SESSION 6: NEW APPLICATIONS

SYNTHETIC DATA AUGMENTATION FOR MITIGATING BIAS IN REAL WORLD DATA

uOttawa

Replica Analytics
AN AETION COMPANY

Presented by:

Lamin Juwara,
Postdoctoral Researcher,
Electronic Health Information Laboratory (EHIL),
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.

Replica Analytics
AN AETION COMPANY

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.

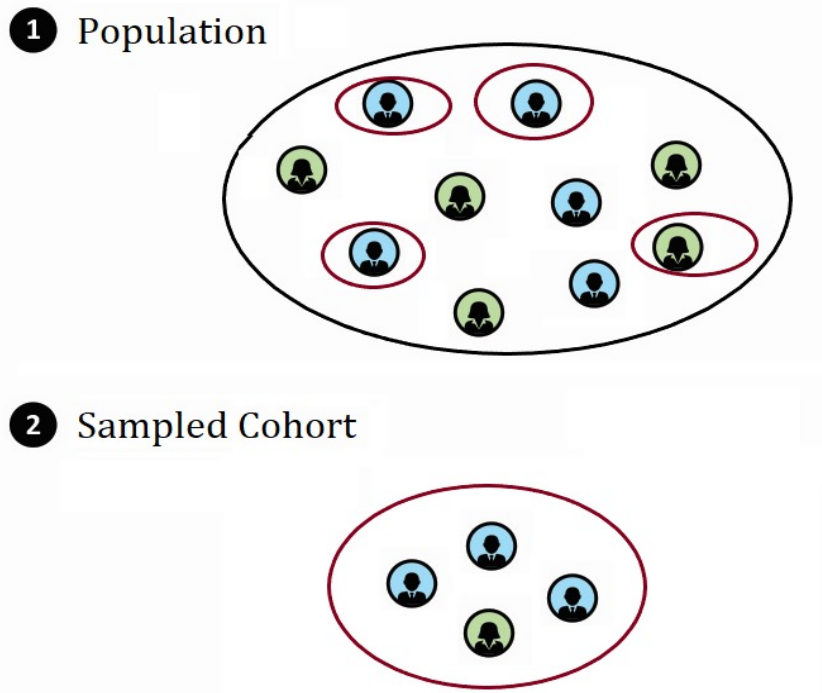


Fig 1. (1)->(2) Hypothetical example of sample selection bias

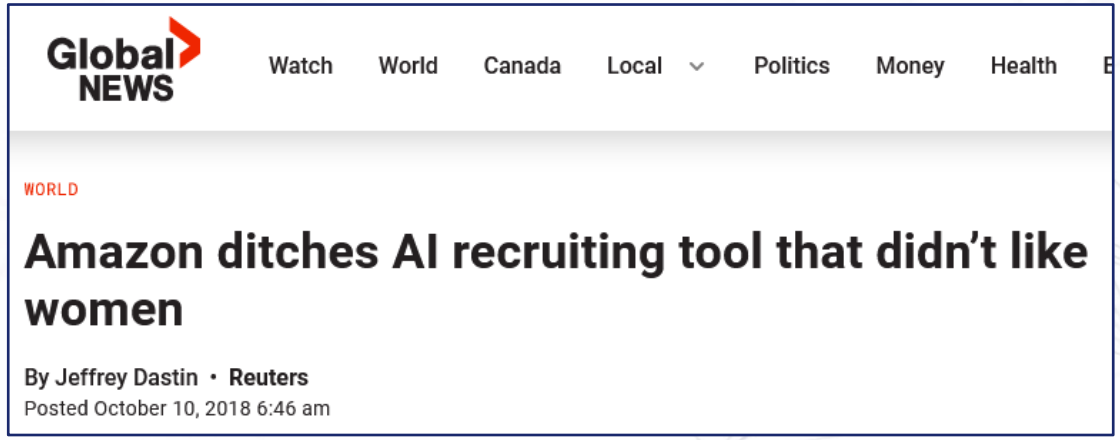
Replica Analytics
AN AETION COMPANY

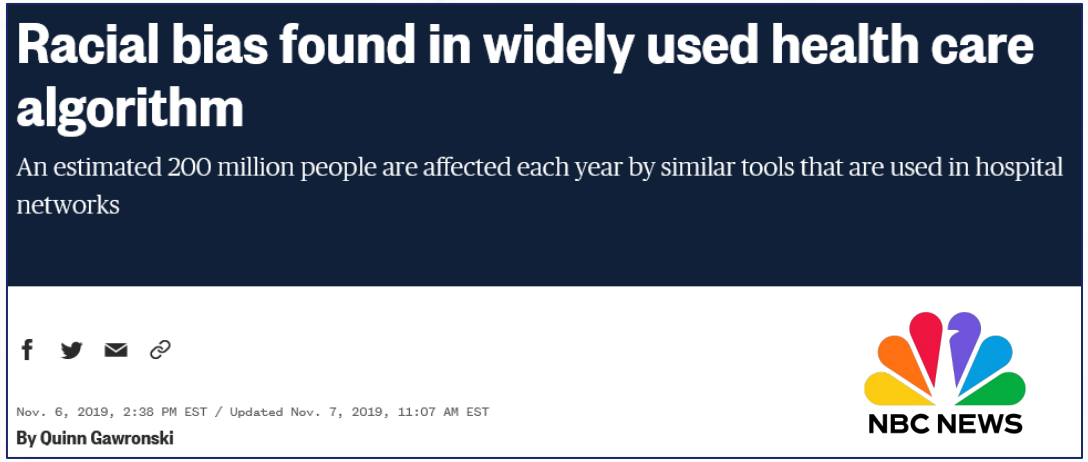
Data Bias Cont'd

- 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 generation 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

Notable Implications

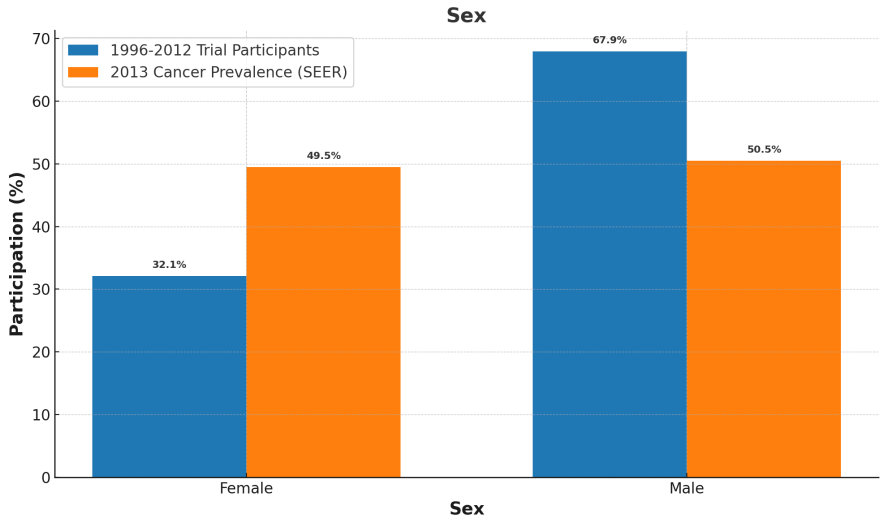


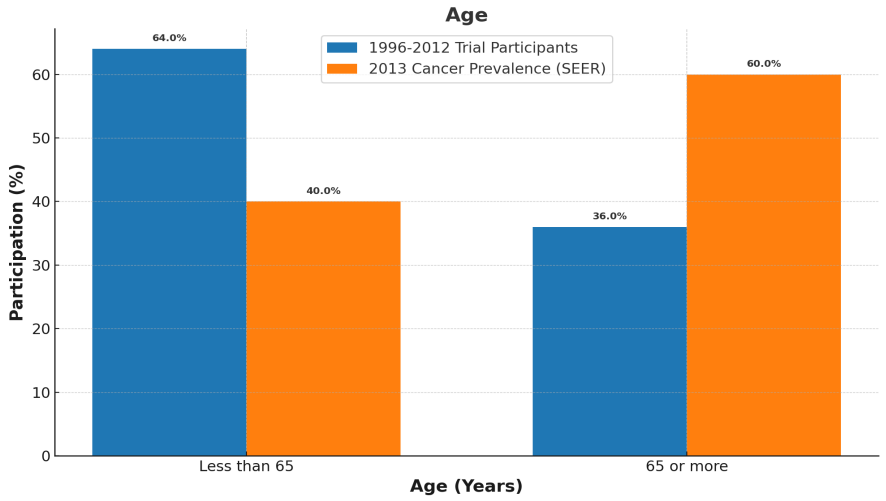


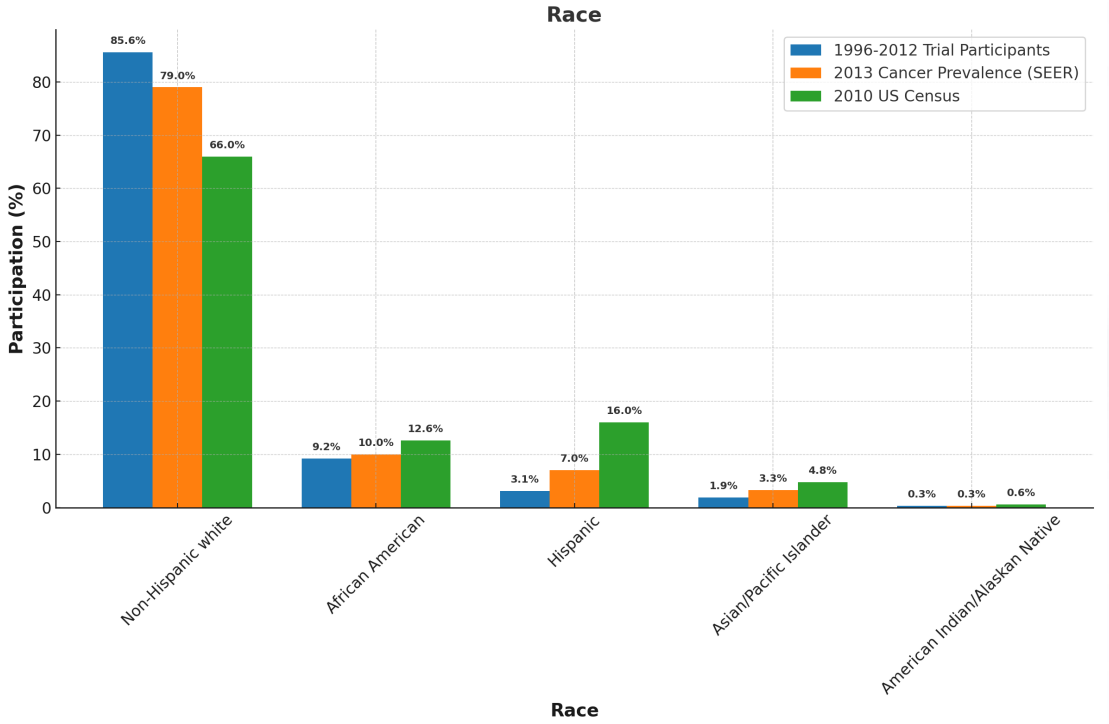
uOttawa

Examples in biomedical research

Participants in all Therapeutic Cancer Trials, 2003-2016 (N = 55,689)







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

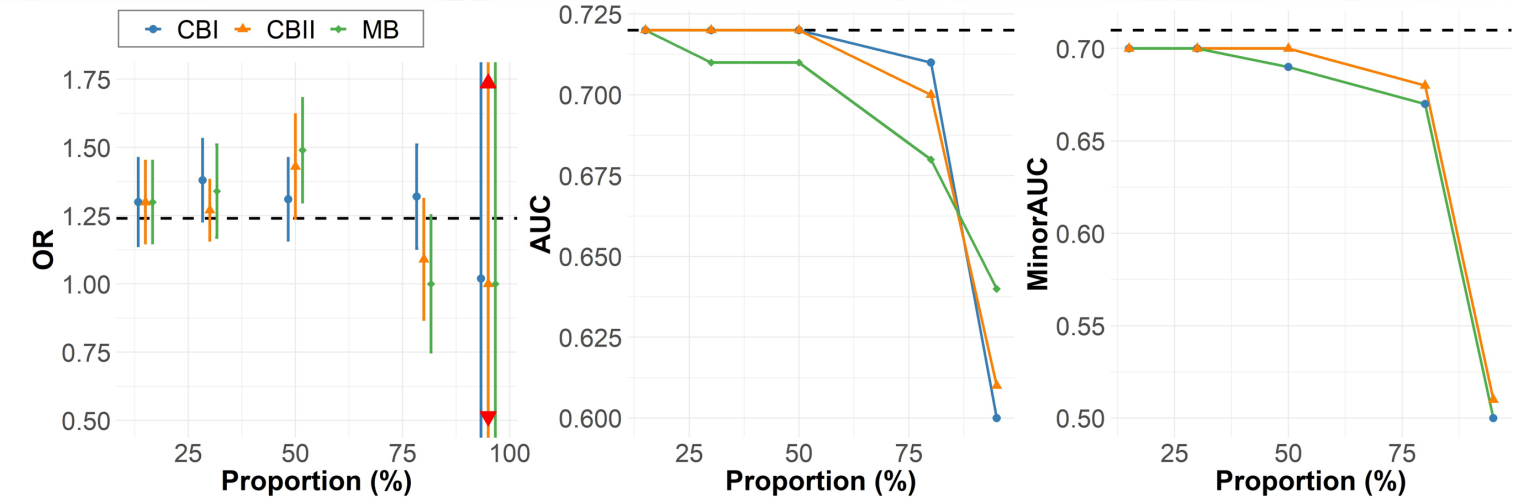
Classifications of biases

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



uOttawa

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 in clinical trials
- From an ethical standpoint, all groups should have a fair participation opportunity

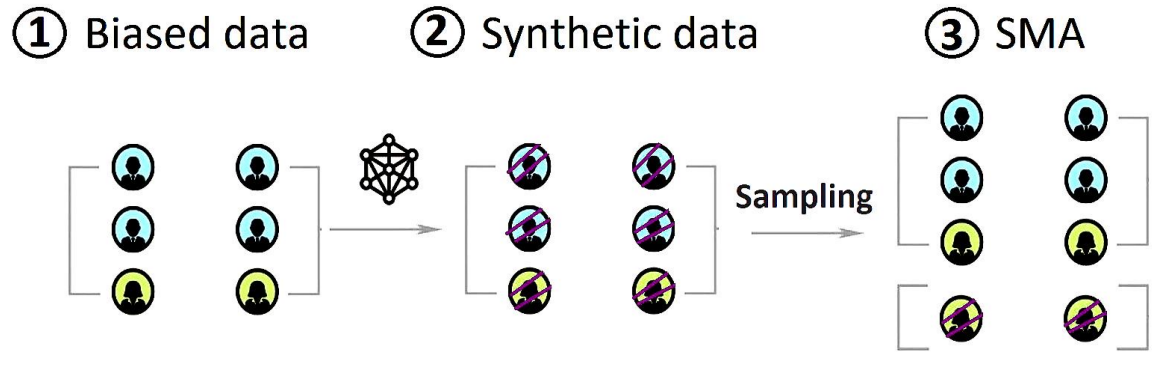
uOttawa

Mitigating Data Bias

Approaches for Mitigating Data Bias

Proposed: Synthetic Minor Augmentation (SMA)

Steps:

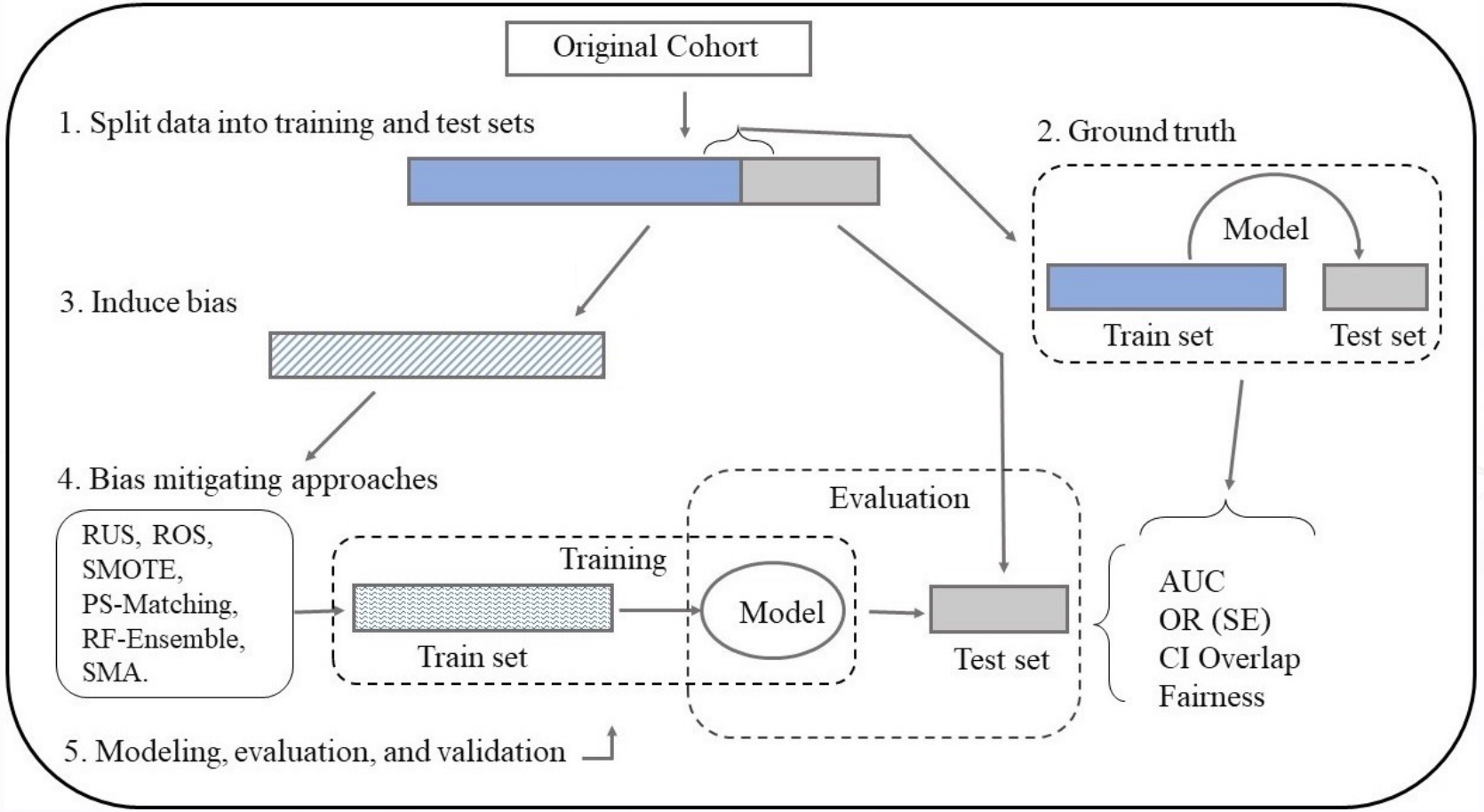


1. Construct a synthetic version of the biased data using sequential synthesis based on gradient boosting decision trees.
2. Sample observations from the bias-inducing (i.e., minor or underrepresented) partition of the generated synthetic dataset.
3. Augment the samples with original biased data to create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

Other approaches

- Random oversampling (ROS)
- Random undersampling (RUS)
- Propensity score (PS) methods (e.g., PS- matching)
- RF ensembles

Model Training & Evaluation



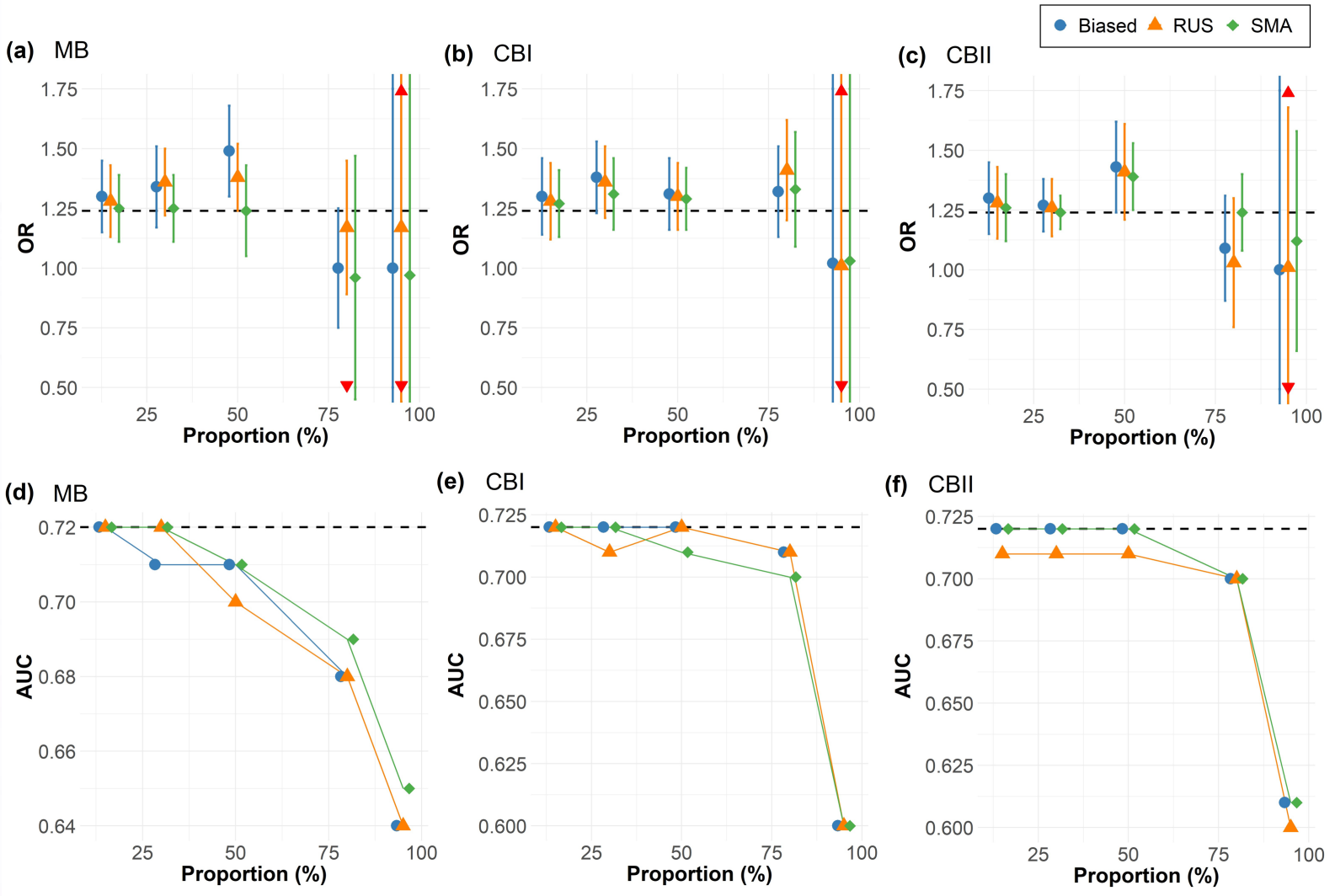
Applications

- We perform two types of analyses:
  - Simulation studies
  - Four real datasets

- The analytical workload assumed is a binary logistic regression model

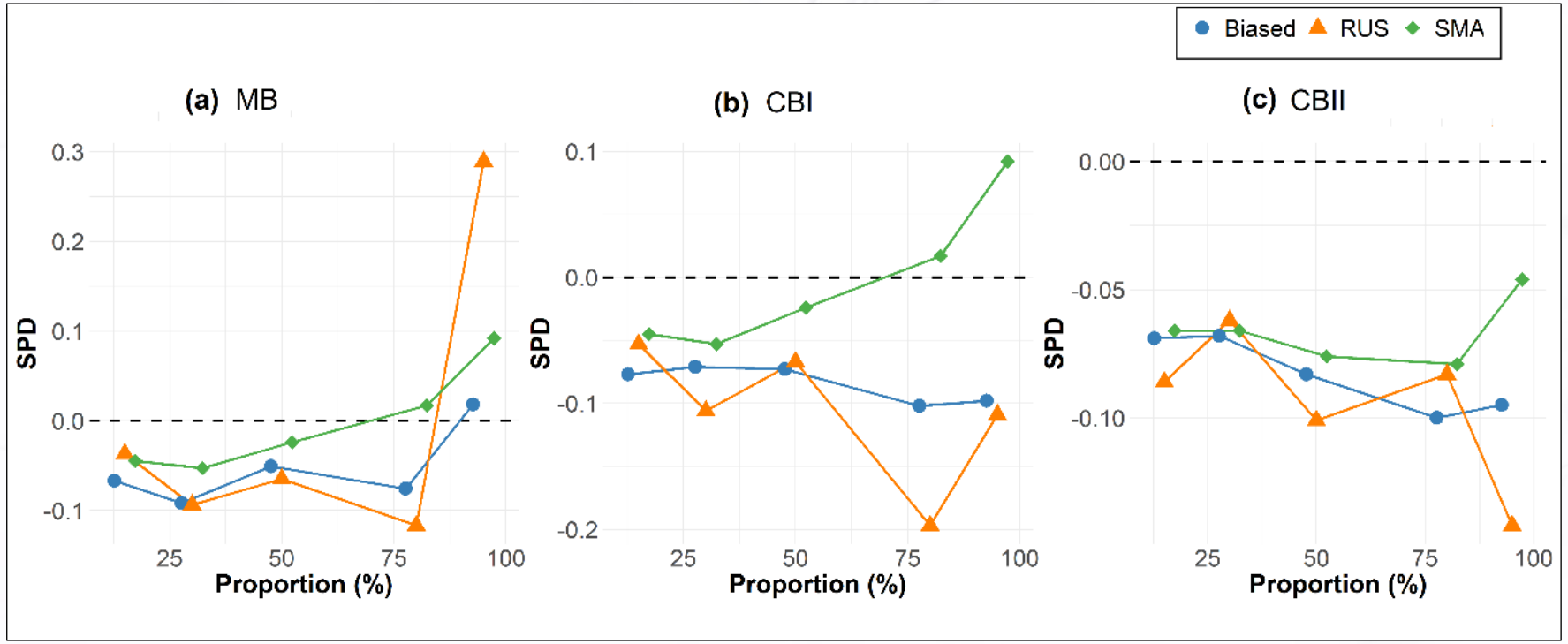
Odds ratio and AUC estimates

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



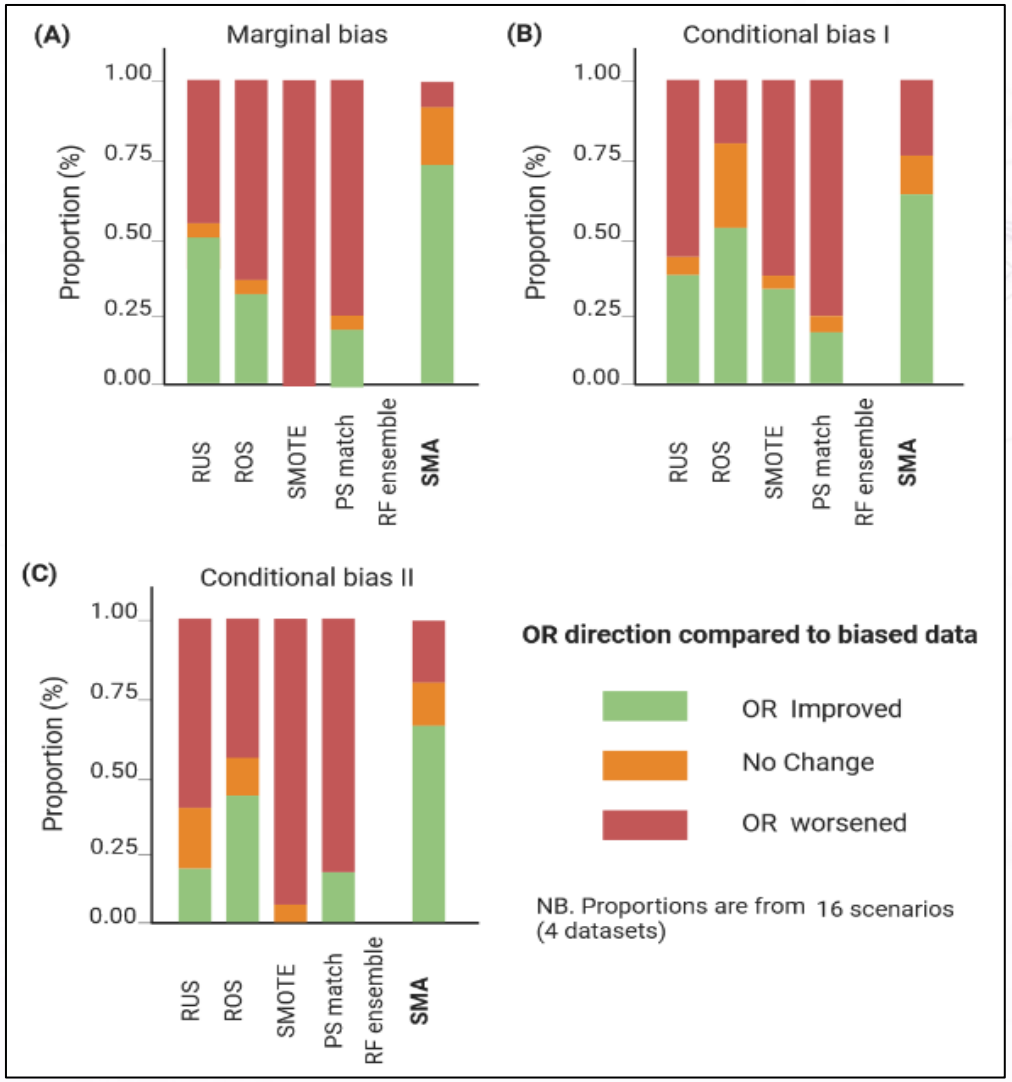
uOttawa

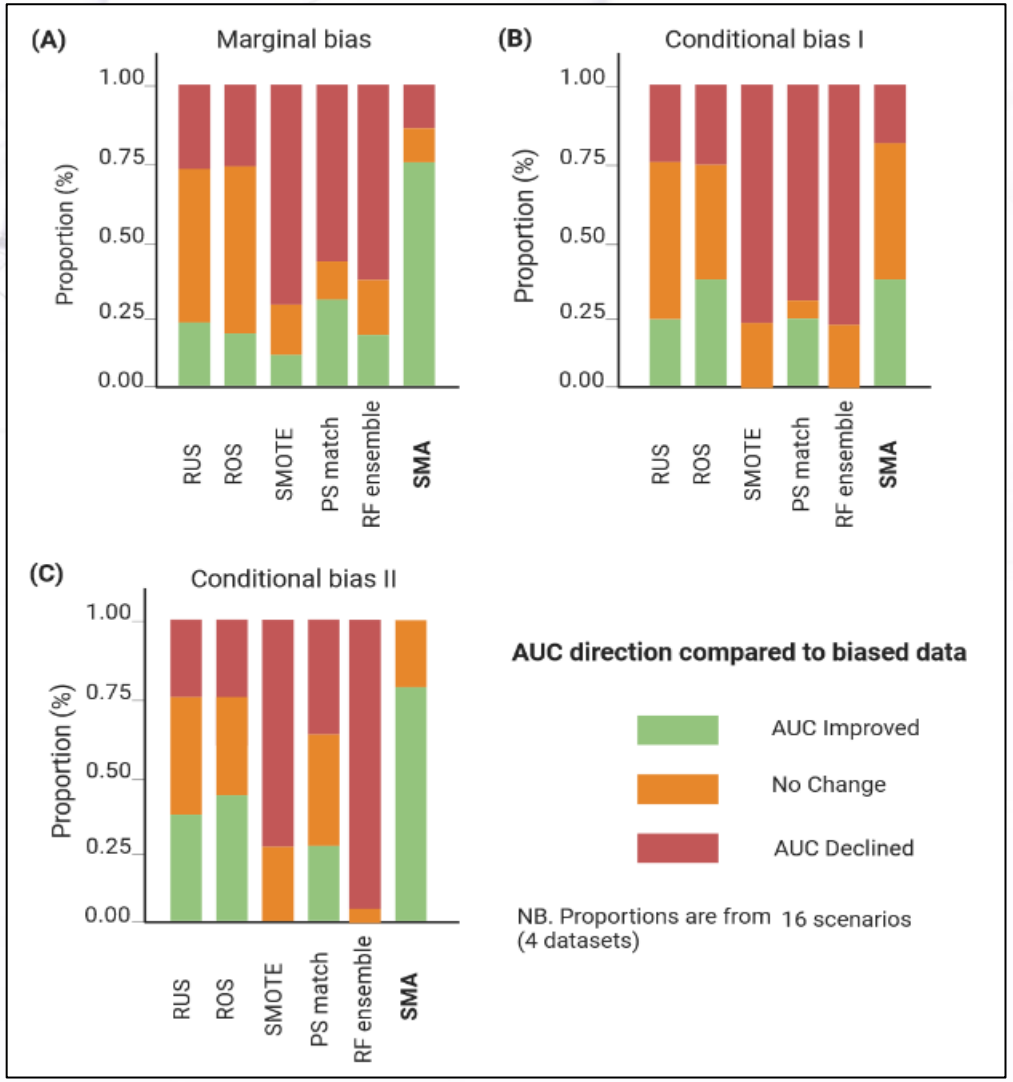
Fairness: Statistical Parity Difference (SPD)



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

Summaries for all datasets: Odds ratio and AUC





uOttawa

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

Questions?