uOttawa

SYNTHETIC DATA AUGMENTATION FOR MITIGATING BIAS IN REAL WORLD DATA

Presented by:

Lamin Juwara
Postdoctoral Researcher
University of Ottawa

Outline

Introduction to Data Bias	1	Define data bias, how it is induced, and some common problems
Examples in Biomedical Research	2	Give some examples in the media and in biomedical research.
Approaches for Mitigating Bias	3	An overview of bias mitigation approaches and the proposed Synthetic Minority Augmentation approach
Model Evaluation & Applications	4	Describe model training and evaluation. Applications to simulated data and case studies.
Conclusions	5	Summarize the study findings and limitations.

What is data bias?

- Data bias is pervasive in biomedical research, especially in large-scale observational datasets.
- In these settings, the rules that govern group assignment are generally unknown or without proper design.

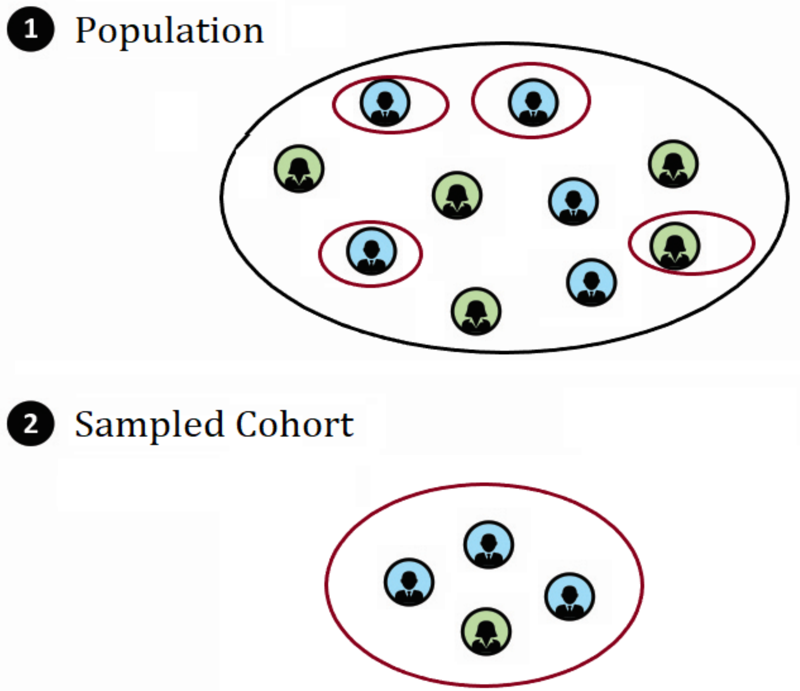


Fig 1. (1)->(2) Hypothetical example of sample selection bias

How data bias occurs

- For example, a sex variable where women are under-represented compared to the population
- Such biases can occur at the data collection or analysis stage:
  - difficulty in collecting data from certain groups due to cost, access, or non-response
  - the data collection process is inherently biased
  - by excluding certain groups during analysis

- It is different from missingness -- entire records are missing instead of specific observations within collected records

Popular examples

Global NEWS

Watch World Canada Local Politics Money Health

WORLD

Amazon ditches AI recruiting tool that didn't like women

By Jeffrey Dastin • Reuters

Posted October 10, 2018 6:46 am

THE GLOBE AND MAIL

INVESTIGATION

Bias behind bars: A Globe investigation finds a prison system stacked against Black and Indigenous inmates

Federal inmates' risk assessments determine everything from where a prisoner is incarcerated to what rehabilitation programs they are offered. After controlling for a number of variables, The Globe found Black and Indigenous inmates are more likely to get worse scores than white inmates, based solely on their race

TOM CARDOSO

PUBLISHED OCTOBER 24, 2020

UPDATED NOVEMBER 11, 2020

Racial bias found in widely used health care algorithm

An estimated 200 million people are affected each year by similar tools that are used in hospital networks

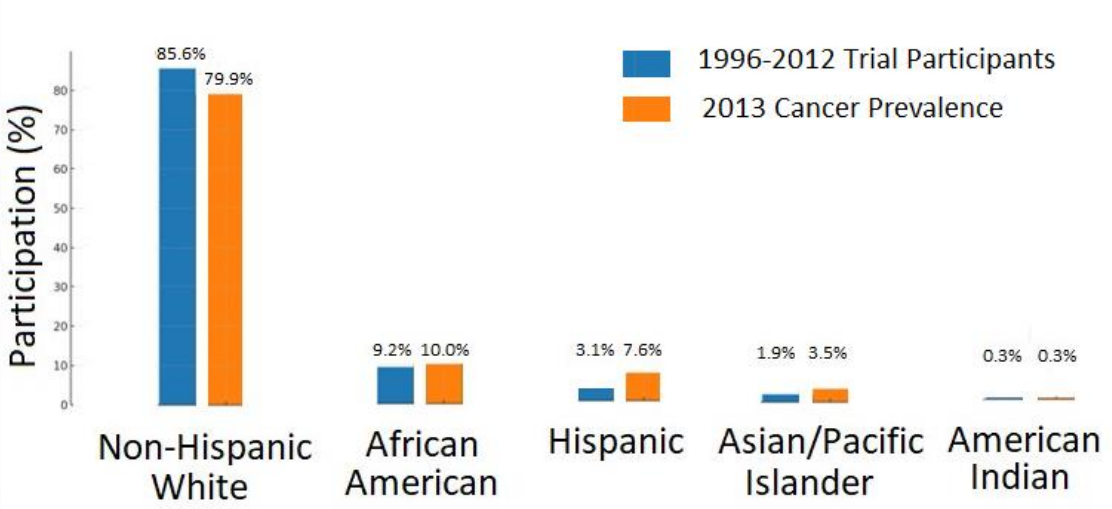
Nov. 6, 2019, 2:38 PM EST / Updated Nov. 7, 2019, 11:07 AM EST

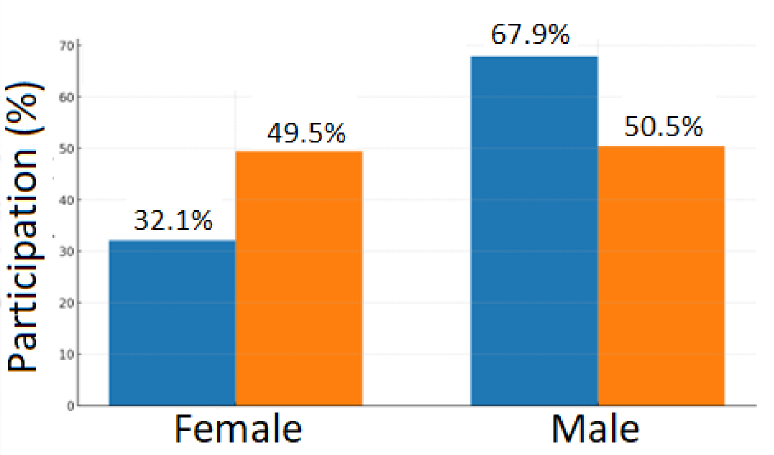
By Quinn Gawronski

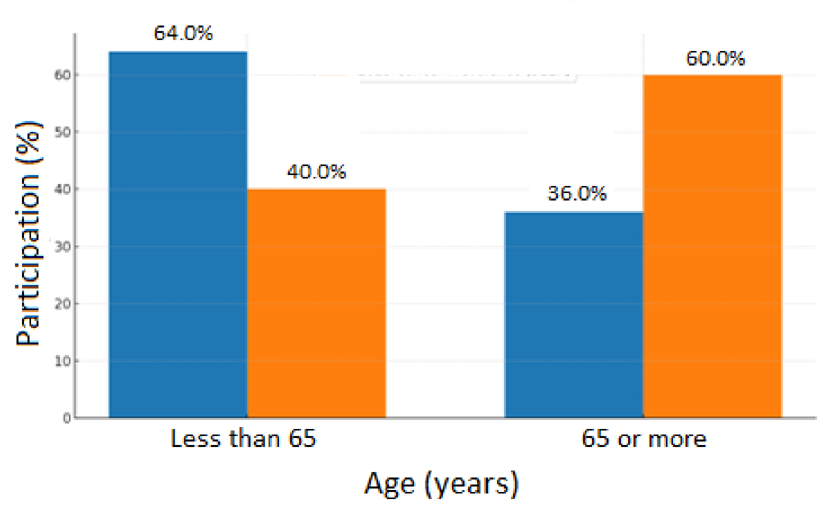
NBC NEWS

Examples in biomedical research

Participants in all Therapeutic Cancer Trials, 1996-2012 (N = 52,170)







Duma, N., et al. "Representation of minorities and women in oncology clinical trials: review of the past 14 years. J Oncol Pract. 2018; 14 (1): e1–e10." Duma et al. conduct a survey of 1012 (2017).

uOttawa

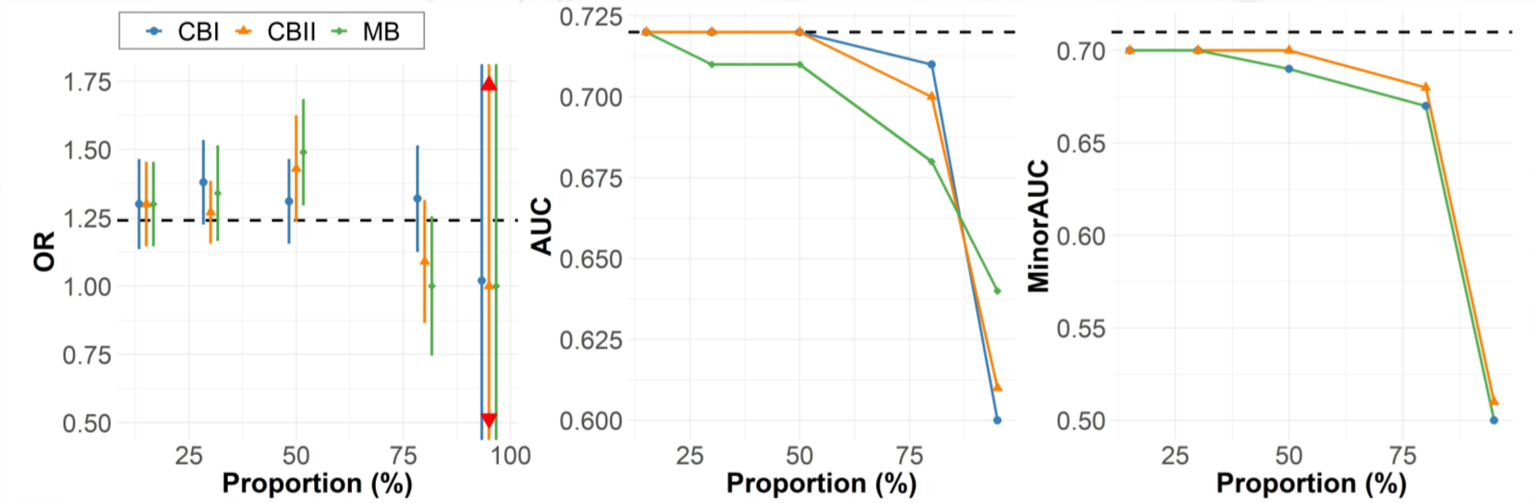
Classification of biases

Type of Bias	Description	Example
Marginal bias	observations from a specific group are omitted from the sampled dataset based solely on the biased variable.	exclude females irrespective of other covariates in the data
Conditional bias I	occurs when an additional covariate that is weakly associated with the biased variable influences the exclusion	exclude female participants with low education level
Conditional bias II	an additional covariate that is strongly associated with the biased variable influences the exclusion	exclude female participants in low income category

Problems with biased datasets

Bias in the training cohort results in:

- Imprecise predictions
- Inconsistent estimations
- Biased estimates of covariate effects



MB = Marginal bias; CBI = Conditional Bias I; CBII = Conditional Bias II.

Why it matters

Representation in biomedical data:

- Ensures results are applicable to the broader population.
- Helps identify potential differences in outcomes. e.g., differences in treatment responses to certain medications.
- From an ethical standpoint, all groups should have a fair participation opportunity.

Mitigating Data Bias

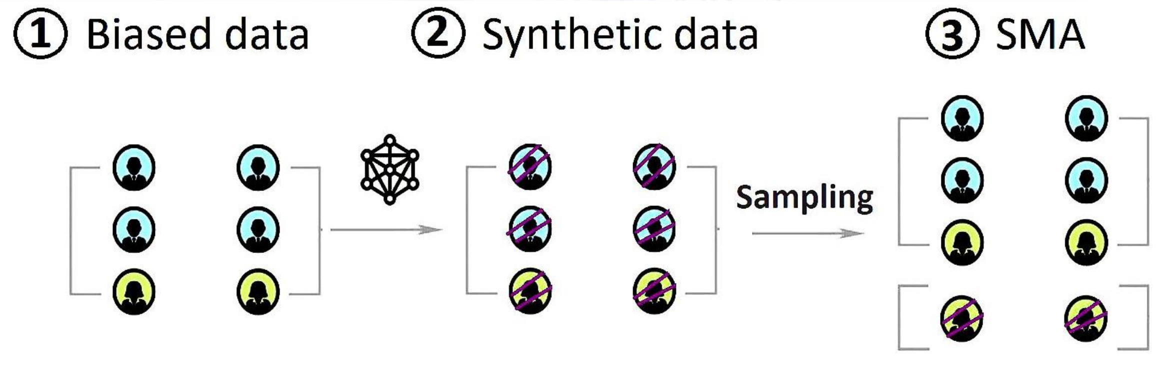
Approaches for Mitigating Data Bias

- Random oversampling (ROS) and undersampling (RUS)
- SMOTE
- Propensity score (PS) methods (e.g., PS- matching)
- RF ensembles
- Proposed: Synthetic Minor Augmentation (SMA)

Proposed Approach

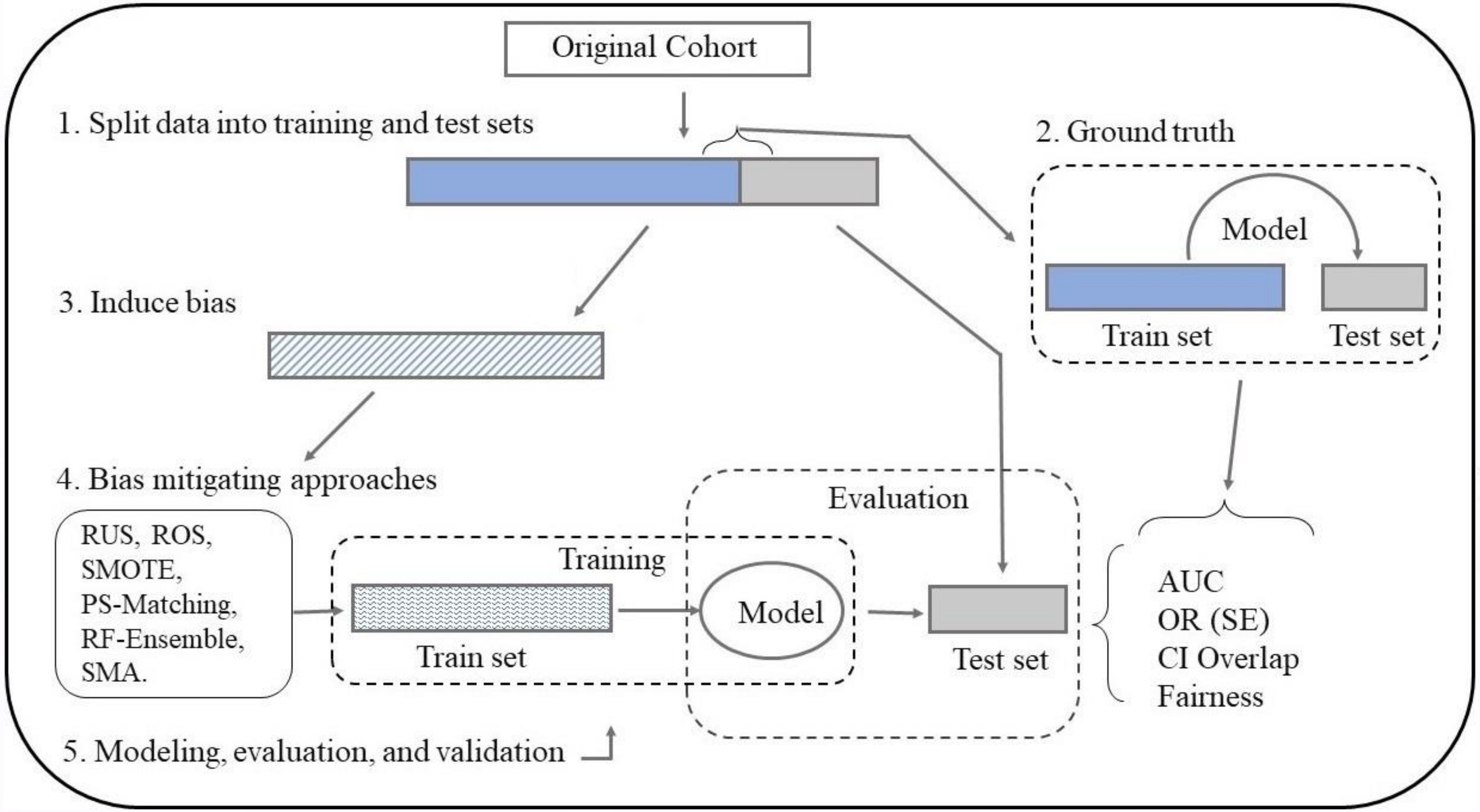
Synthetic Minor Augmentation (SMA)

Steps:



1. Construct a synthetic version of the biased data using sequential synthesis based on gradient boosting decision trees.
2. Sample observations from the bias-inducing (i.e., minor or underrepresented) partition of the generated synthetic dataset.
3. Augment the samples with original biased data to create a complete dataset.

Model Training & Evaluation



uOttawa

Applications

- We perform two types of analyses:
  - Simulation studies
  - Four real datasets
- The analytical workload assumed is a binary logistic regression model

1) Simulation studies

Simulate a binary Outcome data:

- We postulate the logistic regression model:

$$P(Y=1) = \text{expit}(\alpha + \beta_Z Z + \beta_{X1} X_1 + \beta_{X2} X_2 + \beta_{ZX2} Z X_2 + \beta_U U)$$

- Z~binomial(p=0.5), $X_2|Z=1$~binomial(p=0.4), and $X_2|Z=0$~binomial(p=0.39)
- U~log-Normal(12,3.5), a strong predictor of Y and independent of Z, X_1, X_2
- Set of parameters:

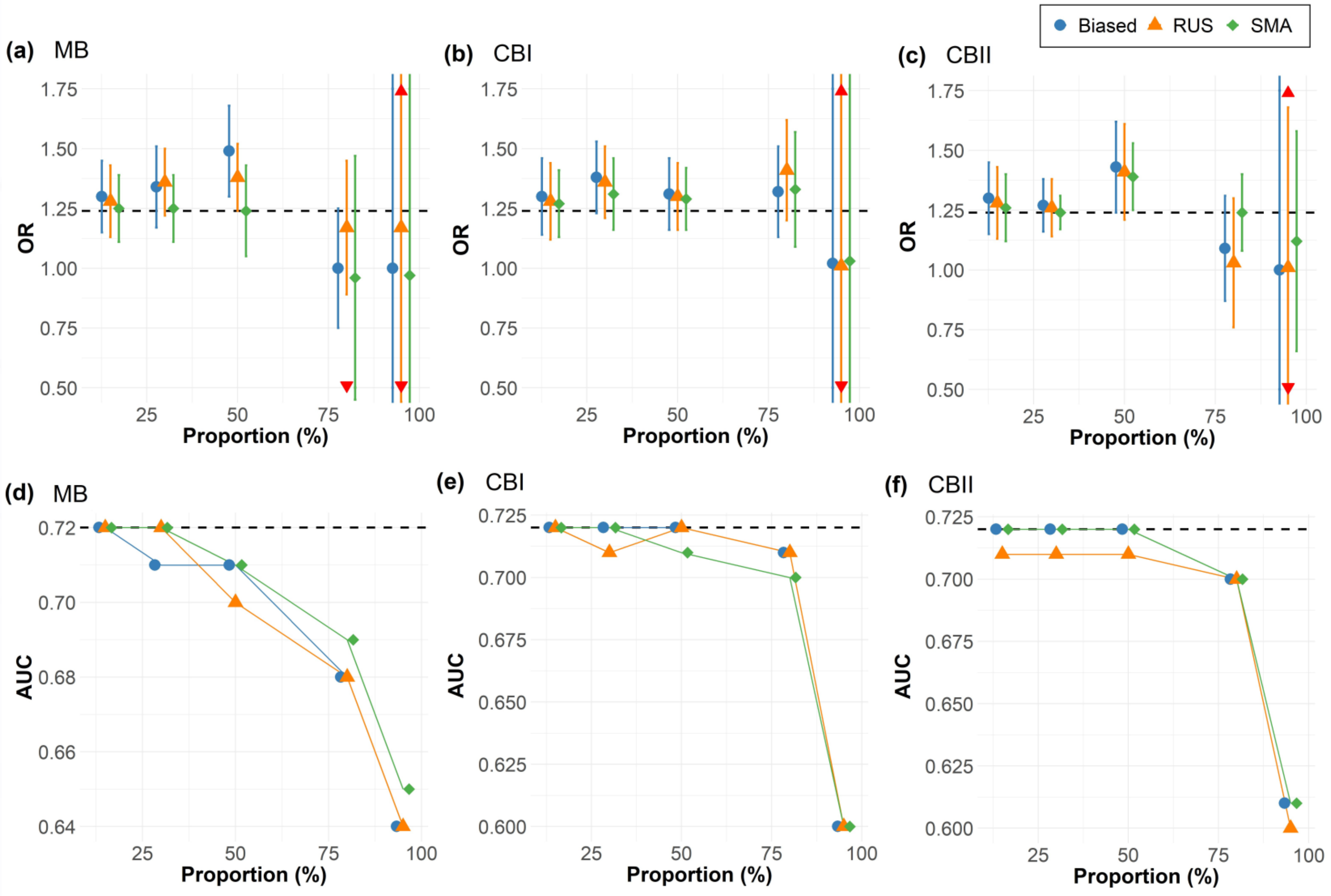
  $\alpha=\log(0.5)$, $\boldsymbol{\beta_Z = \log(1.25)}$, $\beta_{X1} = \log(0.3)$, $\beta_{X2} = \log(2)$, $\beta_{ZX2} = -0.47$, and $\beta_U = \log(0.5)$.

- Generate 500 data cohorts of n=5000 each.

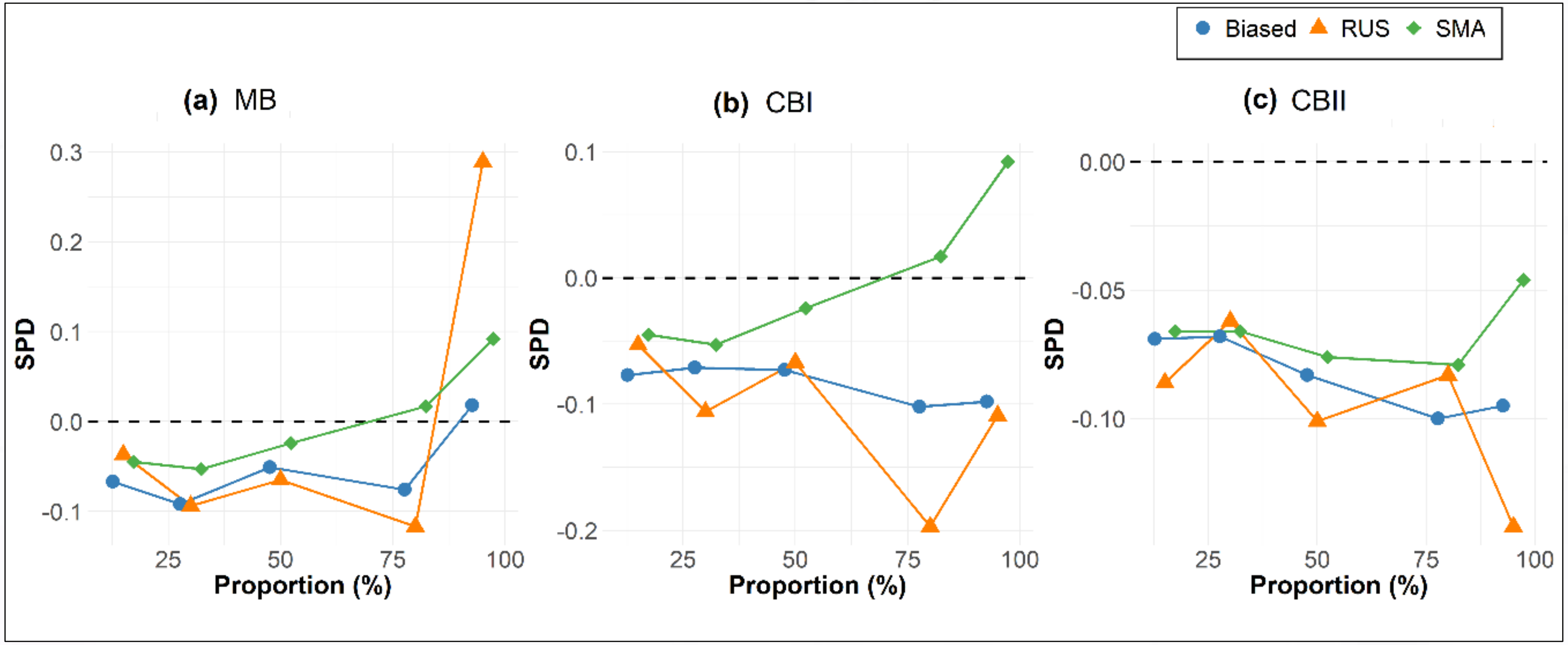
NB. Values of β_Z were also varied to assess robustness.

Odds ratio and AUC estimates

MB = Marginal bias; CBI = Conditional Bias I; CBII = Conditional Bias II.



Fairness: Statistical Parity Difference (SPD)



MB = Marginal bias; CBI = Conditional Bias I; CBII = Conditional Bias II.

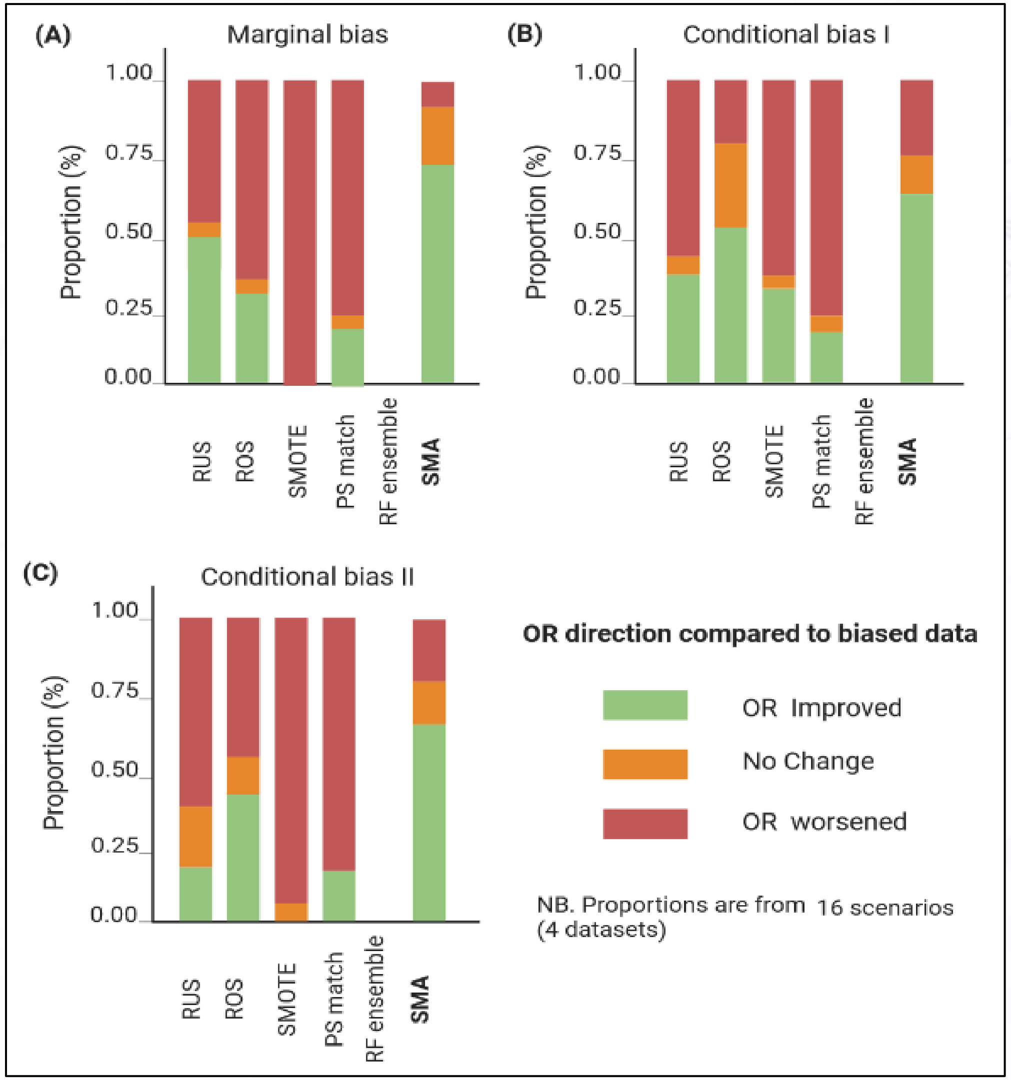
2) Real datasets

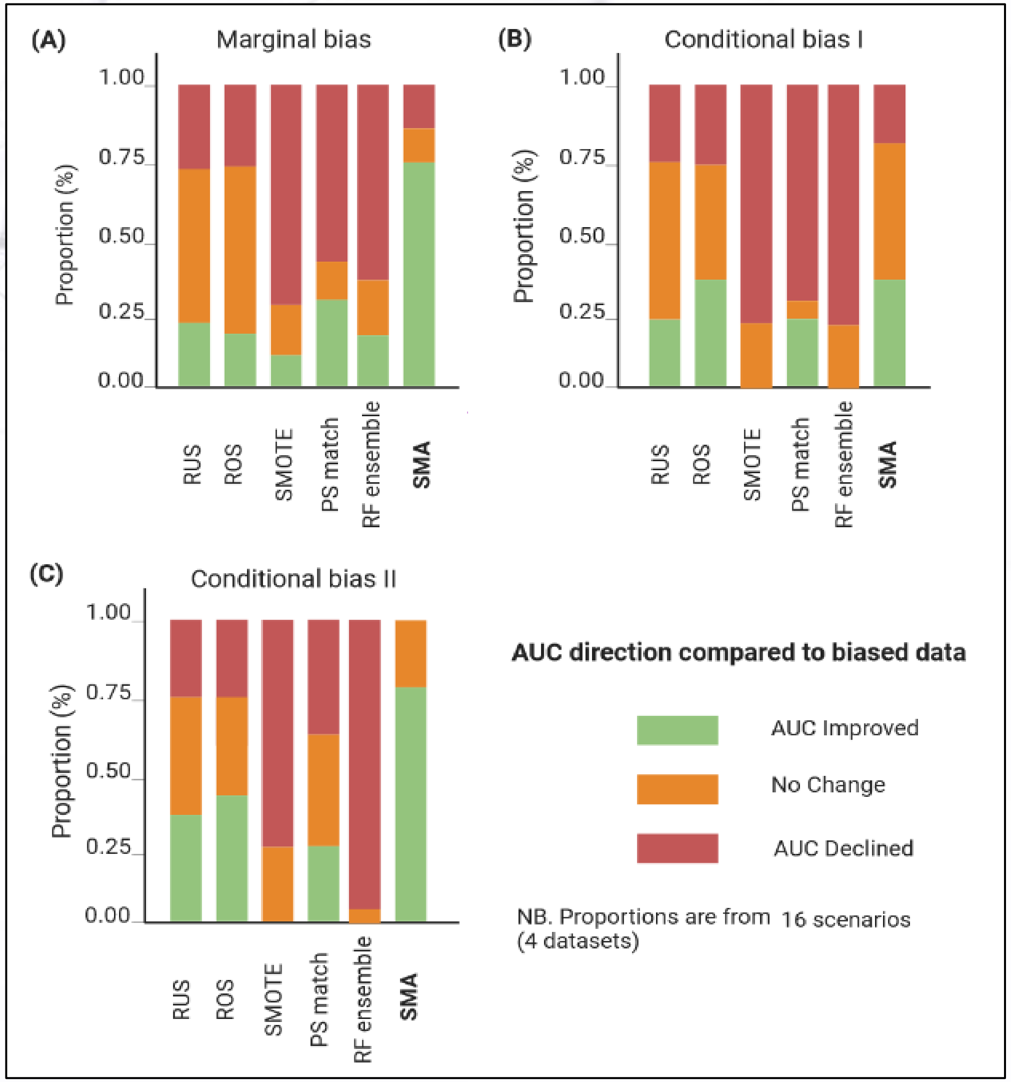
Summary of datasets

Dataset	Description	Outcome	Biased covariate	Conditioning Covariate
1) Cardiovascular Health (CCHS)	63522 observations 8 variables	CCHS status	Gender (Female = 45.2%)	CBI: New immigrants CBII: Marital status
2) N0147 Colon Cancer trial (PDS)	1543 observations 10 variables	Death	Bowel obstruction (No = 83.8%)	CBI: Gender CBII: BMI
3) Danish Colon Cancer data (DCCG)	12855 observations 192 variables (total) 9 selected	Postoperative complications	Gender (Female = 55.9%)	CBI: P-PN stage CBII: ASA
4) Breast Cancer (UCI)	277 observations 10 variables	BC class	Age (20-49 = 45.5%)	CBI: Left/Right breast CBII: Menopause

NB. CBI represents Conditional Bias I and CBII is Conditional Bias II.

Summaries for all datasets: Odds ratio and AUC





Conclusions

- Model parameters are significantly affected by bias
- AUC is not significantly affected by bias
- In low to medium bias severity (less than 50% missing proportion), SMA produces the results with:
  - the least bias (difference between the model estimate and ground truth).
  - the best precision (smallest standard errors) in estimating the regression coefficient than other approaches.
- Above 50% bias, there isn’t an obvious best method
- Above 80% bias, mitigation methods generally perform poorly – it is difficult to compensate for extreme bias irrespective of the method is chosen
- SMA gives the best fairness estimates among groups

How should SMA be adopted?

- Use as a sensitivity analysis tool
- If the biased mitigated estimates matches the biased estimates, the results could be reported with more confidences
- If the mitigated results are different, the results should be reported with caution
- Ideally, steps should be taken to recruit more individuals

Questions?

Notes on the synthesis stage

- The type of generative model used was a sequential tree-based synthesizer
- Each model in the sequence was trained using a gradient boosted decision tree
- Bayesian optimization for hyperparameter selection
- Each combination of hyperparameters was evaluated using 5-fold cross validation on the training dataset during tuning.
- For the synthesis of categorical variables, synthetic values are generated based on predicted probabilities.
- boosted trees do not output correct probabilities and these need to be calibrated, especially as the number of iterations increases
- For example, beta calibration for imbalanced categorical outcomes.