imputation (HDMI)

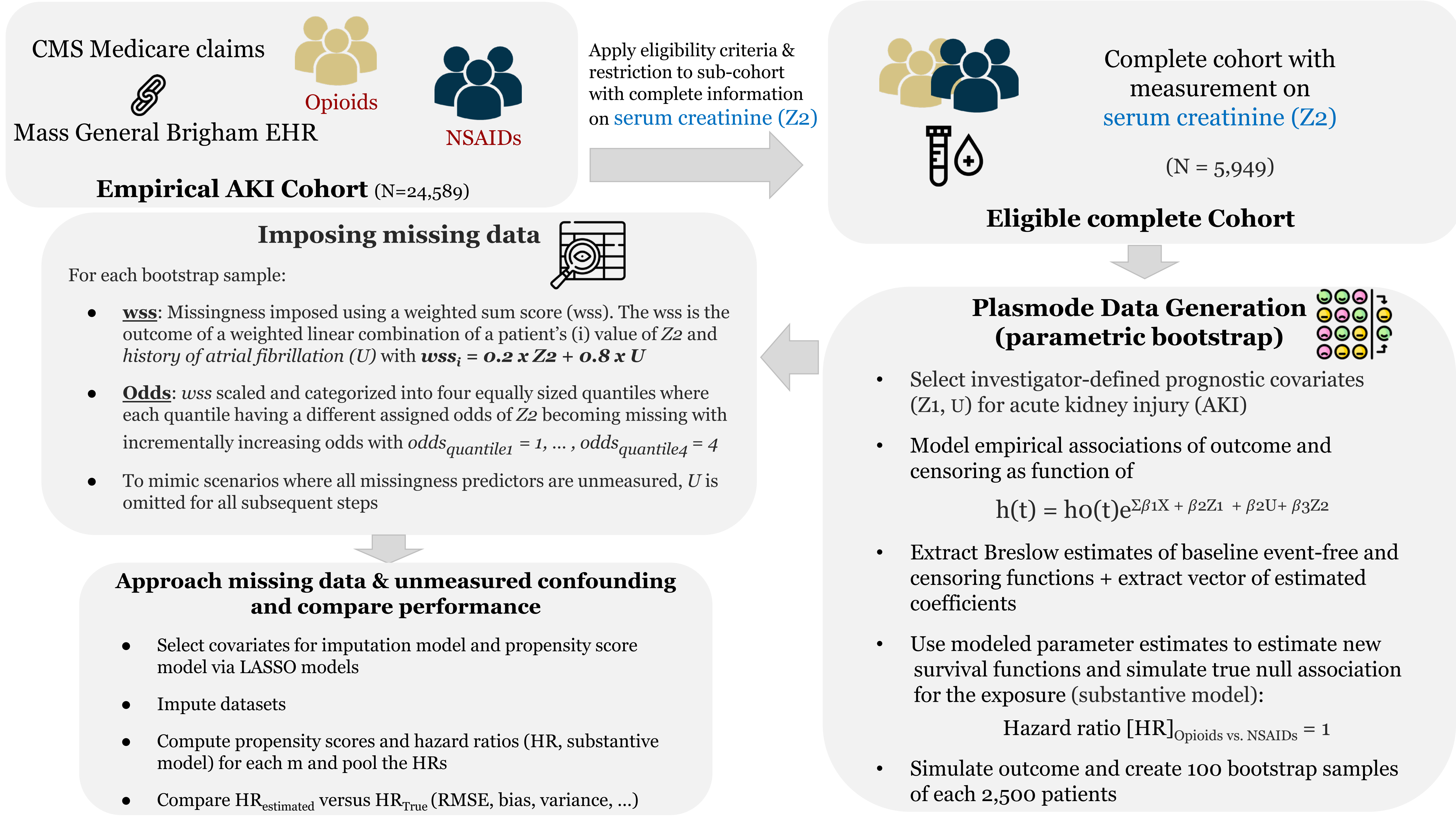
- Idea: High-dimensional data (structured + unstructured) to systematically identify and prioritize ACs that can approximate …
  - Potentially unobserved reasons for missingness in partially observed confounders (and thereby mitigate bias by missing not at random mechanisms)
  - Completely unobserved confounders (see HDPS Schneeweiss et al., Epidemiology 2009;20: 512–522)

- Hypothesis: HDMI can increase statistical efficiency and reduce bias in settings where missingness depends on unobserved factors



U = Unmeasured confounder, X = Exposure, Y = Outcome, Z1 = Completely observed confounders, Z2 = Partially observed confounder(s)

Weberpals, et al. arXiv preprint arXiv:2405.10925 (2024).



CMS Medicare claims

Mass General Brigham EHR

Opioids

NSAIDs

Empirical AKI Cohort (N=24,589)

Apply eligibility criteria & restriction to sub-cohort with complete information on serum creatinine (Z2)

Complete cohort with measurement on serum creatinine (Z2)

(N = 5,949)

Eligible complete Cohort

Plasmode Data Generation (parametric bootstrap)

- Select investigator-defined prognostic covariates (Z1, U) for acute kidney injury (AKI)
- Model empirical associations of outcome and censoring as function of

$$h(t) = h_0(t)e^{\Sigma\beta_1 X + \beta_2 Z_1 + \beta_2 U + \beta_3 Z_2}$$

- Extract Breslow estimates of baseline event-free and censoring functions + extract vector of estimated coefficients
- Use modeled parameter estimates to estimate new survival functions and simulate true null association for the exposure (substantive model):

$$\text{Hazard ratio } [\text{HR}]_{\text{Opioids vs. NSAIDs}} = 1$$

- Simulate outcome and create 100 bootstrap samples of each 2,500 patients

Imposing missing data

For each bootstrap sample:

- **wss**: Missingness imposed using a weighted sum score (wss). The wss is the outcome of a weighted linear combination of a patient's (i) value of *Z2* and *history of atrial fibrillation (U)* with $\boldsymbol{wss_i = 0.2 \times Z2 + 0.8 \times U}$
- **Odds**: *wss* scaled and categorized into four equally sized quantiles where each quantile having a different assigned odds of *Z2* becoming missing with incrementally increasing odds with $odds_{quantile1} = 1, \ldots, odds_{quantile4} = 4$
- To mimic scenarios where all missingness predictors are unmeasured, *U* is omitted for all subsequent steps

Approach missing data & unmeasured confounding and compare performance

- Select covariates for imputation model and propensity score model via LASSO models
- Impute datasets
- Compute propensity scores and hazard ratios (HR, substantive model) for each m and pool the HRs
- Compare $\text{HR}_{\text{estimated}}$ versus HR_{True} (RMSE, bias, variance, …)

Specification of Data Dimensions

Structured

Complete case & Baseline model

Investigator-derived Z1 covariates

HDMI claims

Structured binary empirical claims covariates

Unstructured

HDMI unigram

Binary indicator covariates for presence of a word

{-23.4, 5.2, -4.56, 0.51, ...}

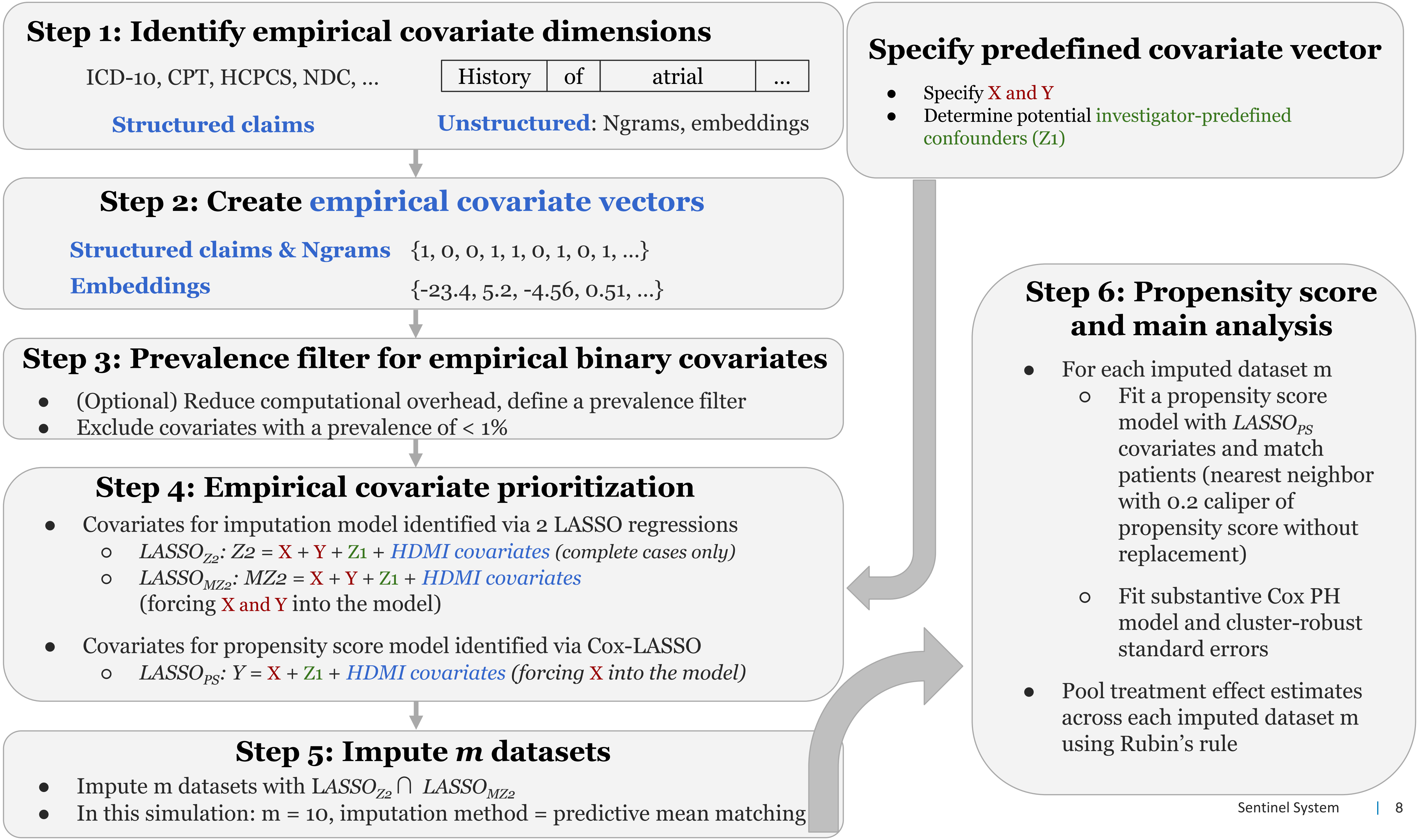
HDMI sentence

Mean pooled BERT sentence embeddings

$N_{\text{Baseline}} = 13$ $N_{\text{HDMI claims}} = 28{,}874$ $N_{\text{HDMI unigrams}} = 19{,}993$ $N_{\text{HDMI sentence}} = 128$

Number of candidate covariates

Weberpals, et al. arXiv preprint arXiv:2405.10925 (2024).



Step 1: Identify empirical covariate dimensions

ICD-10, CPT, HCPCS, NDC, ...

History | of | atrial | ...

Structured claims

Unstructured: Ngrams, embeddings

Step 2: Create empirical covariate vectors

Structured claims & Ngrams {1, 0, 0, 1, 1, 0, 1, 0, 1, ...}

Embeddings {-23.4, 5.2, -4.56, 0.51, ...}

Step 3: Prevalence filter for empirical binary covariates

- (Optional) Reduce computational overhead, define a prevalence filter
- Exclude covariates with a prevalence of < 1%

Step 4: Empirical covariate prioritization

- Covariates for imputation model identified via 2 LASSO regressions
  - $LASSO_{Z2}$: *Z2 = X + Y + Z1 + HDMI covariates (complete cases only)*
  - $LASSO_{MZ2}$: *MZ2 = X + Y + Z1 + HDMI covariates*
    (forcing X and Y into the model)
- Covariates for propensity score model identified via Cox-LASSO
  - $LASSO_{PS}$: *Y = X + Z1 + HDMI covariates (forcing X into the model)*

Step 5: Impute *m* datasets

- Impute m datasets with $LASSO_{Z2} \cap LASSO_{MZ2}$
- In this simulation: m = 10, imputation method = predictive mean matching

Specify predefined covariate vector

- Specify X and Y
- Determine potential investigator-predefined confounders (Z1)

Step 6: Propensity score and main analysis

- For each imputed dataset m
  - Fit a propensity score model with $LASSO_{PS}$ covariates and match patients (nearest neighbor with 0.2 caliper of propensity score without replacement)
  - Fit substantive Cox PH model and cluster-robust standard errors
- Pool treatment effect estimates across each imputed dataset m using Rubin's rule

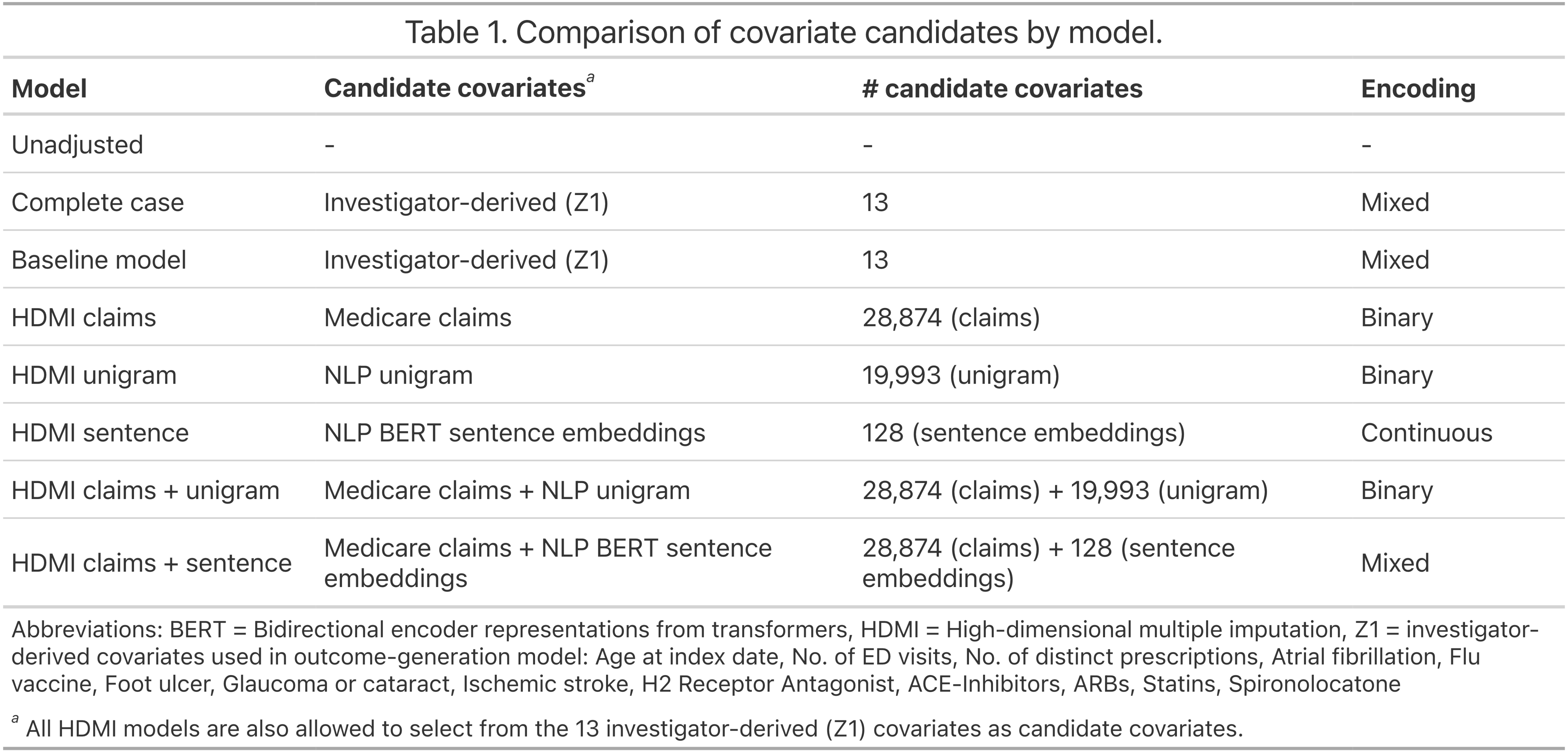


Table 1. Comparison of covariate candidates by model.

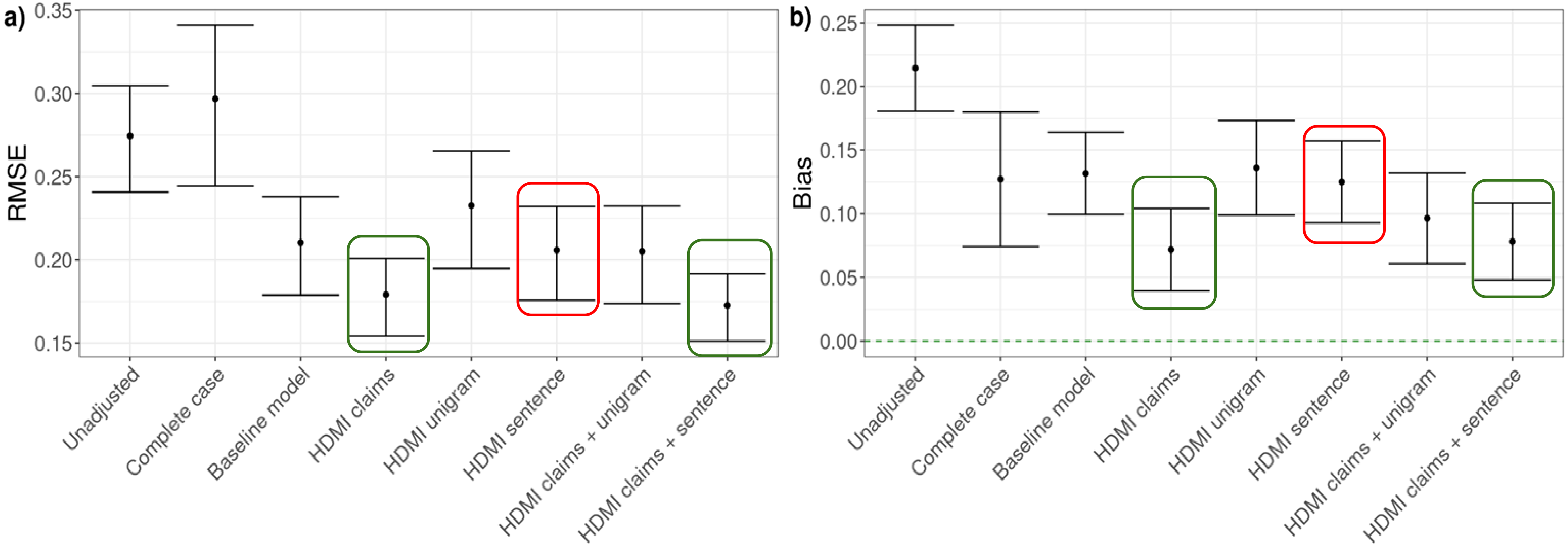
Model	Candidate covariates[a]	# candidate covariates	Encoding
Unadjusted	-	-	-
Complete case	Investigator-derived (Z1)	13	Mixed
Baseline model	Investigator-derived (Z1)	13	Mixed
HDMI claims	Medicare claims	28,874 (claims)	Binary
HDMI unigram	NLP unigram	19,993 (unigram)	Binary
HDMI sentence	NLP BERT sentence embeddings	128 (sentence embeddings)	Continuous
HDMI claims + unigram	Medicare claims + NLP unigram	28,874 (claims) + 19,993 (unigram)	Binary
HDMI claims + sentence	Medicare claims + NLP BERT sentence embeddings	28,874 (claims) + 128 (sentence embeddings)	Mixed

Abbreviations: BERT = Bidirectional encoder representations from transformers, HDMI = High-dimensional multiple imputation, Z1 = investigator-derived covariates used in outcome-generation model: Age at index date, No. of ED visits, No. of distinct prescriptions, Atrial fibrillation, Flu vaccine, Foot ulcer, Glaucoma or cataract, Ischemic stroke, H2 Receptor Antagonist, ACE-Inhibitors, ARBs, Statins, Spironolocatone

[a] All HDMI models are also allowed to select from the 13 investigator-derived (Z1) covariates as candidate covariates.

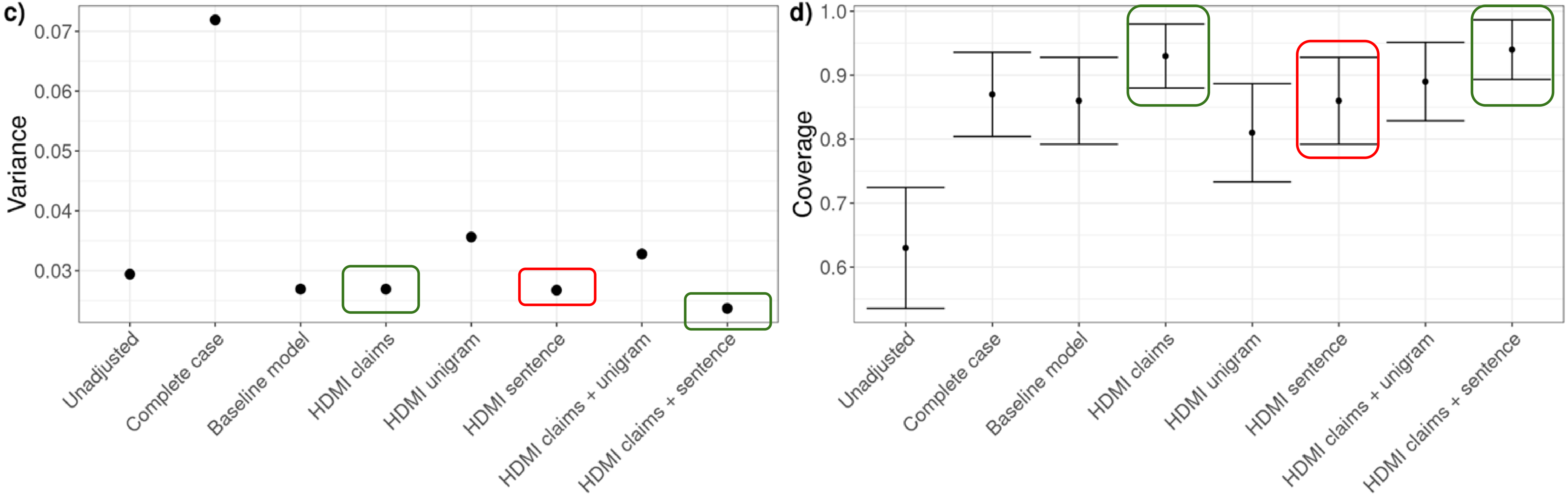
Weberpals, et al. arXiv preprint arXiv:2405.10925 (2024).

HDMI Main Results: Illustrating the a) root-mean-squared-error (RMSE), b) bias, c) variance and d) coverage of the nominal 95% confidence interval (CI) between analytical methods to account for partially observed serum creatinine (Z2) measurements and unmeasured confounding



Weberpals, et al. arXiv preprint arXiv:2405.10925 (2024).

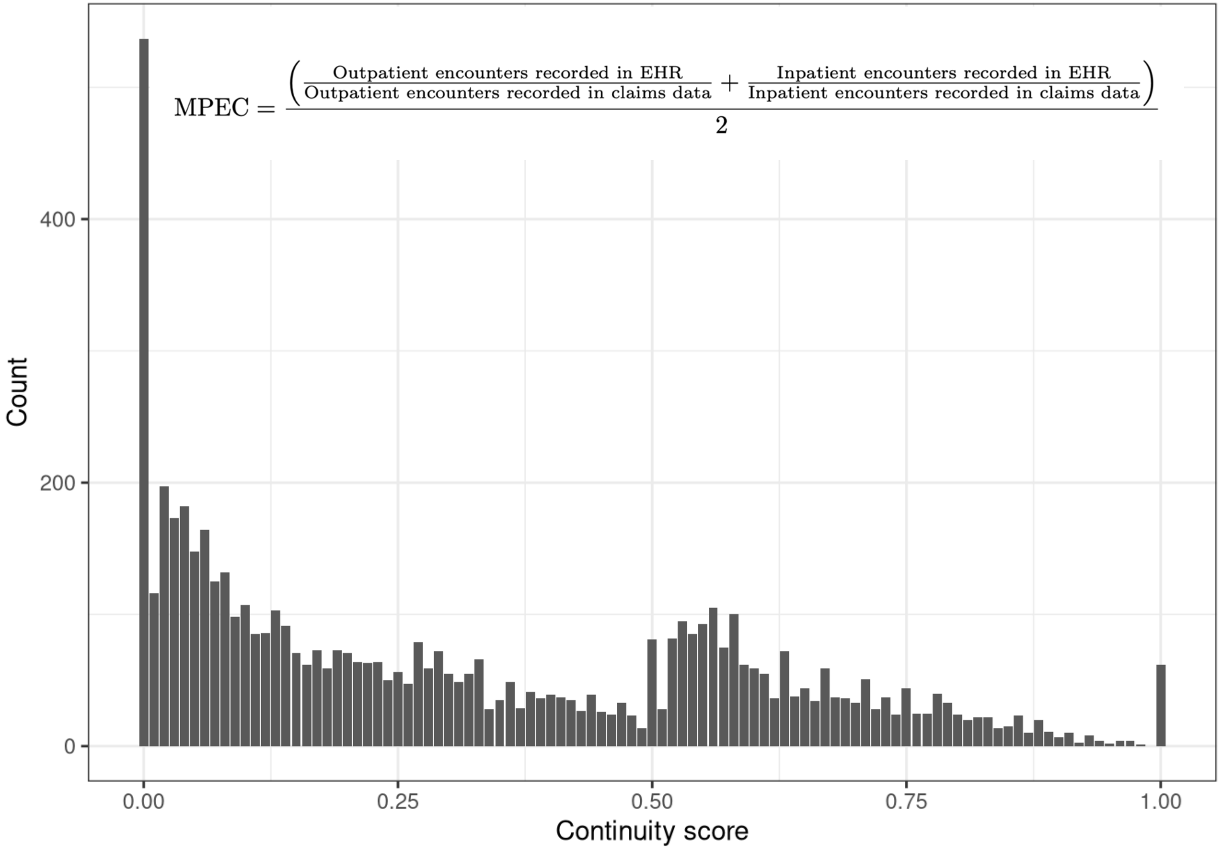
HDMI Main Results Continued: Illustrating the a) root-mean-squared-error (RMSE), b) bias, c) variance and d) coverage of the nominal 95% confidence interval (CI) between analytical methods to account for partially observed serum creatinine (Z2) measurements and unmeasured confounding



Weberpals, et al. arXiv preprint arXiv:2405.10925 (2024).

Data Continuity in Electronic Healthcare Records

- Continuity score: mean proportion of encounters captured in linked EHR-claims data[1] among all patients in the eligible complete cohort
- *Mass General Brigham* is a tertiary care provider
- **Lack of observability of EHR data for a larger proportion of patients**
- Similar observation: prediction performance of clinical risk scores is substantially worse in patients with lower vs. high EHR-continuity[2]



$$\text{MPEC} = \frac{\left(\frac{\text{Outpatient encounters recorded in EHR}}{\text{Outpatient encounters recorded in claims data}} + \frac{\text{Inpatient encounters recorded in EHR}}{\text{Inpatient encounters recorded in claims data}}\right)}{2}$$

Weberpals, et al. arXiv preprint arXiv:2405.10925 (2024).

[1] Lin KJ et al., Clin Pharmacol Ther 2018

[2] Jin Y, Weberpals J, et al., Clin Pharmacol Ther 2023

Conclusions

- HDMI approaches can decrease bias and increase statistical efficiency in studies with partially observed confounders where missingness depends on unobserved factors
- Practicality depends on access to different data dimensions

- Future directions:
  - Gain more experience in applied studies
  - Streamline implementation using R package (in development)
  - Explore other data modalities, e.g., radiomics/imaging data, digital biomarkers, etc.

Weberpals, et al. arXiv preprint arXiv:2405.10925 (2024).

Sentinel

Study repository

https://gitlab-scm.partners.org/drugepi/hdmi-manuscript

Study protocol & report with annotated R code

https://drugepi.gitlab-pages.partners.org/hdmi-manuscript/

Data Availability

Original CMS data cannot be shared but simulated using the *generate_data()* function, see:

https://drugepi.gitlab-pages.partners.org/bias_simulation_missing_data/

Sentinel

Thank You

Questions?