Medicines & Healthcare products Regulatory Agency

Synthetic data applications

Puja Myles

16 June 2022

Synthetic data- more than a privacy enhancing technology?

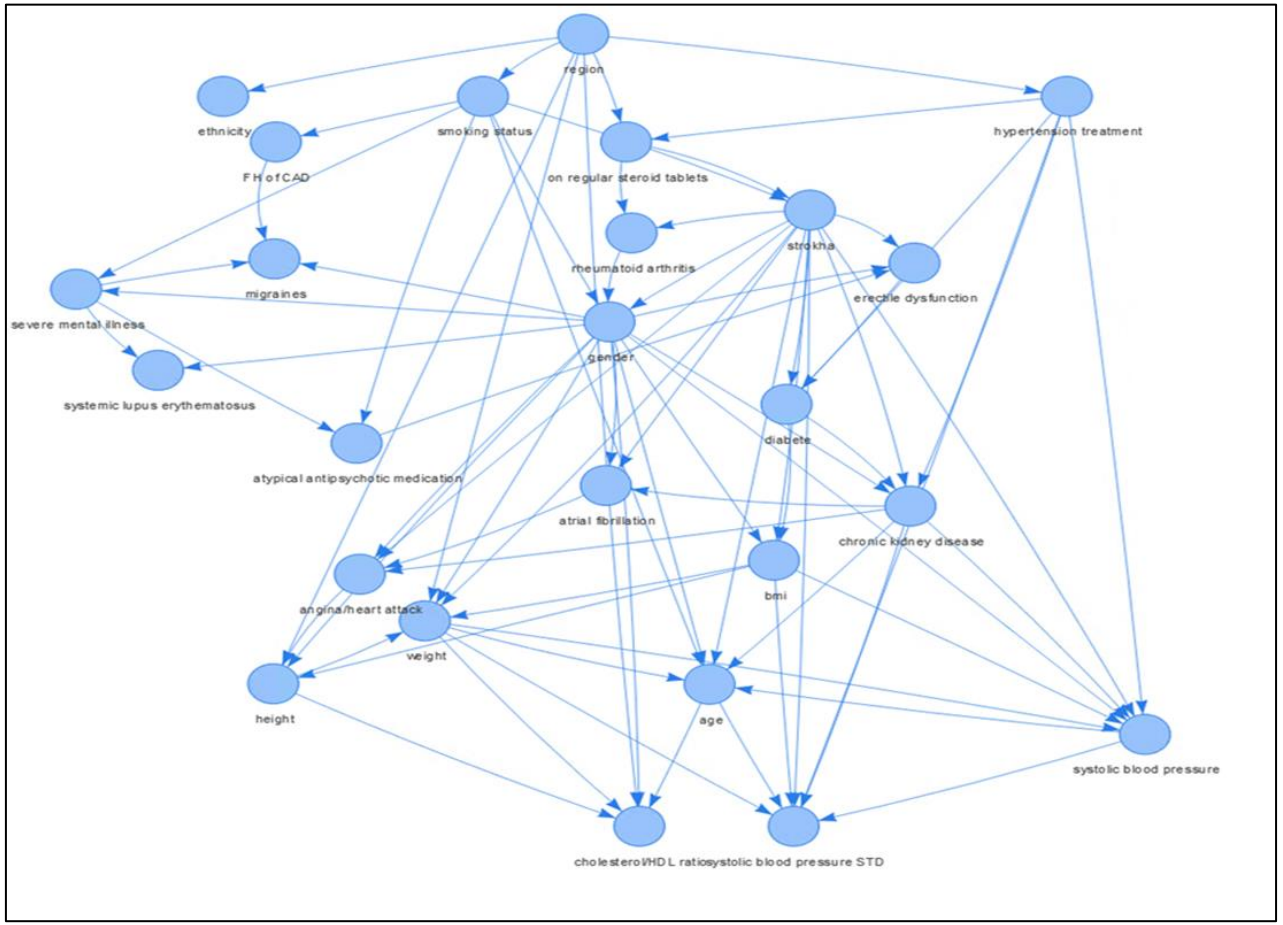
- Advances in synthetic data generation methods have opened up other potential applications in addition to the use of synthetic data as a privacy enhancing technology (PET)
- The potential applications will differ based on the utility or fidelity of the synthetic dataset
- In the context of patient health care data, a high-fidelity synthetic dataset would be able to capture complex clinical relationships and be clinically indistinguishable from real patient data
- Low to medium fidelity data → useful for understanding data structure, developing programming code, analytics tools and machine learning workflows for use with 'real' patient data on which the synthetic data is based, teaching data management/wrangling etc.
- High-fidelity synthetic data opens up more applications…

High-fidelity synthetic data: application 1

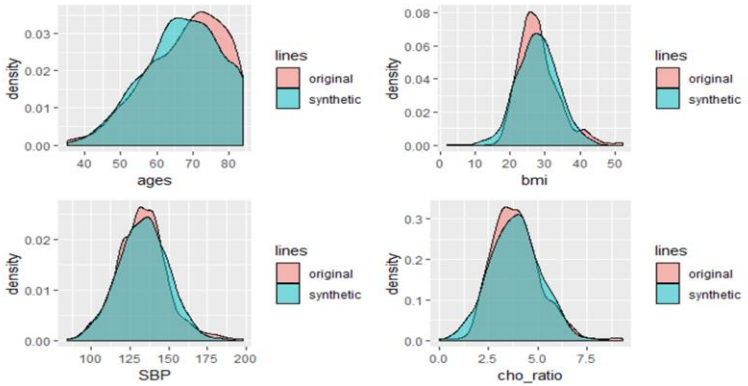
How can we assess machine learning healthcare algorithms if we don’t have access to suitable datasets for validation purposes?

The MHRA’s work in this area

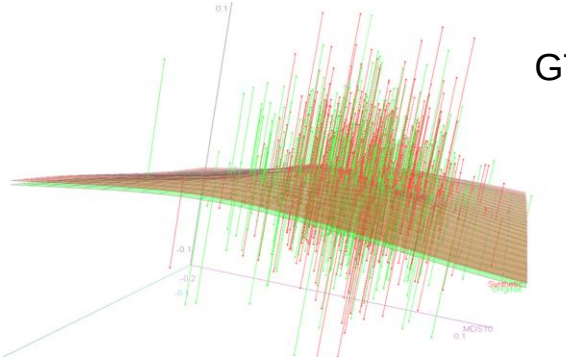
- Bayesian network analysis used to discover relationships between multiple data fields in ‘real’ anonymised ground truth (GT) patient data
- These learned relationships/patterns in the GT data are used to generate 100% artificial synthetic (SYN) data



How do we know that the synthetic data is any good?

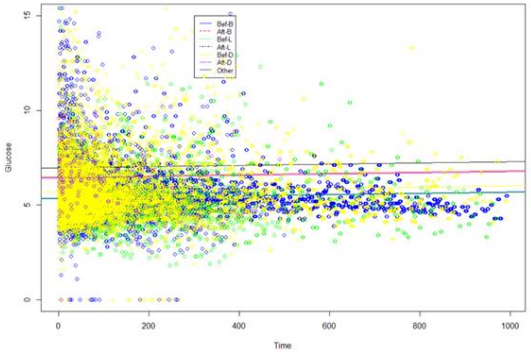


Look for overlap between distributions in real and synthetic data- one variable at a time

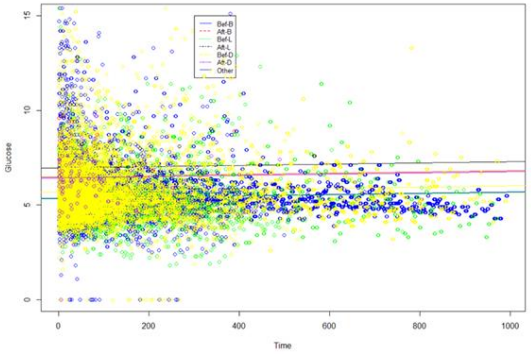


GT: Green; SYN: Red

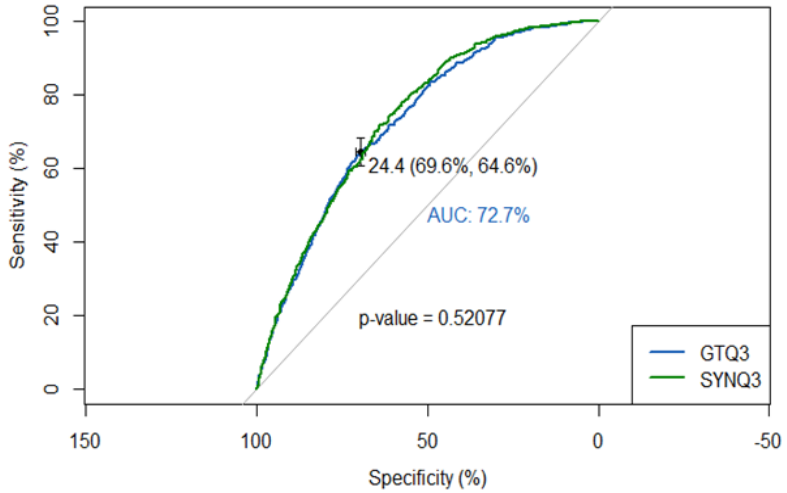
3D mapping of SYN & GT data points



Ground truth data



Synthetic data



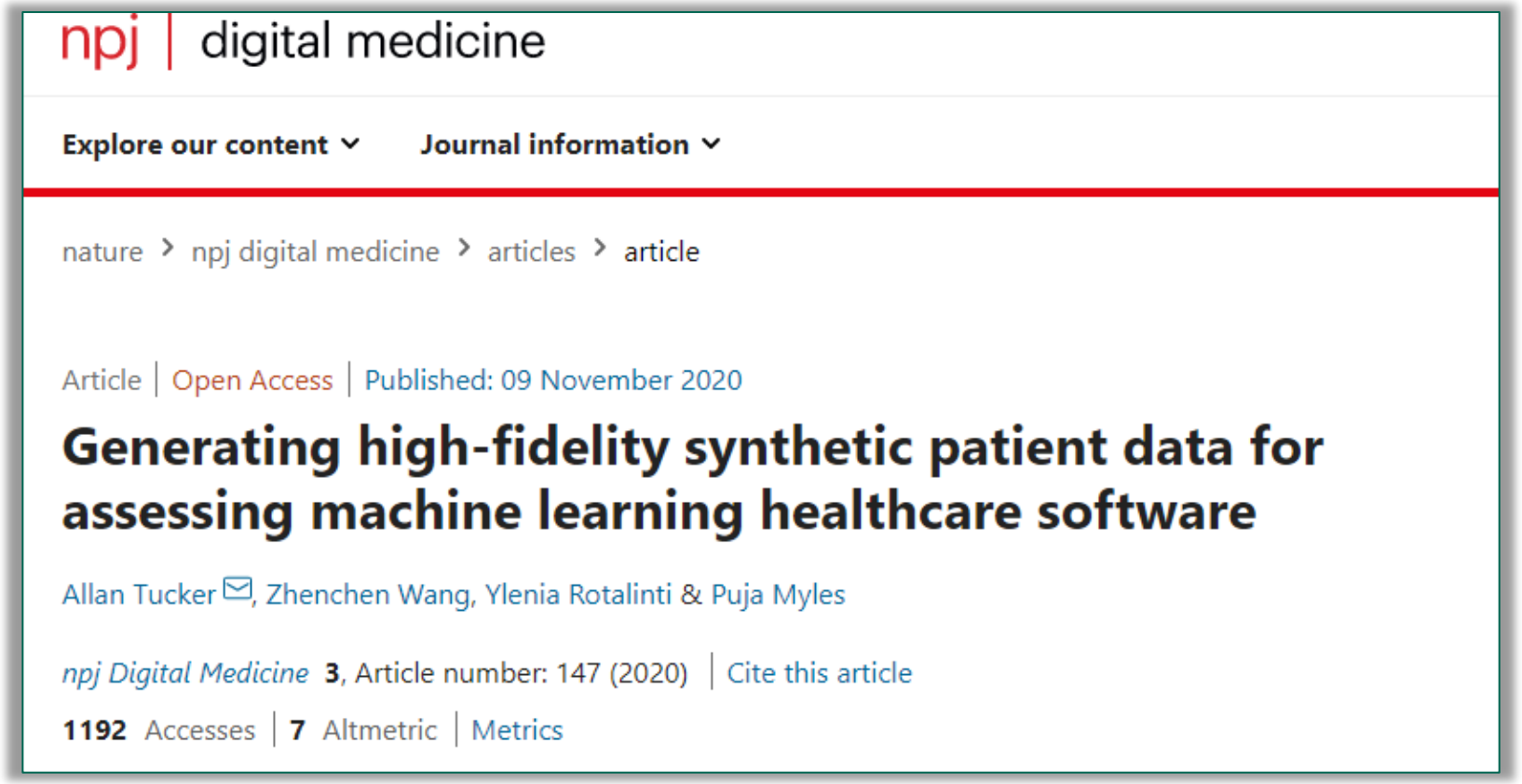
Comparison of ML algorithm performance in GT & SYN data

Clinical validation

- 2 independent medical assessors
- Review of randomly selected sample (n=100) of equal number of synthetic and real patient records
- Clinical experts able to classify real patient records correctly with high degree of accuracy but tended to misclassify synthetic records as being real

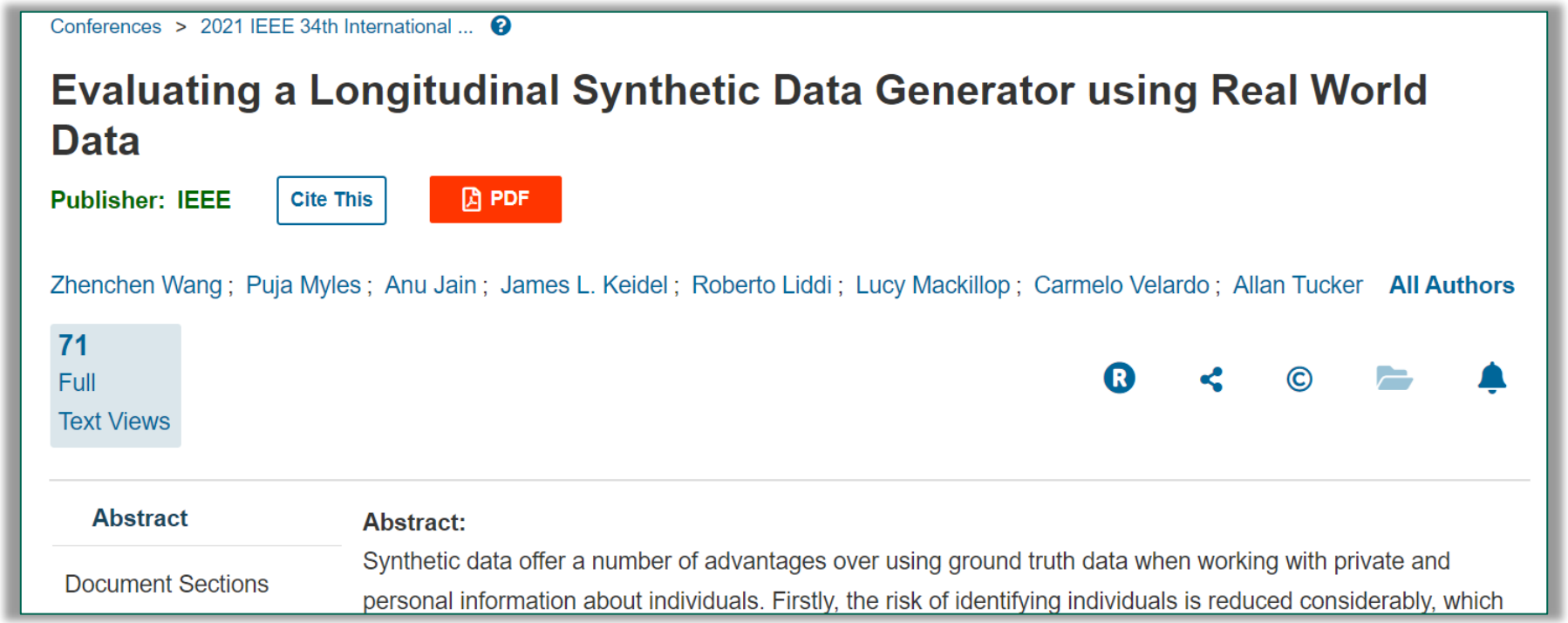
Type of records evaluated	**Expert 1 results**		**Expert 2 results**	
	Correctly classified	Incorrectly classified	Correctly classified	Incorrectly classified
Total records	15/24 (62.5%)	9/24 (37.5%)	17/34 (50%)	17/34 (50%)
Synthetic records	2/10 (20.0%)	8/10 (80.0%)	1/15 (6.7%)	14/15 (93.3%)
Real records	13/14 (92.9%)	1/14 (7.1%)	16/19 (84.2%)	3/19 (15.8%)

Validation and benchmarking



https://www.nature.com/articles/s41746-020-00353-9

Synthetic data for training and validation



Conferences > 2021 IEEE 34th International ...

Evaluating a Longitudinal Synthetic Data Generator using Real World Data

Publisher: IEEE Cite This PDF

Zhenchen Wang ; Puja Myles ; Anu Jain ; James L. Keidel ; Roberto Liddi ; Lucy Mackillop ; Carmelo Velardo ; Allan Tucker **All Authors**

71
Full
Text Views

Abstract

Document Sections

Abstract:
Synthetic data offer a number of advantages over using ground truth data when working with private and personal information about individuals. Firstly, the risk of identifying individuals is reduced considerably, which

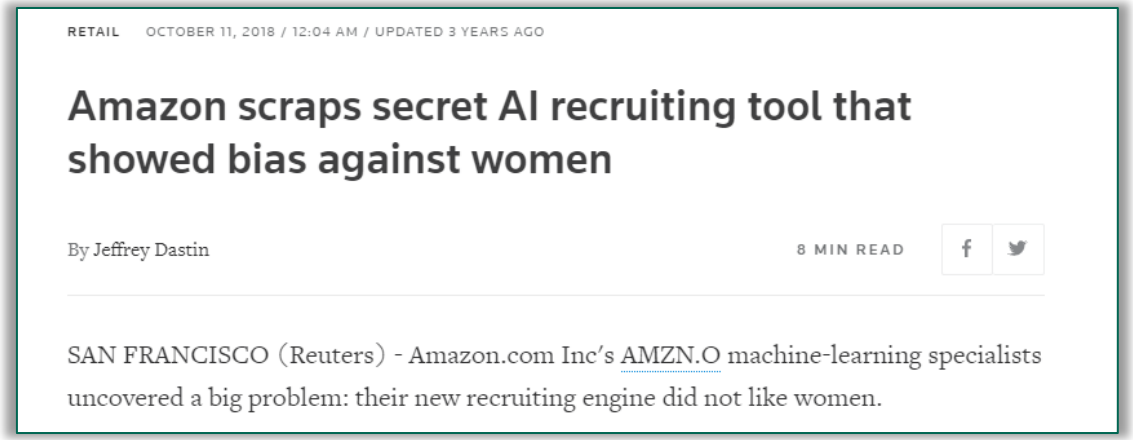
High-fidelity synthetic data can be used for both training and validation of ML algorithms

High-fidelity synthetic data: application 2

- Sample size boosting
- Is this informative?

https://doi.org/10.1111/coin.12427

Biased data leads to biased AI algorithms

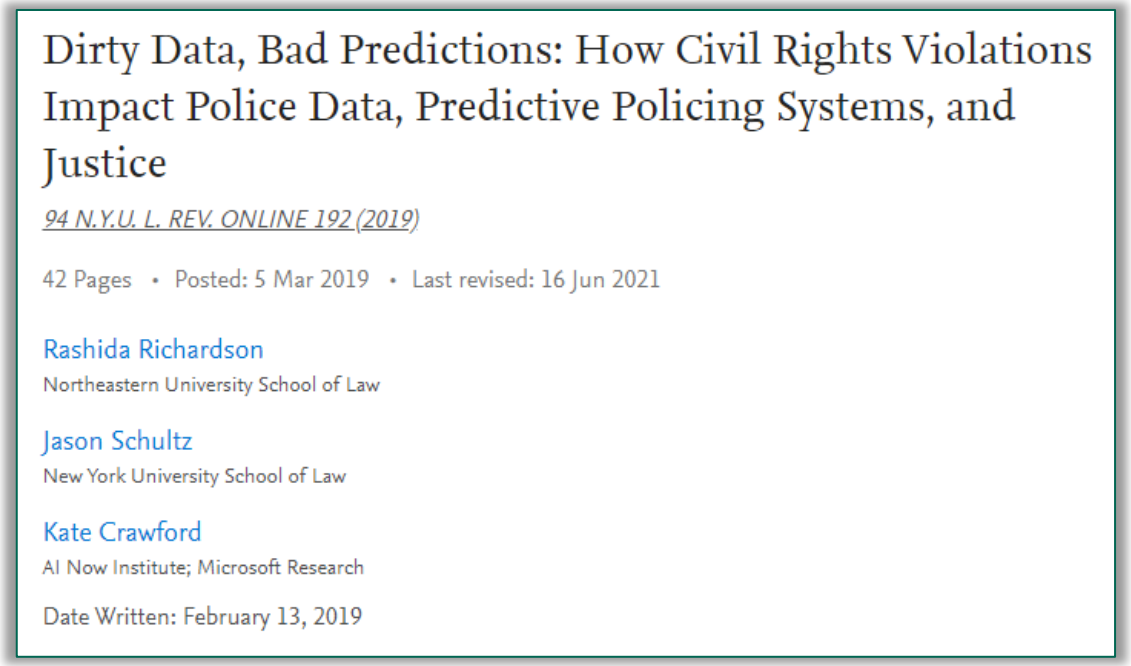


RETAIL OCTOBER 11, 2018 / 12:04 AM / UPDATED 3 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.



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

94 N.Y.U. L. REV. ONLINE 192 (2019)

42 Pages • Posted: 5 Mar 2019 • Last revised: 16 Jun 2021

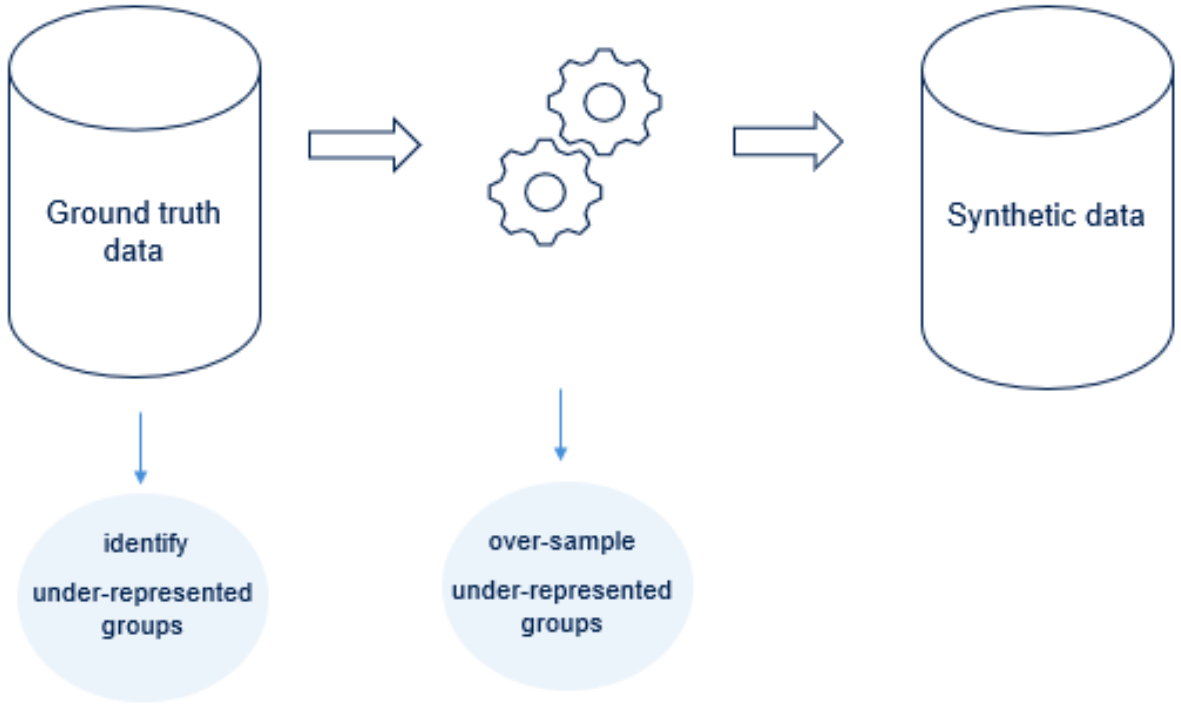
Rashida Richardson
Northeastern University School of Law

Jason Schultz
New York University School of Law

Kate Crawford
AI Now Institute; Microsoft Research

Date Written: February 13, 2019

High-fidelity synthetic data application 3: correcting biases



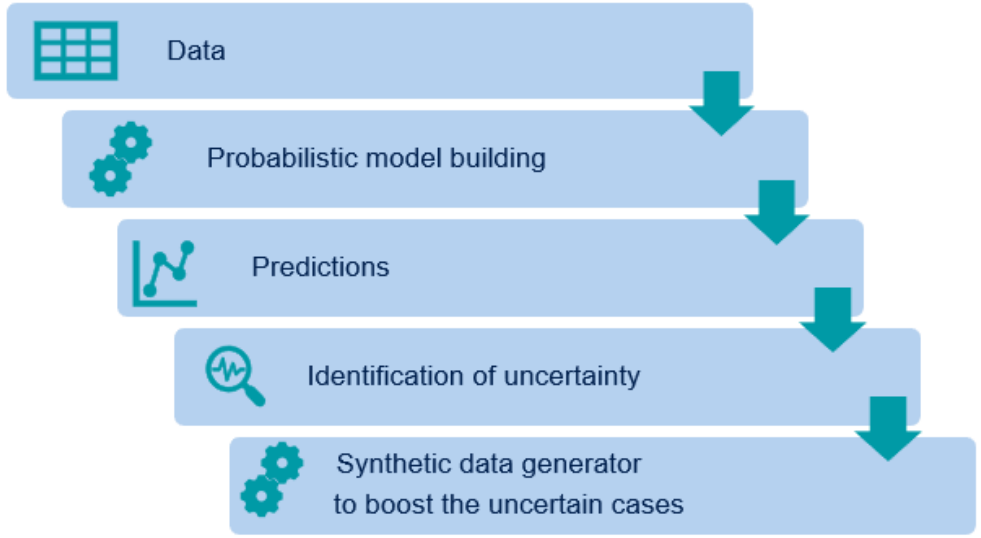
Framework for detecting and correcting for bias

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

BayesBoost: Identifying and Handling Bias Using Synthetic Data Generators

Barbara Draghi BARBARA.DRAGHI@MHRA.GOV.UK
Zhenchen Wang ZHENCHEN.WANG@MHRA.GOV.UK
Puja Myles PUJA.MYLES@MHRA.GOV.UK
Medicine and Healthcare products Regulatory Agency, London UK

Allan Tucker ALLAN.TUCKER@BRUNEL.AC.UK
Brunel University London, UK



BayesBoost Framework for detection and correction of biases

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

Other potential high-fidelity synthetic data applications

- Generation of synthetic patient cohorts to support in silico trials to simulate intervention effects in sub-groups not typically included in RCTs
- External control groups for clinical testing or benchmarking data from single-arm trials
- Causal effect estimation by generating synthetic factual and counterfactual outcomes

Synthetic data and regulation of AI medical devices

RF QUARTERLY

Synthetic data and the innovation, assessment, and regulation of AI medical devices

01 June 2021 | By Puja Myles, PhD, MPH; Johan Ordish, MA; and Richard Branson, MSc, MA

Synthetic data are artificial data that mimic the properties of and relationships in real data. They show promise for facilitating data access, validation, and benchmarking, addressing missing data and under-sampling, sample boosting, and the creation of control arms in clinical trials. The UK Medicines and Healthcare products Regulatory Agency (MHRA) is using its current research into the development of high-fidelity synthetic data to develop its regulatory position on ar...

Synthetic datasets available from CPRD

- High-fidelity synthetic datasets
  - Cardiovascular disease risk
  - COVID-19 symptoms and risk factors

  (can be used for ML/AI research applications)
- Medium-fidelity dataset based on primary care records
  - CPRD Aurum sample dataset

  (can be used to understand the structure and utility of the anonymised primary care data, as a data management teaching/training resource, to develop/validate/test analytics tools for use with real data, develop machine learning workflows that can be applied to real patient data)

More information: https://cprd.com/synthetic-data