AI for Multiple Long-term Conditions
Research Support Facility

The Alan Turing Institute

Synthetic data for a health research programme

Dr Rachael Stickland & Dr Batool Almarzouq

Trustworthy Synthetic Data in Practice Workshop

23rd January 2024

DATA ARE PEOPLE

Batool Almarzouq

Research Project Manager

AI for Multiple Long-term Conditions
Research Support Facility

UNIVERSITY OF LIVERPOOL

Honorary Research Fellow

The Turing Way

Core Contributor

Einstein Award Jury

EINSTEIN Foundation.de

INTERNATIONAL ASSOCIATION OF UNIVERSITIES
INTERNATIONAL UNIVERSITIES BUREAU

IAU Expert Group on Open Science

Talk Overview

Batool

- **Introduction** to the research programme (AIM)
- **Introduction** to the Research Support Facility (RSF)

Rachael

- **Large population health datasets** used by AIM
- Why can **synthetic versions of health datasets** help research?
- **Our experiences** with synthetic health datasets
  - Current progress and future plans
  - Learnings and reflections
- Summary and acknowledgements

Stickland, R., & Almarzouq, B. (2024, January 22). Synthetic data for a health research programme. Trustworthy Synthetic Data in Practice, Warwick University. Zenodo. https://doi.org/10.5281/zenodo.10552753

All the slides can be accessed online

£23 million investment by the NIHR

Research Support Facility (RSF)

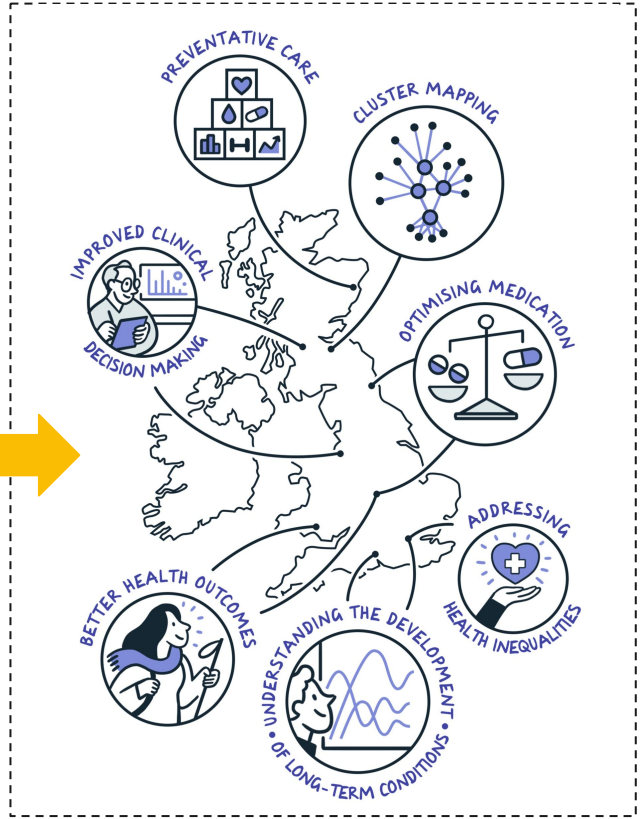
£23 million investment by the NIHR

Research Support Facility (RSF)

AI for Multiple Long Term Conditions (AIM) Community

£23 million investment by the NIHR

Research Support Facility (RSF)



AI for Multiple Long Term Conditions (AIM) Community

27 Universities

12 NHS Trusts

& charities, city councils, and public involvement organisations.

AI for Multiple Long-term Conditions
Research Support Facility

Slide re-used from RSF lightning talk in AI-UK by Sophia Batchelor: DOI 10.5281/zenodo.8082165.

£23 million investment by the NIHR

Research Support Facility (RSF)

Transformed MLTC research culture and environments

MLTC = Multiple-Long Term Conditions

AI for Multiple Long Term Conditions (AIM) Community

£23 million investment by the NIHR

Research Support Facility (RSF)

Transformed MLTC research culture and environments

MLTC = Multiple-Long Term Conditions

AI for Multiple Long Term Conditions (AIM) Community

Swansea University
Prifysgol Abertawe

The Alan Turing Institute

THE UNIVERSITY of EDINBURGH

Slide re-used from presentation developed by Eirini Zormpa

Swansea University
Prifysgol Abertawe

The Alan Turing Institute

THE UNIVERSITY of EDINBURGH

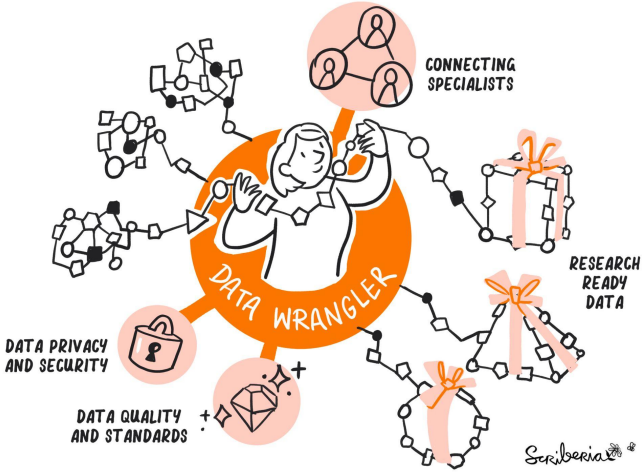
Theme 2: Research-ready data

Slide re-used from presentation developed by Eirini Zormpa

I am talking from the perspective of a **researcher and data scientist**, not a data provider

I am a **Data Wrangler** at The Alan Turing Institute

- Team: **Data for Research (Data Wrangling)**
- Project: **Artificial Intelligence for Multiple Long-Term Conditions (AIM)** for Theme 2: Research ready data
- Recently joined **Innovate UK BridgeAI** as an Independent Scientific Advisor

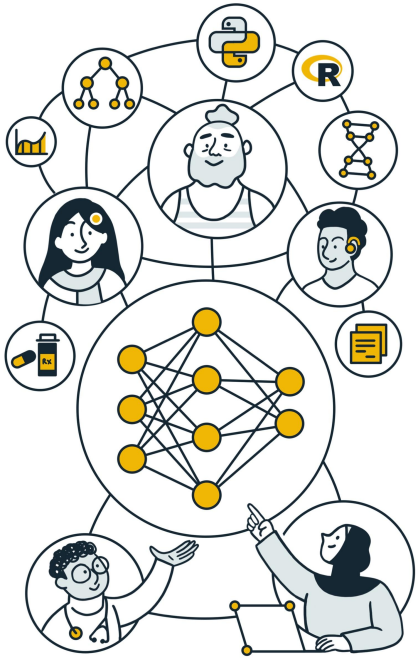


The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 licence. DOI: 10.5281/zenodo.3332807.

AI for Multiple Long-term Conditions
Research Support Facility

Using AI to address important questions about human health

- ★ Which health conditions **cluster together** and why?
- ★ Which **conditions** come first?
  - Which one might come next?
- ★ Does **clustering or trajectory change** for different groups?
- ★ Can we **identify** those most at risk of a condition?
  - Or most at risk of high severity?
- ★ What is the **impact of polypharmacy**?
  - Can we offer a better combination of drugs?
- ★ How can we have a **holistic (patient centered)** approach rather than a single disease focus?



AI for Multiple Long-term Conditions
Research Support Facility

These questions require large population health datasets

SAIL DATABANK

Welsh population
From ONS, NHS & more

DataLoch

Primary and secondary health care in the Lothian region (Scotland)

CPRD

Primary care data (and linked secondary) across UK

UK Data Service

Longitudinal studies
e.g. birth cohorts, ageing populations

biobank uk

Biomedical database of genetic, lifestyle and health information from sample of UK population

AI for Multiple Long-term Conditions
Research Support Facility

How can synthetic versions of health datasets help?

FOR RESEARCH TEAMS

- **Dataset suitability**
  - Application time & money
  - What infrastructure, software, personnel? (particularly if no TRE)
- **Dataset familiarity** → (transferable) preliminary analytical workflows
- **Research applications** → train models, bias correction, sample boost

FOR DATA PROVIDERS

- **Promote datasets** (with accompanying educational resources)
- **Receive better applications** for the real datasets

Sharing synthetic data, and derived outputs, should have lower disclosure risk
Ensuring privacy is a priority for health datasets

Clinical Practice Research Datalink

CPRD

Delivered by the Medicines & Healthcare products Regulatory Agency

with support from NIHR | National Institute for Health and Care Research

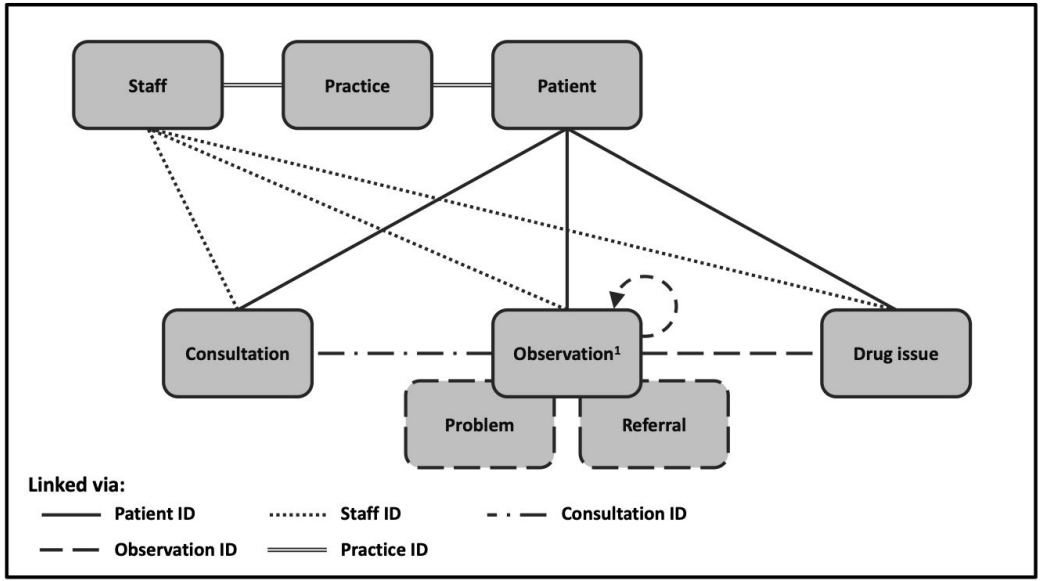
- **Anonymised longitudinal primary care data** from GPs across UK
- **Linkage** to secondary care
- **Over 60 million patients,** at least **16 million** currently registered with GP
- For **>30 years** it has been used for research
- Resulting in **>3,000 peer-reviewed** publications
- **Two main datasets** - Gold & Aurum

https://cprd.com

Clinical Practice Research Datalink

CPRD

CPRD Aurum dataset structure



- Fully-coded patient **electronic health records**
- **Tabular** data
- 8 **linked** tables
- 2 data **dictionaries** (medical & drug)
- Multiple **lookup tables**

https://cprd.com

AI for Multiple Long-term Conditions
Research Support Facility

CPRD synthetic datasets

Medium: resemble data types, values, formats & structure, and table relationships

High: replicate complex clinical relationships in real data while protecting patient privacy

		MSL holder	Non-MSL holder
Medium-fidelity datasets *x2*	CPRD GOLD and CPRD Aurum Sample Datasets together	£0	£1,500
High-fidelity datasets* *x2*	CPRD Cardiovascular disease synthetic dataset	£200	£200
	CPRD COVID-19 symptoms and risk factors synthetic dataset	£200	£200

*For non-teaching applications like complex statistical analyses as well as machine learning and artificial intelligence (AI) research applications. Where the COVID-19 and Cardiovascular synthetic datasets will be used for teaching, there is a £1,100 fee applicable per dataset.

Real datasets cost **x10** or **x100** more, depending on type of license

CPRD operates on a cost-recovery basis

Source:
https://cprd.com/pricing

AI for Multiple Long-term Conditions
Research Support Facility

CPRD synthetic datasets

High fidelity synthetic dataset generation & evaluation methodology (fidelity & privacy):

ORIGINAL ARTICLE | Open Access

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

Zhenchen Wang, Puja Myles, Allan Tucker

First published: 03 January 2021 | https://doi.org/10.1111/coin.12427 | Citations: 12

Article | Open access | Published: 09 November 2020

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

Allan Tucker, Zhenchen Wang, Ylenia Rotalinti & Puja Myles

npj Digital Medicine **3**, Article number: 147 (2020) | Cite this article

- Probabilistic Graphical Modelling (Bayesian Networks)
- Approaches for modelling missing data
- Use-case: complex statistical analysis and ML research applications

Medium fidelity synthetic dataset

- Generated with similar algorithms, did not maintain multivariate relationships
- Use-case: data management & training, develop/validate tools, improve algorithms

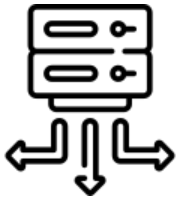
AI for Multiple Long-term Conditions
Research Support Facility

Our goals with CPRD (medium fidelity) synthetic datasets

GOALS

- **Streamline** the process for researchers using real CPRD data
- **Document** data management and processing strategies
- **Share** code in open languages (postgreSQL, Python, R)

OUTPUTS



- Collation of CPRD **metadata**, available tools & resources
- **Pre-processing workflow**

  File conversions → QC → tables in a relational database
- **.ipynb** for familiarisation with tables & to build a sample cohort

Our goals with CPRD (medium fidelity) synthetic datasets

FUTURE GOALS

- Create **educational resources** & work with CPRD to maximise impact of their synthetic data for health research
- Address most **common challenges** across research groups
- **Harmonise** data wrangling approaches across data sources

CHALLENGES

- Synthetic data is notably **smaller in size**
- **Data license agreement** has restrictions
- Hosting the data - **"is it personal data?"**
- Do not reinvent the wheel

These experiences are without a TRE model, but that will change in the future:

cprd.com/cprd-trusted-research-environment

AI for Multiple Long-term Conditions
Research Support Facility

Synthetic Health Data: Reflections & Learnings

- **Synthetic datasets are being created** by many organisations (e.g., CPRD, UKBiobank, NHS, ONS)

- **Low fidelity** → similar benefits to excellent metadata?
- **Medium/high fidelity** → starts to replicate research activities

- Synthetic data used to ***address*** **bias & disclosure risk, but can also** ***introduce*** **it!**

- **Data governance needs adapting** to maximise impact
- **Consensus is needed** on terminology & methodology

Machine learning

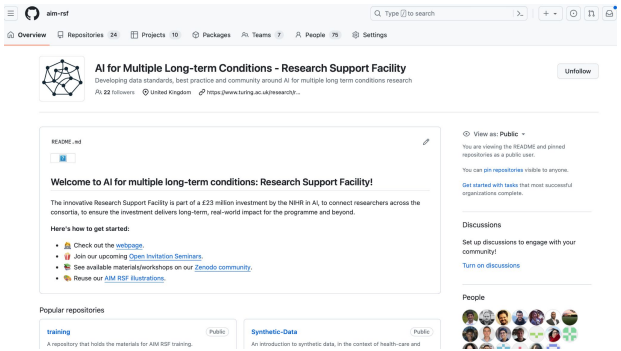
Data augmentation

Utility, Fidelity, Privacy

Partially versus Fully Synthetic

AI for Multiple Long-term Conditions
Research Support Facility

AIM RSF GitHub Organisation



AIM Newsletter

These slides used flaticon.com for icons

RSF Team

FUNDED BY
NIHR | National Institute for Health and Care Research

Kirstie Whitaker

Ronan Lyons

Evelina Gabasova

David Ford

Monica Fletcher

Ann-Marie Mallon

Aziz Sheikh

Gabriella Linning

Sydney Ambrose

Bastian Greshake Tzovaras

Emma Karoune

Sophia Batchelor

Mahwish Mohammad

Jon Smart

Chris Orton