SESSION 2: PERSPECTIVES FROM HEALTH REGULATORS

HIGH-FIDELITY SYNTHETIC DATA APPLICATIONS FOR DATA AUGMENTATION

uOttawa

Replica Analytics
AN AETION COMPANY

Presented by:

Puja Myles,
Director, Clinical Practice Research Datalink (CPRD)
Safety and Surveillance group.
UK Medicines and Healthcare products Regulatory Agency

Medicines & Healthcare products Regulatory Agency

High-fidelity synthetic data applications for data augmentation

Puja Myles

30 November 2023

Scope of presentation

An overview of the Medicines and Healthcare products Regulatory Agency's (MHRA) research into:

- high-fidelity synthetic data generation and evaluation
- high-fidelity synthetic data for data augmentation
- emerging regulatory perspective on data augmentation using synthetic data generation (SDG) approaches

MHRA motivation to develop synthetic data

- Initial funding from Innovate UK in 2018 for proof-of-concept project to develop a high-fidelity synthetic dataset that captures the complex clinical relationships in real data and be used to validate machine learning algorithms
- Key driver: to explore whether synthetic data could support regulation of ML algorithms used in healthcare
- Further funding from NHSX and Regulators Pioneer Fund to extend methodology and test other synthetic data applications, notably for data augmentation

Defining high-fidelity synthetic data

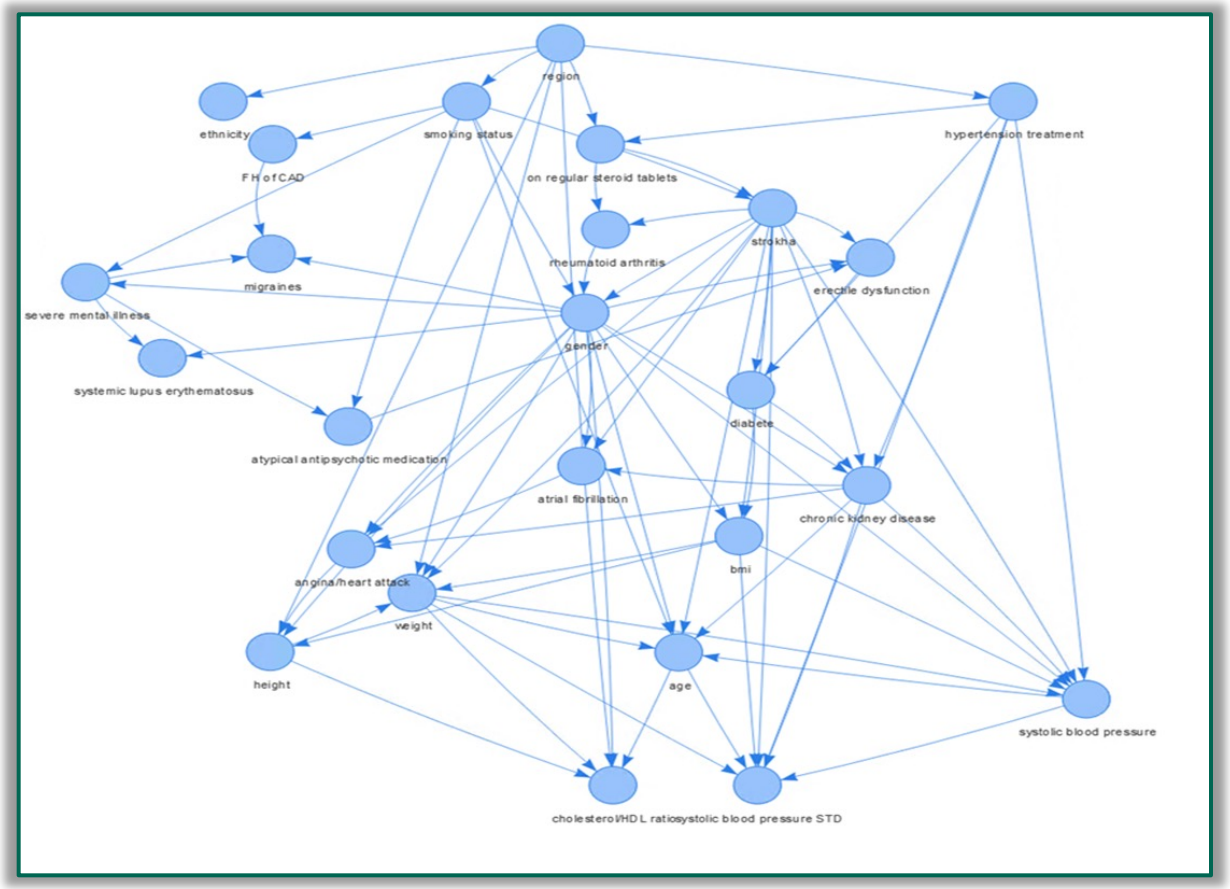
- High-fidelity synthetic data capture both the complex inter-relationships between various data fields and the statistical properties of real data
- In the context of patient data, high-fidelity synthetic data would capture complex clinical relationships and be clinically indistinguishable from ‘real’ patient data

Data augmentation- definition and potential benefits

- A technique used to enhance an existing dataset by generating variations of the available samples using external data sources or synthetic data.
- Aims to increase diversity, volume and quality of the training data
- Could improve a model's ability to generalise and make accurate predictions on unseen instances
- Could help a model learn how to handle different scenarios and improve performance
- Could help address biases resulting from undersampling

MHRA approach to synthetic data generation

- Bayesian network analysis used to discover relationships between multiple data fields in ‘real’ anonymised ground truth (GT) patient data
- These learned relationships/patterns in the GT data are used to generate 100% artificial synthetic (SYN) data



Source: [Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy - Wang - 2021 - Computational Intelligence - Wiley Online Library]

Overview of synthetic data generation approach

Allan Tucker, Zhenchen Wang, Ylenia Rotalinti and Puja Myles

Generating high-fidelity synthetic patient data for assessing machine learning healthcare software.

npj Digit. Med. 3, 147 (2020).
https://doi.org/10.1038/s41746-020-00353-9

Methodological enhancements

- Sample size boosting
- Handling missing data
- Heterogeneous data sources
- Evaluation approach covering utility and privacy metrics

Zhenchen Wang, Puja Myles and Allan Tucker

Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy.

Computational Intelligence. 2021; 37: 819– 851.
https://doi.org/10.1111/coin.12427

Data augmentation for sample size boosting

- Initial experiment using copula modelling to scale up a dataset ten-fold using synthetic data without duplicated records
- Correlational direction between the variables was well preserved in the scaled-up dataset
- Similar predictions obtained from original and augmented datasets
- Copula modelling could not deal with high dimensional datasets
- Bayesian approaches to synthetic data generation now preferred
- Key questions:
  - Is the boosted sample size informative?
  - Are any existing biases preserved or even amplified?

Biased data leads to biased AI algorithms

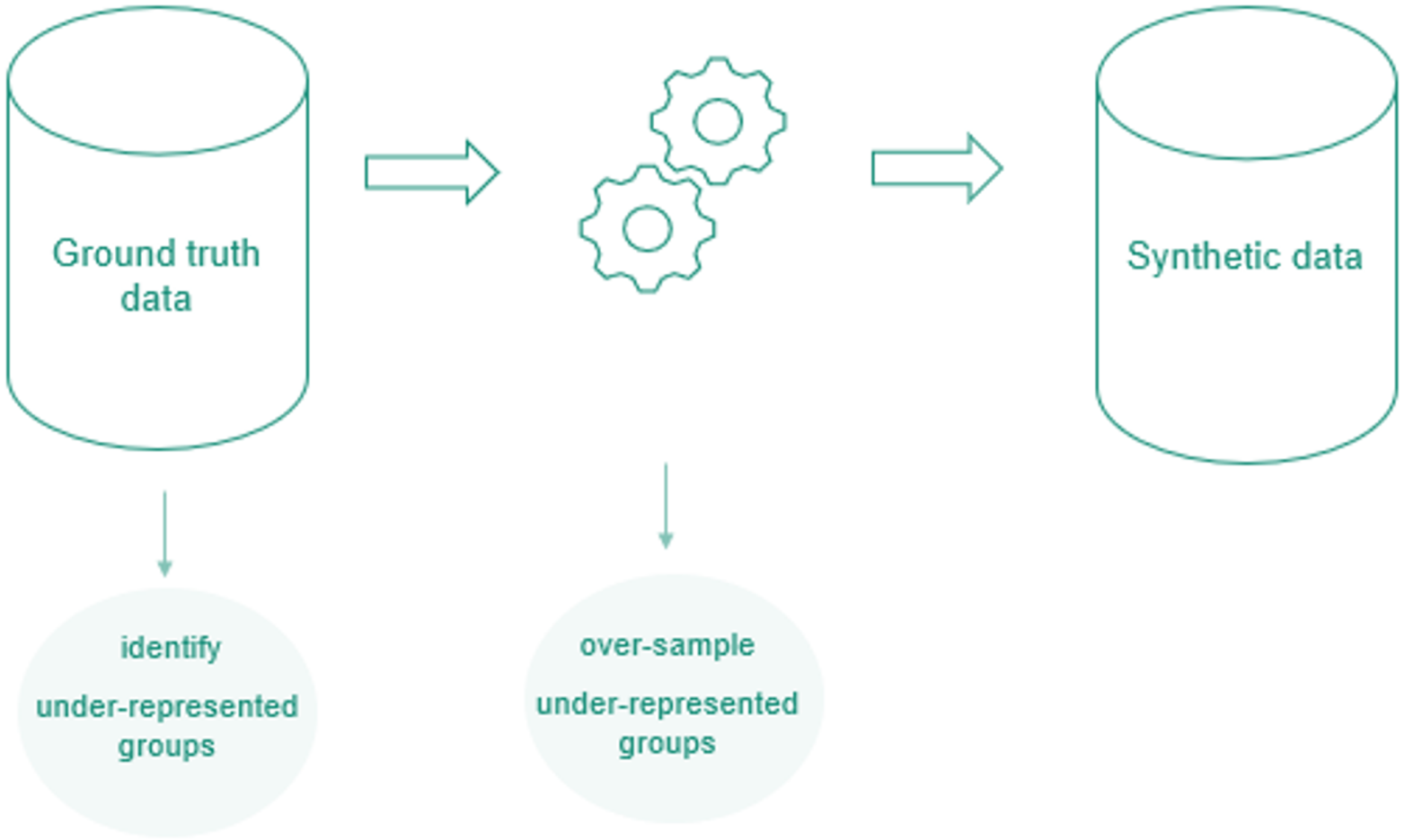
According to media reports in 2018, a large company stopped using an AI recruiting tool because it showed bias against women.

Rashida Richardson, Jason Schultz and Kate Crawford

Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice.

(February 13, 2019). 94 N.Y.U. L. REV. ONLINE 192 (2019)
https://ssrn.com/abstract=3333423

How synthetic data can correct for bias

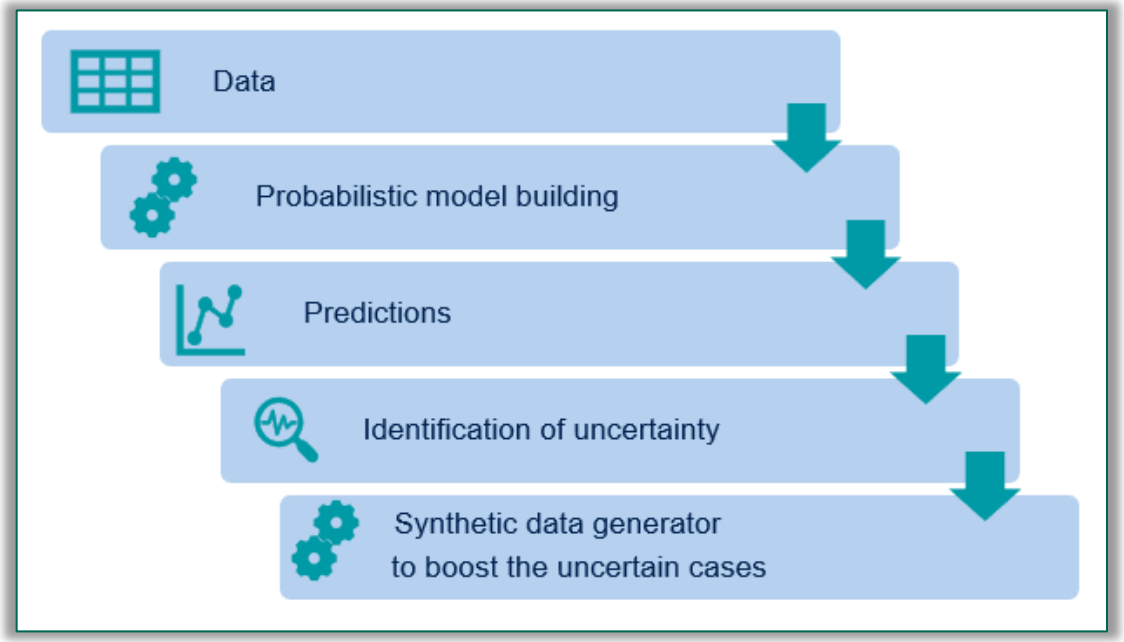


Framework for detecting and correcting for bias

Proceedings of Machine Learning Research 1-13, 2021 Learning and Imbalanced Domains: Theory and Applications

Bayesboost: Identifying and Handling Bias Using Synthetic Data Generators

Barbara Draghi, Zhenchen Wang, Puja Myles and Allan Tucker



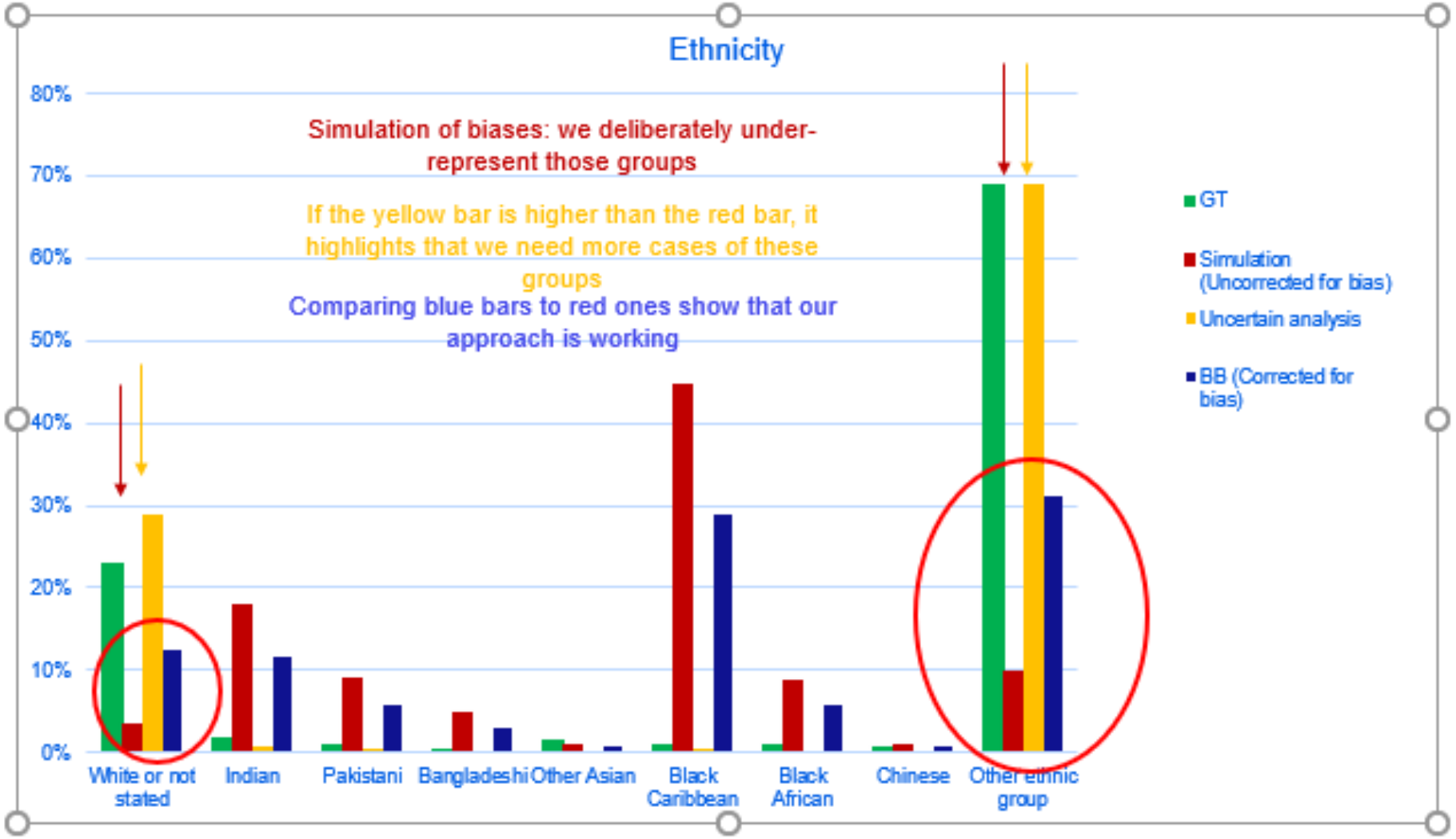
BayesBoost is a novel approach using Bayesian approaches and synthetic data generation methods to detect and correct for known and unknown biases within data

BayesBoost Framework for detecting and correcting bias

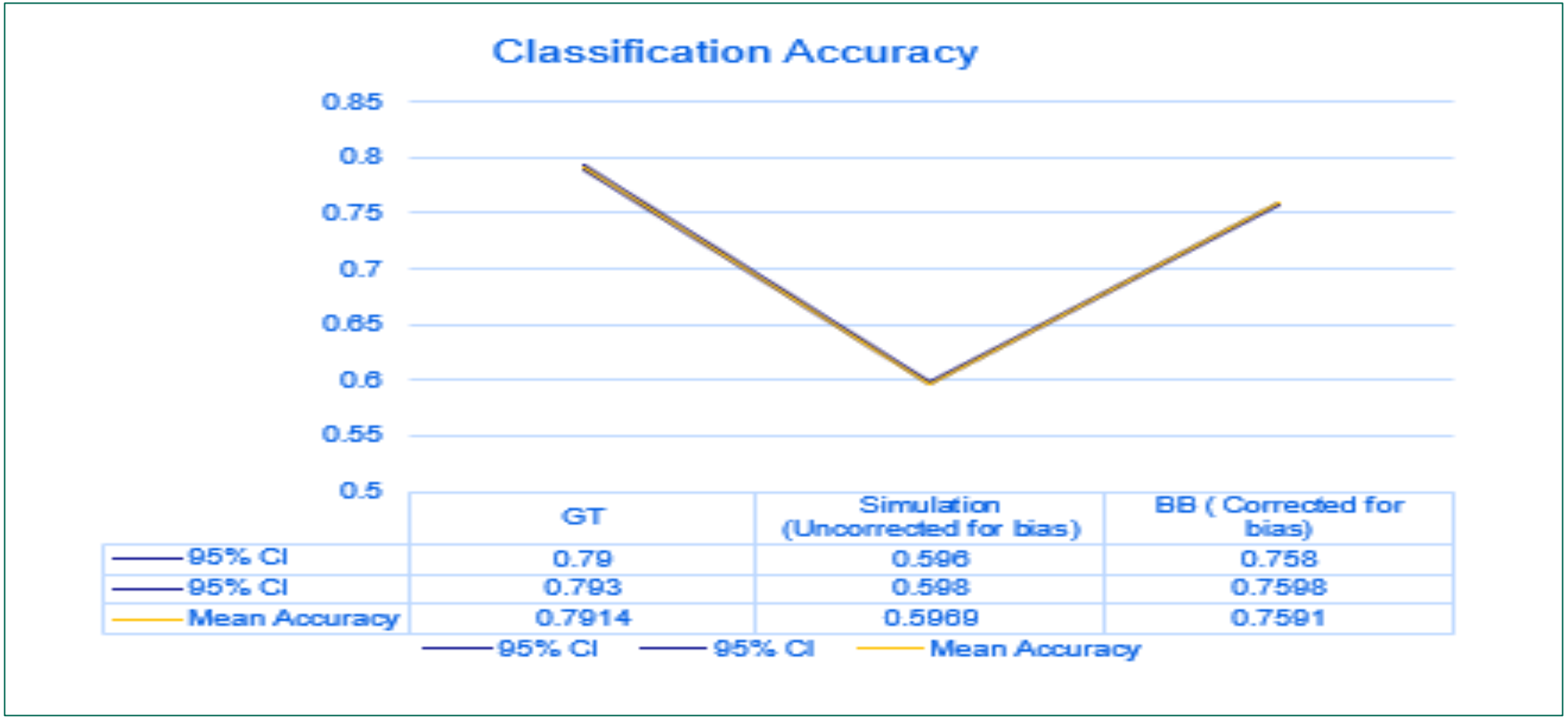
Case study

- Cardiovascular disease risk prediction algorithm
- Cardiovascular risk factors and cardiovascular disease outcomes (combined outcome of stroke and/or heart attacks; type 2 diabetes)

Simulating biases in CVD data to test our detection and correction approach



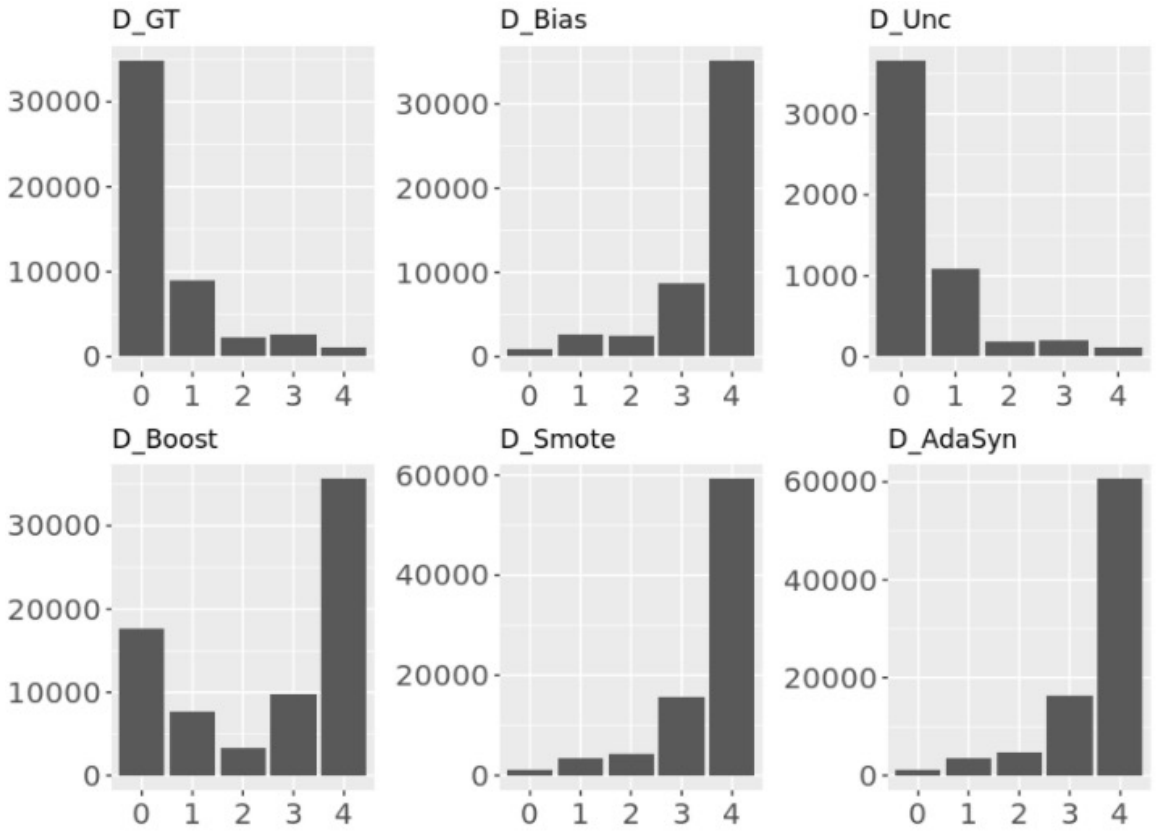
Impact on AI algorithm performance



	GT	Simulation (Uncorrected for bias)	BB (Corrected for bias)
95% CI	0.79	0.596	0.758
95% CI	0.793	0.598	0.7598
Mean Accuracy	0.7914	0.5969	0.7591

How does BayesBoost compare to other bias correction methods?

- Other methods considered- SMOTE and Adaptive Synthetic Sampling (AdaSyn)
- These methods aim to balance the classification problem by balancing the target variable, even if it results in deviating from the correct distribution of the data.
- Data distributions obtained through BayesBoost better resemble the original distribution (see right for smoking data distributions)



Sample size boosting for clinical trials

- Ongoing research funded by the Regulators' Pioneer Fund to use high-fidelity synthetic data generation methods for:
  i. Boost a subset of a previously completed clinical trial dataset (data augmentation) and validate against the full clinical trial dataset
  ii. Generate an entirely artificial control patient arm that could either replace or be an adjunct to a real control arm, with validation against a previously completed controlled clinical trial
  iii. Methodological studies on impact of conditional generation on a particular feature and handling of underrepresentation on more than one feature simultaneously

Regulatory perspective on data augmentation using SDG methods

- AI as a Medical Device needs to be trained on data representative of the entire intended purpose. This must cover all patient populations and claims made by the manufacturer.
- This methodology is valuable because it helps to identify bias and helps developers to understand and explore their data.
- MHRA continues to monitor the progress of these experiments to mitigate bias and understand how this impacts product performance

Medicines & Healthcare products Regulatory Agency

Thank you

Questions?

Puja Myles

puja.myles@mhra.gov.uk

https://www.cprd.com/synthetic-data

Copyright information