Medicines & Healthcare products Regulatory Agency

CPRD

Brunel University London

Detect and Mitigate Bias in Patient Data Using Synthetic Data Generators

2nd ENFIELD Webinar

Bias in Medical AI: Identifying Risks and Ensuring Fairness

23 May 2025, 9:30 - 13:30 (CET)

Barbara Draghi

barbara.draghi@brunel.ac.uk

barbara.draghi@mhra.gov.uk

Agenda

- CPRD and Synthetic Data
- Bias in Real-World data
- How synthetic data can help
- Emulate

CPRD and Synthetic Data

Clinical Practice Research Datalink (CPRD) is a UK government health data research service supporting observational and interventional public health and clinical studies by academics, industry and regulators worldwide.

CPRD operates on a cost-recovery basis

GP practices voluntarily contribute data to CPRD

Researchers must apply and studies be approved to use CPRD data

CPRD

>60 million
Patients for observational studies

>16 million
Patients for trials & clinical studies

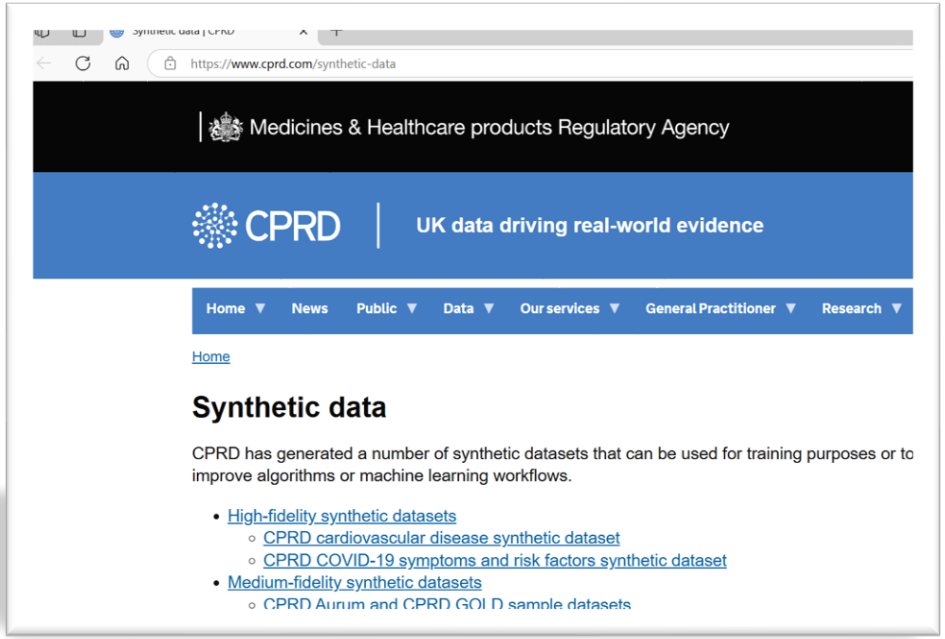
Median 10 years follow-up
25% 20 years follow-up

GP Network
1 in every 4 GP practices in UK

- One of the largest sources of primary care data for public health and research
- Service based on >30 years collecting primary care EHR
- Daily data collection
- ~25% UK population coverage

Synthetic Data

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

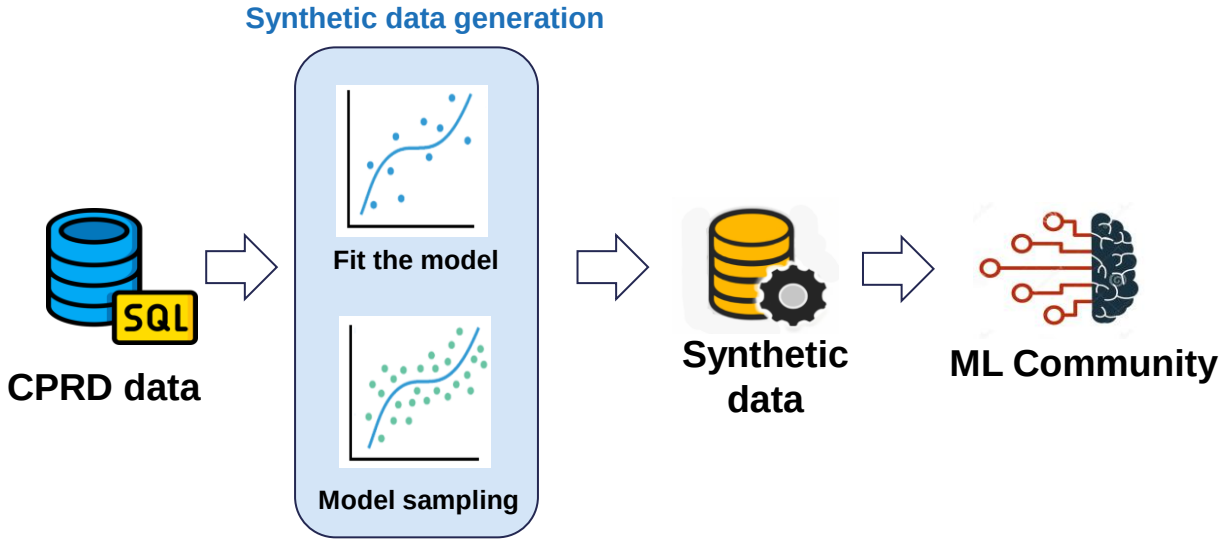


What is Synthetic data?

Synthetic data mimics the characters of real data without containing a one-to-one mapping with real individuals.

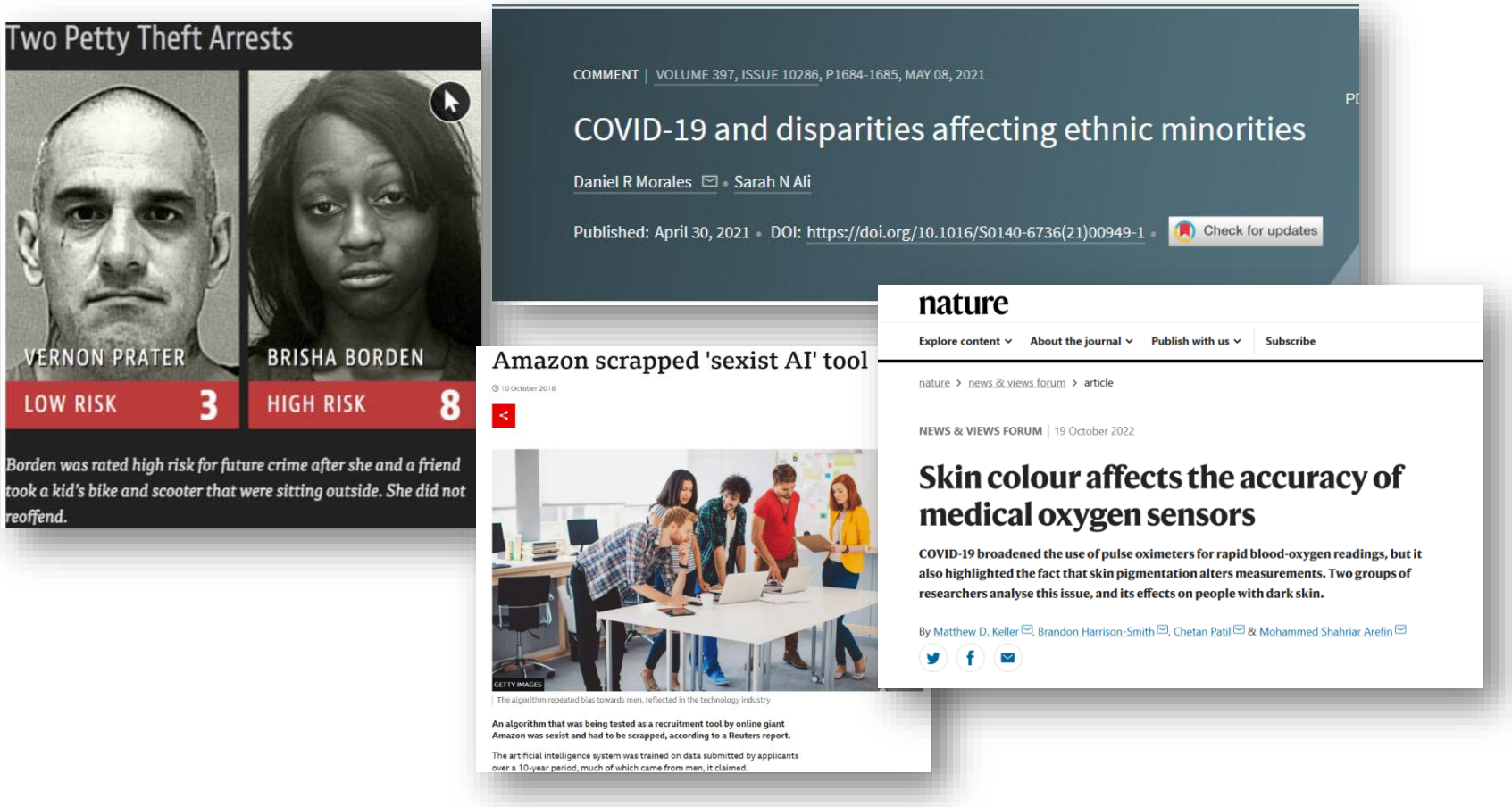
Key advantages

Ease of access, cost and test efficiency, privacy, data augmentation, …

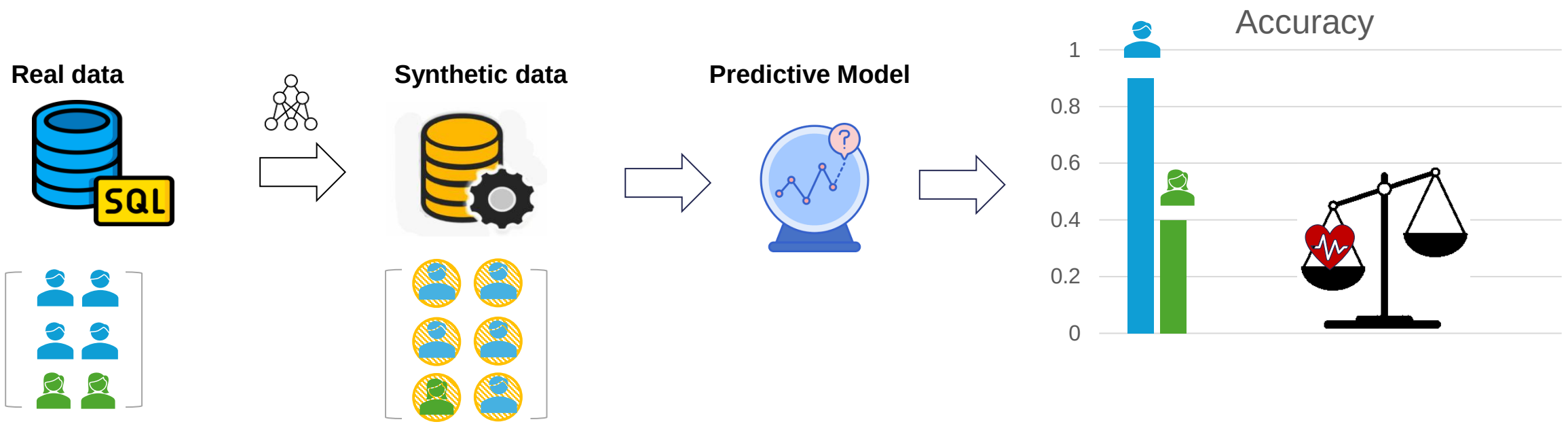


Bias in Real-World data

The problem of bias in real-world data

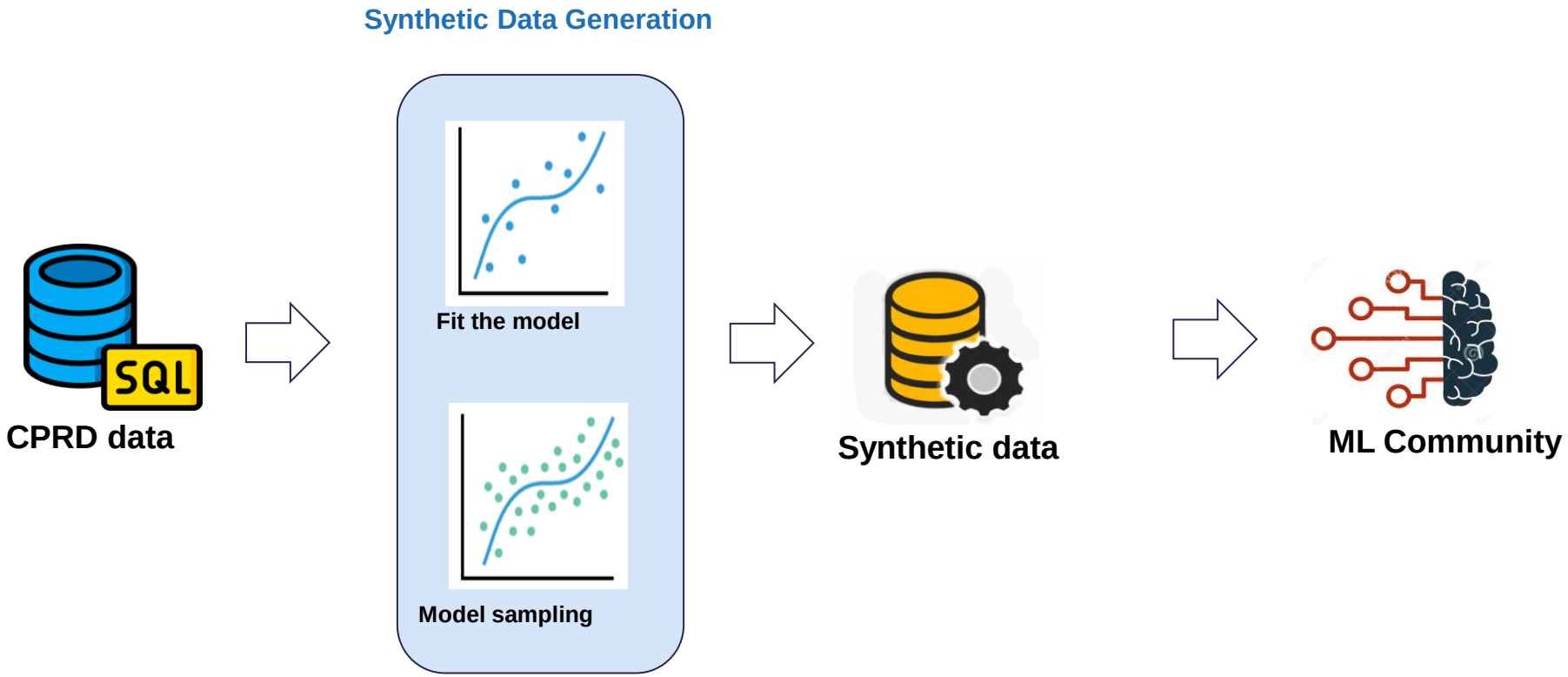


From bias in real-world healthcare to unfair clinical models



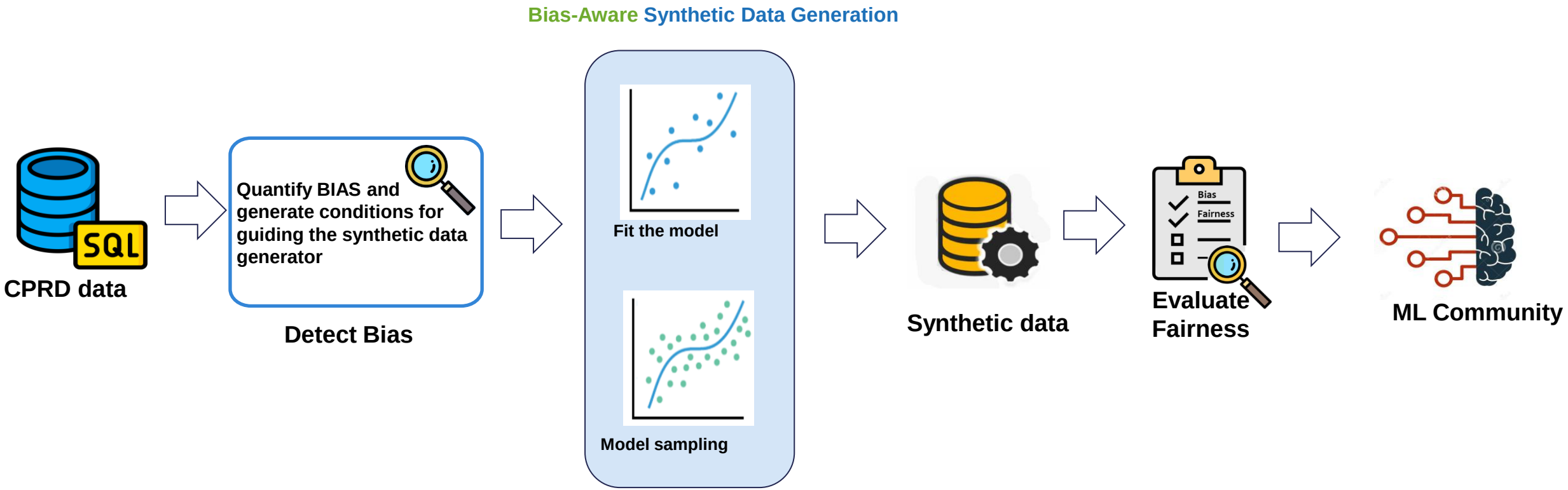
How synthetic data can help

From Synthetic Data Generation…



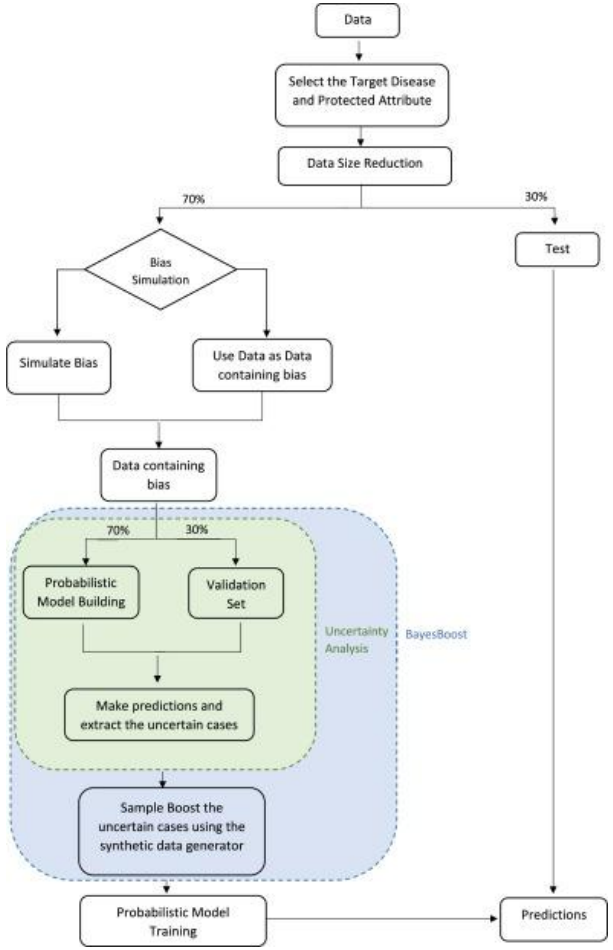
From Synthetic Data Generation…

… to Bias-Aware Synthetic Data Generation

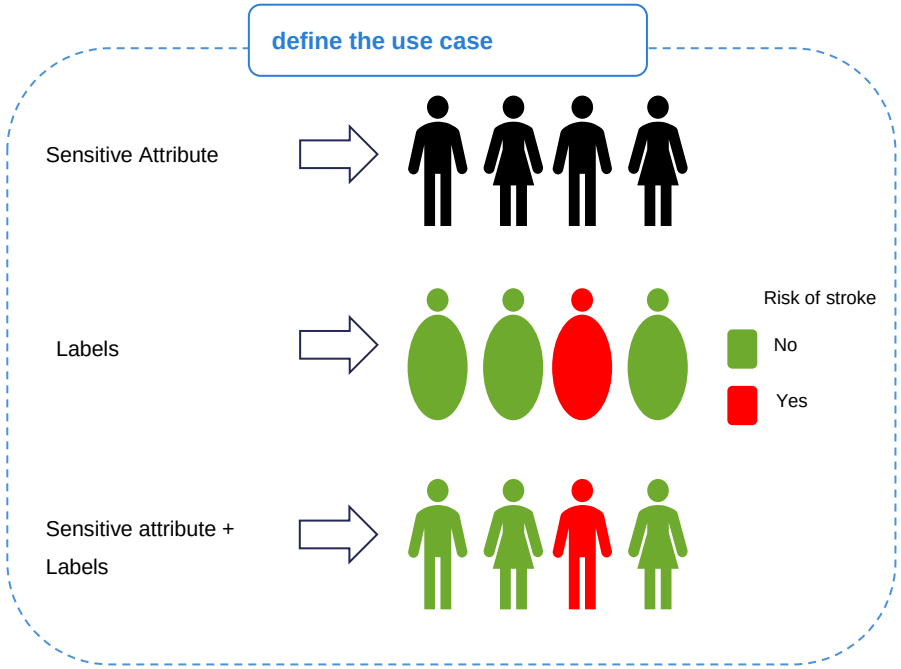


Quantify bias in large (A) and small (B) data

(A) Uncertainty Analysis

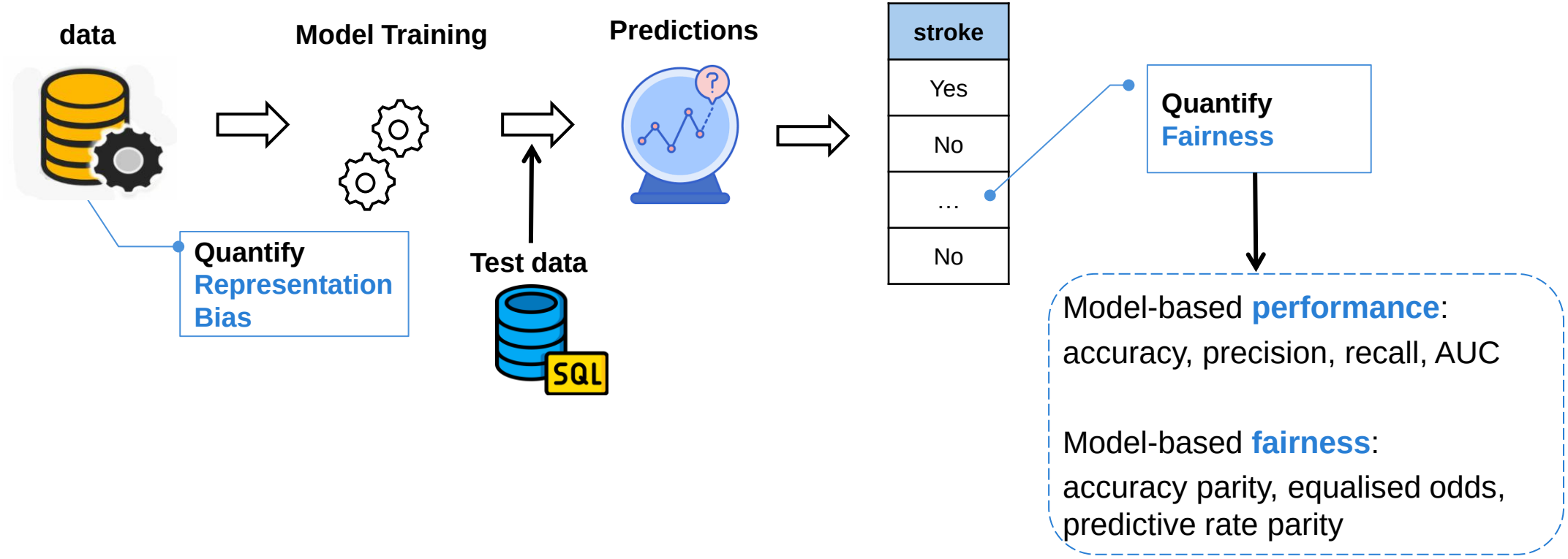


(B) Representation Bias



Coverage metrics help identify whether specific patient groups are under-represented or over-represented in the dataset.

3. Evaluation

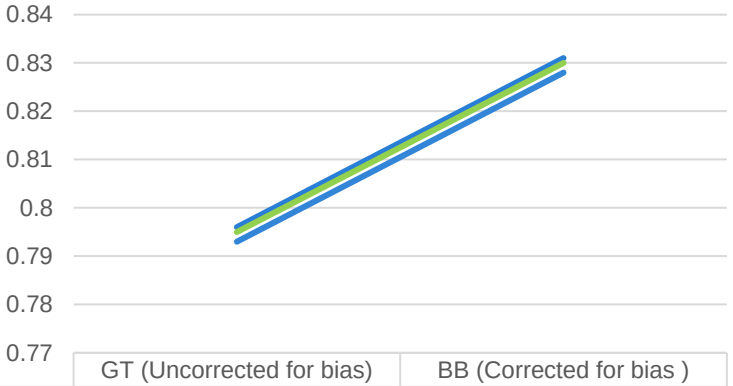


Some key results (1)

- Case study on bias relating to **ethnic** group data. Testing the approach on the **CVD** data.
- Predicting **stroke and heart attacks**.

Some key results (2)

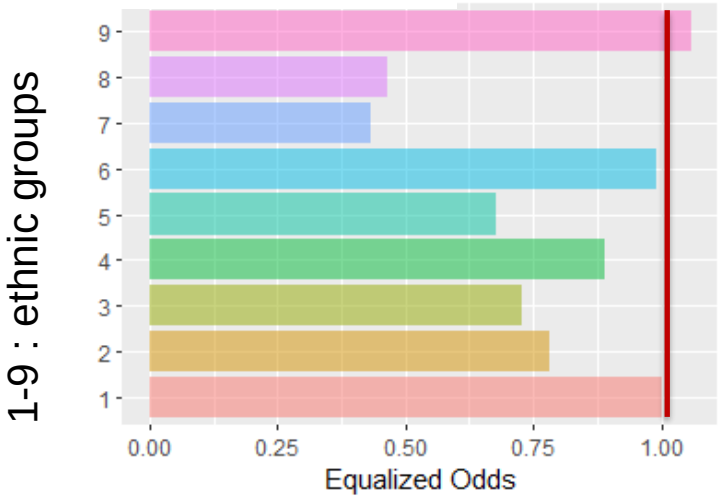
Classification Accuracy



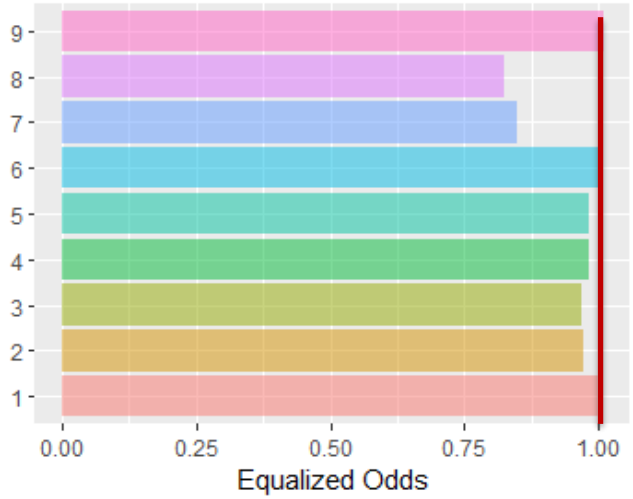
	GT (Uncorrected for bias)	BB (Corrected for bias)
95% Confidence Interval	0.793	0.828
95% Confidence Interval	0.796	0.831
Mean Accuracy	0.795	0.83

Fairness metrics

Original data



Corrected data



Achievements [1]

- Detected biases for one protected attribute
- Improved Classification Accuracy and Fairness for target disease prediction

[1] Draghi B, Wang Z, Myles P, Tucker A. **Identifying and handling data bias within primary healthcare data using synthetic data generators.** *Heliyon*. 2024;10(2):e24164. Published 2024 Jan 10. doi:10.1016/j.heliyon.2024.e24164

EMULATE: a self-service synthetic data generation platform

- Based on the synthetic data research and development undertaken by the MHRA
- Data-driven approach using BN-based approaches
- Web application with ‘no code’ user-friendly interface
- Privacy- utility reports
- Generate high-fidelity synthetic data versions of your own tabular coded datasets
- No storage of your data or data patterns
- You retain the IP and ownership of synthetic data generated from your data

For expressions of interest to try the prototype, contact enquiries@cprd.com with "Emulate EOI" in the subject header

Medicines & Healthcare products Regulatory Agency

Brunel University London

CPRD

Thank you!

barbara.draghi@brunel.ac.uk / barbara.draghi@mhra.gov.uk

www.linkedin.com/in/barbaradraghi

enquiries@cprd.com - "Emulate EOI" in the subject header