HDR UK Reproducible Machine Learning Project

Collaborators Meeting

Date of meeting: **9.30am-12noon on Wednesday 31 March 2021**

Purpose of this meeting: To share progress and identify new opportunities around reproducible machine learning in health data science.

Meeting Schedule:

Time	Item	Speakers
9:30	Why we need reproducible machine learning to support trustworthy clinical insight	**Aiden Doherty** (Oxford)
9:50	How should researchers report machine learning in health data science?	**Gary Collins** (Oxford)
10:05	What are the opportunities and challenges around the generation of synthetic healthcare datasets?	**Allan Tucker** (Brunel)
10:20	How can one conduct reproducible clinically-relevant machine learning in restricted 'safe-haven' environments?	**James Liley** (Edinburgh)
10.35	**Break** (15mins)	
10:50	Tools and open datasets to support training activities around reproducible machine learning: an activity monitoring use-case.	**Shing Chan** (Oxford)
11.05	What opportunities are there to embed reproducible machine learning across national training programmes?	**Thanasis Tsanas** (Edinburgh)
11.20	How can we learn from the public voice to positively contribute to reproducible machine learning in healthcare?	**Sophie Staniszewska** (Warwick)
11.35	How might outputs from this project be more discoverable by researchers across HDR UK?	**Susheel Varma** (HDR UK)
11.50	AOB and Closing Remarks	**Aiden Doherty**
12:00	**Meeting Ends**	

Attended:

	First Name	Surname	Institute
1	**Aiden**	**Doherty (Chair)**	**University of Oxford**
2	**Bella**	**Pratt (Administration)**	**University of Oxford**
3	Louis	Aslett	Durham University / The Alan Turing Institute
4	Shing	Chan	University of Oxford
5	Peter	Charlton	University of Cambridge
6	Gary	Collins	University of Oxford
7	Emanuele	Di Angelantonio	University of Cambridge
8	Tim	Hubbard	King's College London
9	James	Liley	University of Edinburgh
10	Paolo	Missier	Newcastle University
11	Marcus	Munafo	University of Bristol
12	Hollydawn	Murray	HDR UK
13	Heidi	Seibold	Alan Turing Institute
14	Sophie	Staniszewska	University of Warwick
15	Darren	Treanor	University of Leeds
16	Thanasis	Tsanas	University of Edinburgh
17	Allan	Tucker	Brunel University
18	Catalina	Vallejos	Alan Turing Institute
19	Susheel	Varma	HDR UK
20	Sebastian	Vollmer	Alan Turing Institute

The meeting slides will now follow in the order outlined in the programme above.

HDRUK
Health Data Research UK

Why we need reproducible analytics to support trustworthy clinical insights

Aiden Doherty[1], Chris Holmes[2,3], Martin Landray[1], Gary Collins[4], Catalina Vallejos[3,5], Sebastian Vollmer[3], Emanuele Di Angelantonio[6], Louis Aslett[3], Alastair Denniston[7], Verena Heise[4], Bilal Mateen[3], James Rudd[6], Sophie Staniszewska[7], Thanasis Tsanas[5], Kirstie Whittaker[3,6], Gabriella Rustici[2], John Danesh[6], Harry Hemingway[8], Tim Hubbard[8], Cathie Sudlow[5]

Shing Chan[1],
James Liley[3,5],
Arabella Pratt[1]

1 Health Data Research UK Oxford 2 Health Data Research UK
3 Alan Turing Institute
4 University of Oxford
5 Health Data Research UK Scotland
6 Health Data Research UK Cambridge
7 Health Data Research UK Midlands
8 Health Data Research UK London

Why machine learning could transform medicine

HDRUK
Health Data Research UK

Data generation and data acquisition is becoming cheap

- driven by advances in digital measurement technologies
- genomes; images; health records; internet; wearables; EHRs
- BioBanks and longitudinal cohorts e.g. UK Biobank, 100,000 Genomes project

Coupled to increasing raw computing power (GPUs) that facilitate compute hungry algorithms

High level (governmental) recognition of AI and data as a resource

Note: the actual methods, such as deep neural networks, are not too dissimilar to those used in the 1980s

Clinical areas where machine learning can add substantial value

HDRUK
Health Data Research UK

Diagnosis

Risk prediction

Decision support

Typified by large quantities of rich, high-dimensional, and multi-modal data with clearly defined objectives. Areas where large data volumes make human decision analysis laborious.

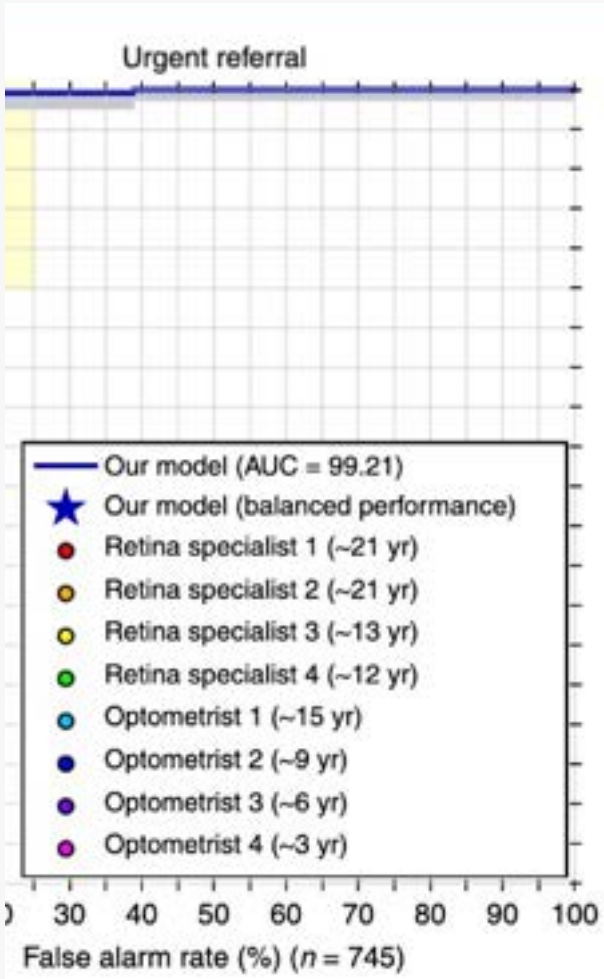
Where we have cryptic “hard to define” features

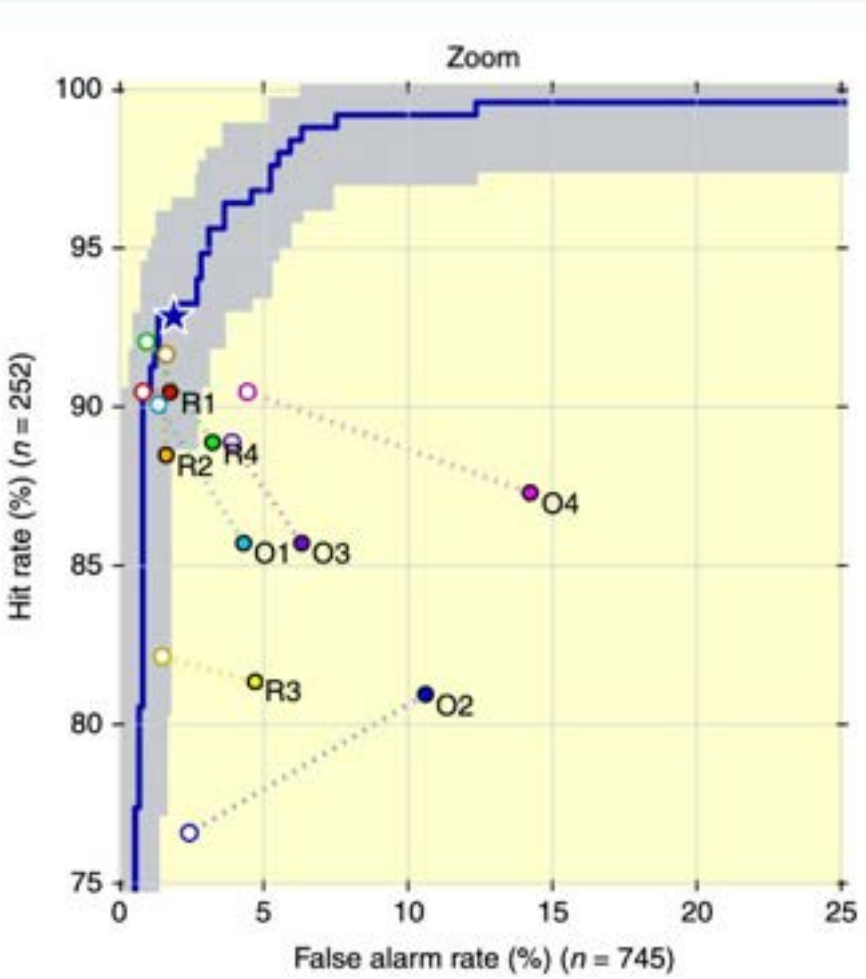
- Imaging
- Electronic health records free text
- Wearables and consumer devices
- Other areas where handcrafted variables would be challenging and traditional statistical models difficult to elicit

Machine learning offers promise to support diagnostic decisions

Imaging - Diabetic retinopathy

HDRUK
Health Data Research UK





De Fauw, et al ***Nature Medicine*** 2018; 24(9) 1342-1350.

Machine learning offers promise for risk prediction
Electronic health records - Acute kidney injury

HDRUK
Health Data Research UK

Artificial intelligence + Add to myFT

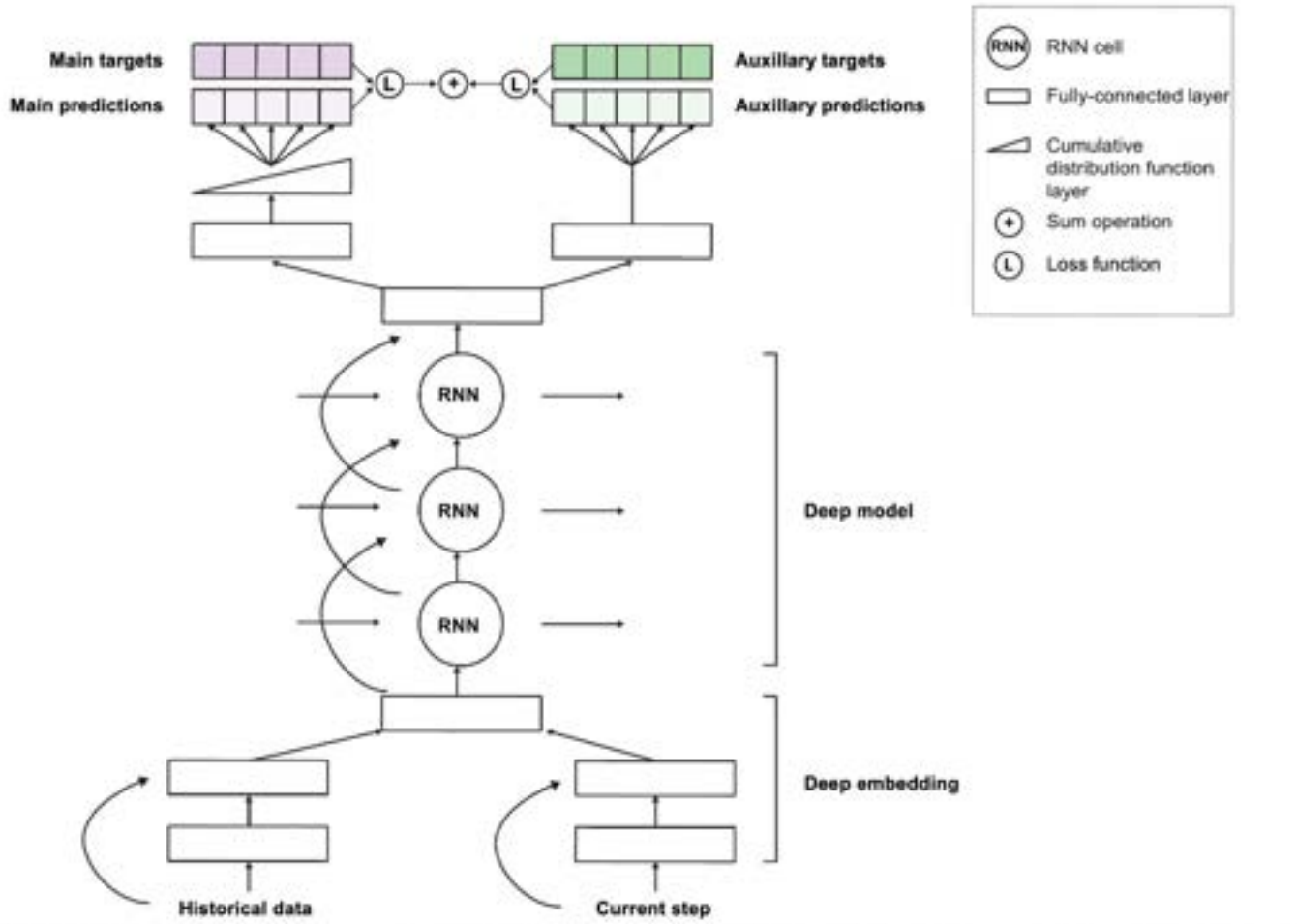
DeepMind creates algorithm to predict kidney damage in advance

AI company's model can give 48 hours' warning of potentially fatal complications

The machine learning model was developed jointly by DeepMind Health, a division of Google, and the US Department of Veterans Affairs

Madhumita Murgia in London JULY 31 2019 28

Artificial intelligence can now warn critical care doctors that their patients are at risk of developing severe kidney damage up to two days early, with the potential to save hundreds of thousands of lives every year.



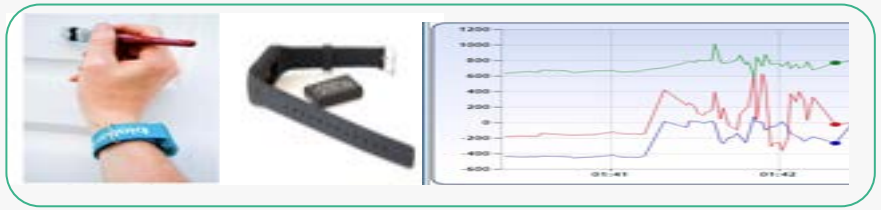
Extended Data Fig. 2 | Architecture of the proposed model. The best performance was achieved by a multi-task deep recurrent highway network architecture on top of an L_1-regularized deep residual embedding component that learns the best data representation end-to-end without pre-training.

Tomašev, N., Glorot, X., Rae, J.W. et al.. ***Nature*** 572, 116–119 (2019).

Machine learning can improve exposure measurement
Physical activity & cardiovascular disease

HDRUK
Health Data Research UK

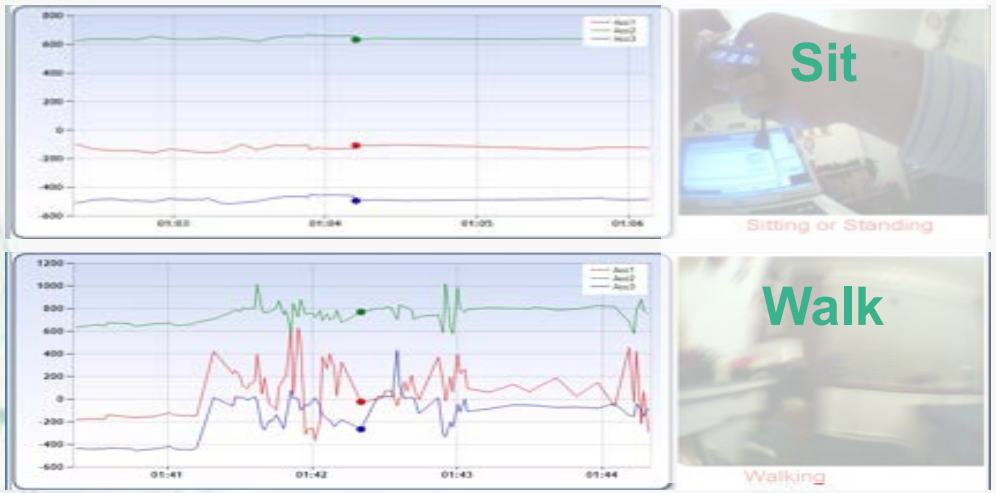
Large health sensor datasets now available



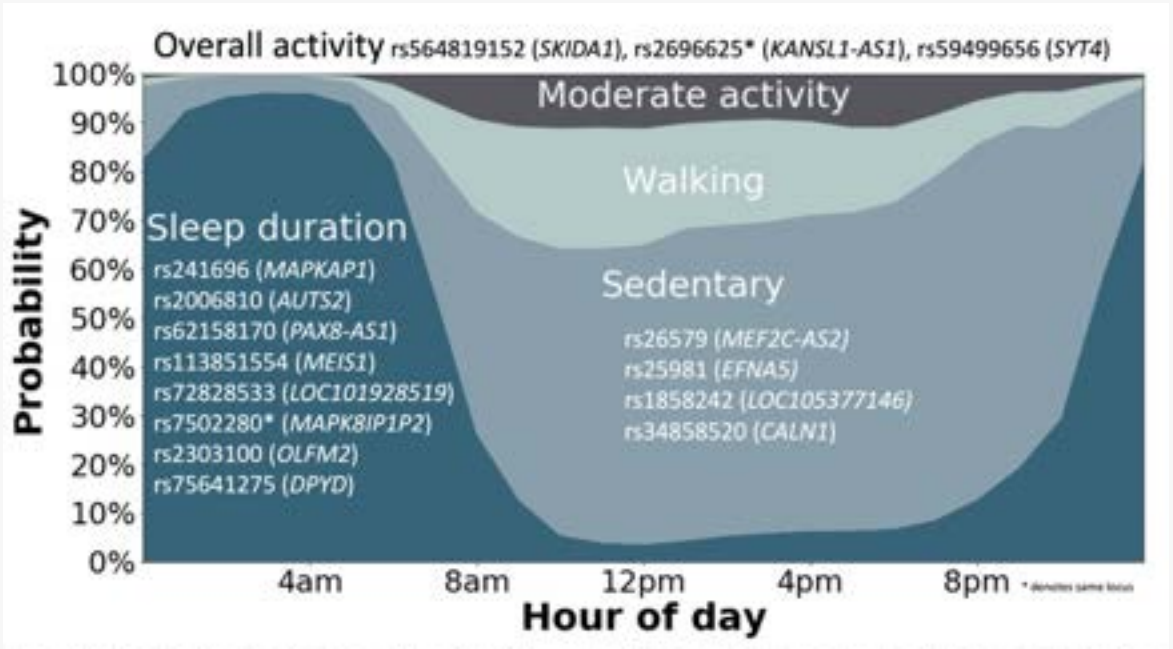
UK Biobank - 100,000 people

- Range of health measurements
- Linkage to clinical outcomes

Classify activities of daily living



New insights on sleep & physical activity



New biological insight

Causal association with blood pressure & depression

Importance of activity has been underestimated

Willetts, M., Hollowell, S., Aslett, L., Holmes, C. & Doherty, A. ***Sci. Rep.*** 8, 7961 (2018).
Doherty, Smith-Bryne, Ferreira, MV Holmes, C Holmes, Pulit, Lindgren ***Nature Communications*** 9:5257 (2018)

Should I always use machine learning methods?

ELSEVIER

Journal of Clinical Epidemiology 110 (2019) 12–22

Journal of Clinical Epidemiology

REVIEW

A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models

Evangelia Christodoulou[a], Jie Ma[b], Gary S. Collins[b,c], Ewout W. Steyerberg[d], Jan Y. Verbakel[a,e,f], Ben Van Calster[a,d,*]

[a]Department of Development & Regeneration, KU Leuven, Herestraat 49 box 805, Leuven, 3000 Belgium
[b]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Windmill Road, Oxford, OX3 7LD UK
[c]Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[d]Department of Biomedical Data Sciences, Leiden University Medical Centre, Albinusdreef 2, Leiden, 2333 ZA The Netherlands
[e]Department of Public Health & Primary Care, KU Leuven, Kapucijnenvoer 33J box 7001, Leuven, 3000 Belgium
[f]Nuffield Department of Primary Care Health Sciences, University of Oxford, Woodstock Road, Oxford, OX2 6GG UK

Accepted 5 February 2019; Published online 11 February 2019

Abstract

Objectives: The objective of this study was to compare performance of logistic regression (LR) with machine learning (ML) for clinical prediction modeling in the literature.

Study Design and Setting: We conducted a Medline literature search (1/2016 to 8/2017) and extracted comparisons between LR and ML models for binary outcomes.

Results: We included 71 of 927 studies. The median sample size was 1,250 (range 72–3,994,872), with 19 predictors considered (range 5–563) and eight events per predictor (range 0.3–6,697). The most common ML methods were classification trees, random forests, artificial neural networks, and support vector machines. In 48 (68%) studies, we observed potential bias in the validation procedures. Sixty-four (90%) studies used the area under the receiver operating characteristic curve (AUC) to assess discrimination. Calibration was not addressed in 56 (79%) studies. We identified 282 comparisons between an LR and ML model (AUC range, 0.52–0.99). For 145 comparisons at low risk of bias, the difference in logit(AUC) between LR and ML was 0.00 (95% confidence interval, −0.18 to 0.18). For 137 comparisons at high risk of bias, logit(AUC) was 0.34 (0.20–0.47) higher for ML.

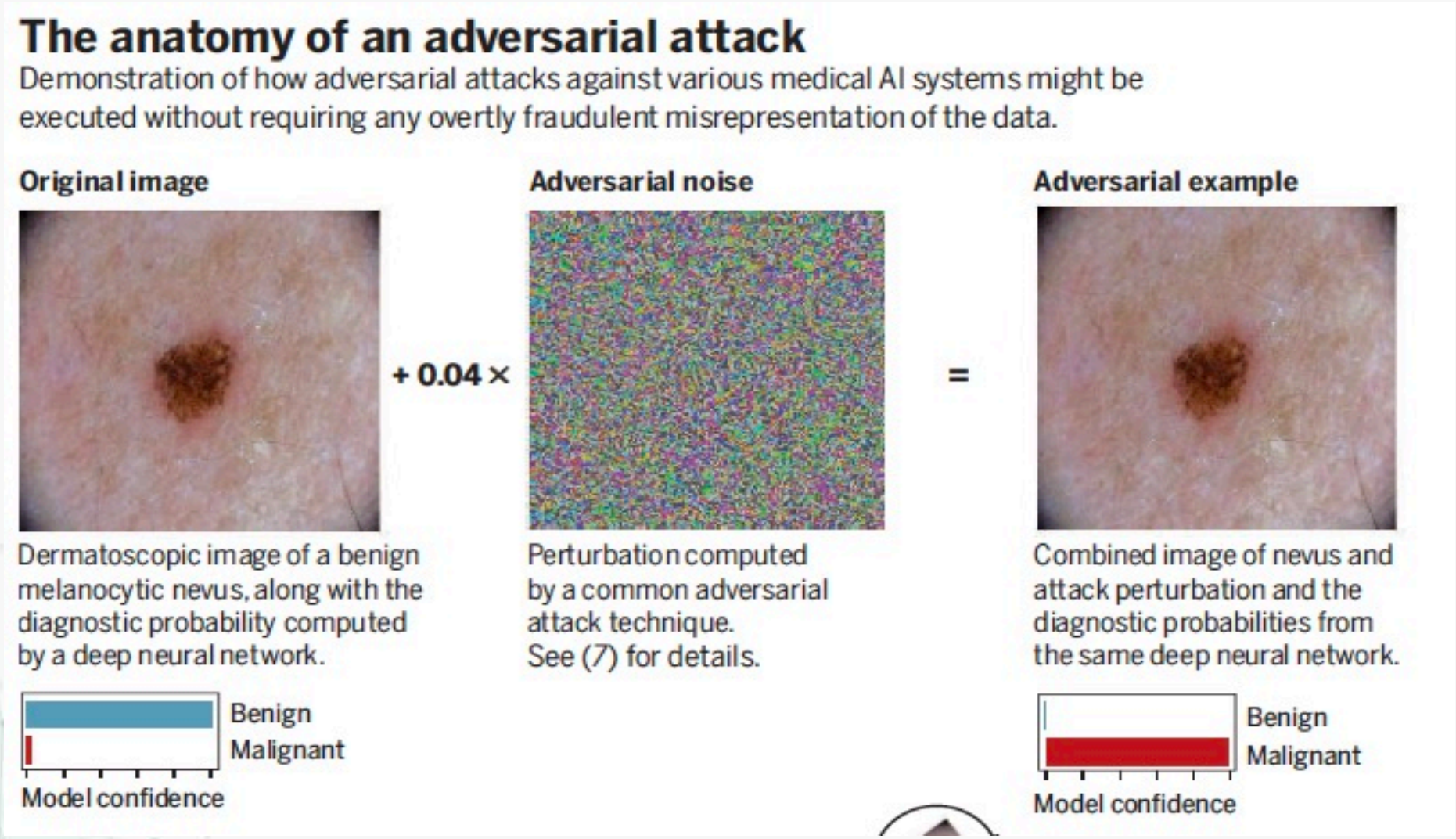
Conclusion: We found no evidence of superior performance of ML over LR. Improvements in methodology and reporting are needed for studies that compare modeling algorithms.

Keywords: Clinical prediction models; Logistic regression; Machine learning; AUC; Calibration; Reporting

HDRUK
Health Data Research UK

Is the machine learning model robust to erroneous input?

A small change to the input causes a large, highly confident, change to the prediction



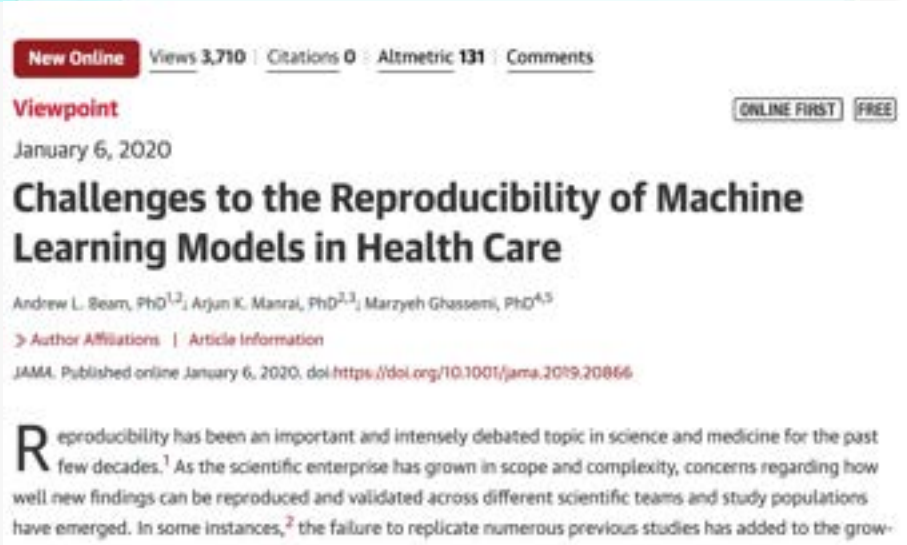
Finlayson et al ***Science*** 2019; 22 Mar 2019: Vol. 363, Issue 6433, pp. 1287-1289

There is not an engrained culture of reproducible practices when using machine learning in medicine

HDRUK
Health Data Research UK

- Reporting of machine learning methods is not standardised[1]
- Researchers using the open MIMIC dataset, for the same challenge, report vastly different sample sizes[2]
- While completing the same neuroimaging task, different analysis tools provide widely different results[3]
- Data sharing is rarely done and often restricted[4]

		Data: Same	Data: Different
Analysis: Same		Reproducible	Replicable
Analysis: Different		Robust	Generalisable



New Online | Views 3,710 | Citations 0 | Altmetric 131 | Comments

Viewpoint

ONLINE FIRST | FREE

January 6, 2020

Challenges to the Reproducibility of Machine Learning Models in Health Care

Andrew L. Beam, PhD[1,2]; Arjun K. Manrai, PhD[2,3]; Marzyeh Ghassemi, PhD[4,5]

> Author Affiliations | Article Information

JAMA. Published online January 6, 2020. doi:https://doi.org/10.1001/jama.2019.20866

Reproducibility has been an important and intensely debated topic in science and medicine for the past few decades.[1] As the scientific enterprise has grown in scope and complexity, concerns regarding how well new findings can be reproduced and validated across different scientific teams and study populations have emerged. In some instances,[2] the failure to replicate numerous previous studies has added to the grow-

1 Rajkomar, A., Oren, E., Chen, K. et al. ***npj Digital Med*** 1, 18 (2018).
2 Johnson, et al. ***Machine Learning for Healthcare Conference***, 2017.
3 Bowring et al. ***Hum Brain Mapp*** 2019;40:3362-3384
4 *Miyakawa* ***Molecular Brain*** (2020) 13:24
5 Hutson ***Science*** 2018; 16 Feb 2018: Vol. 359, Issue 6377, pp. 725-726

Reproducible machine learning for health lags behind other fields

An evaluation of 511 scientific papers across several machine learning subfields

HDRUK
Health Data Research UK

SCIENCE TRANSLATIONAL MEDICINE | PERSPECTIVE

Evaluation metrics

- **A** Technical reproducibility
  - 1 Code available
  - 2 Public dataset
- **B** Statistical reproducibility
  - 1 Variance reported
- **C** Conceptual reproducibility (replicability)
  - 1 Multiple datasets

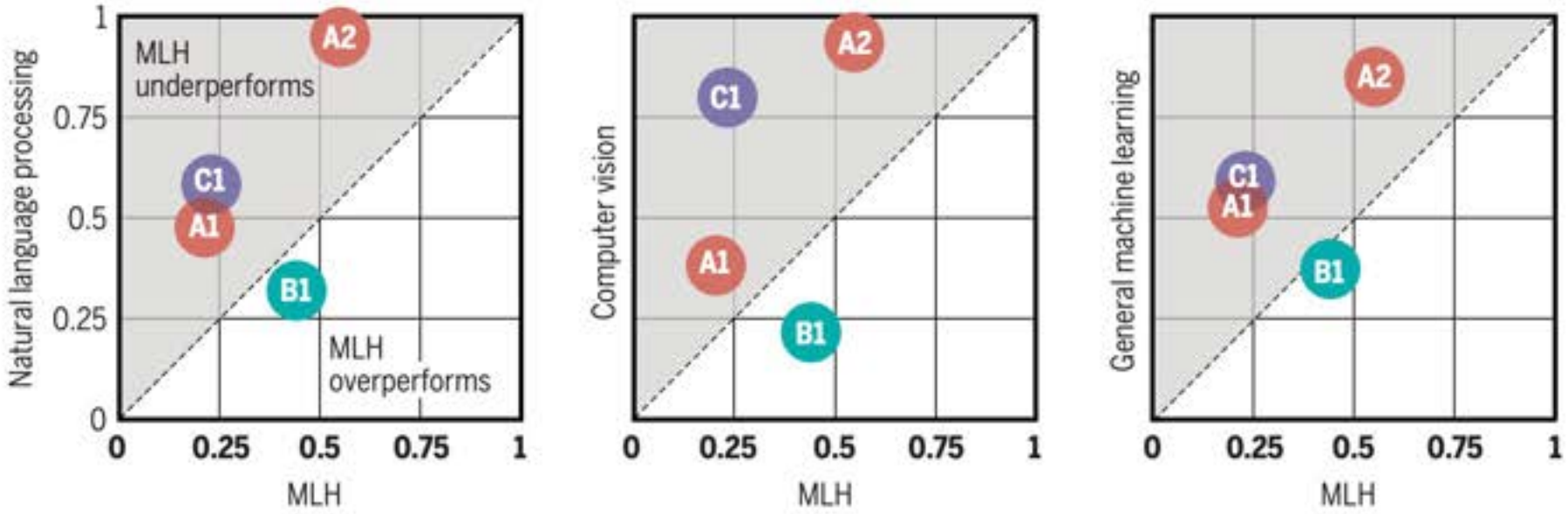


Fig. 1. Reproducibility metrics for machine learning applications. Shown are reproducibility metrics (A, B, and C) for evaluating scientific papers from four machine learning subspecialties: machine learning in health (MLH), natural language processing, computer vision, and general machine learning. Presented is the fraction of papers in a given subspecialty (*y* axis) versus those in MLH (*x* axis) that release their code (A1), release their data (A2), report their variance (B1), and leverage multiple datasets (C1). MLH consistently lags other subfields of machine learning on all measures of reproducibility apart from inclusion of proper statistical variance.

McDermott et al., ***Sci. Transl. Med.*** 13, eabb1655 (2021) 24 March 2021

Proposed solutions to improve reproducibility of ML in health

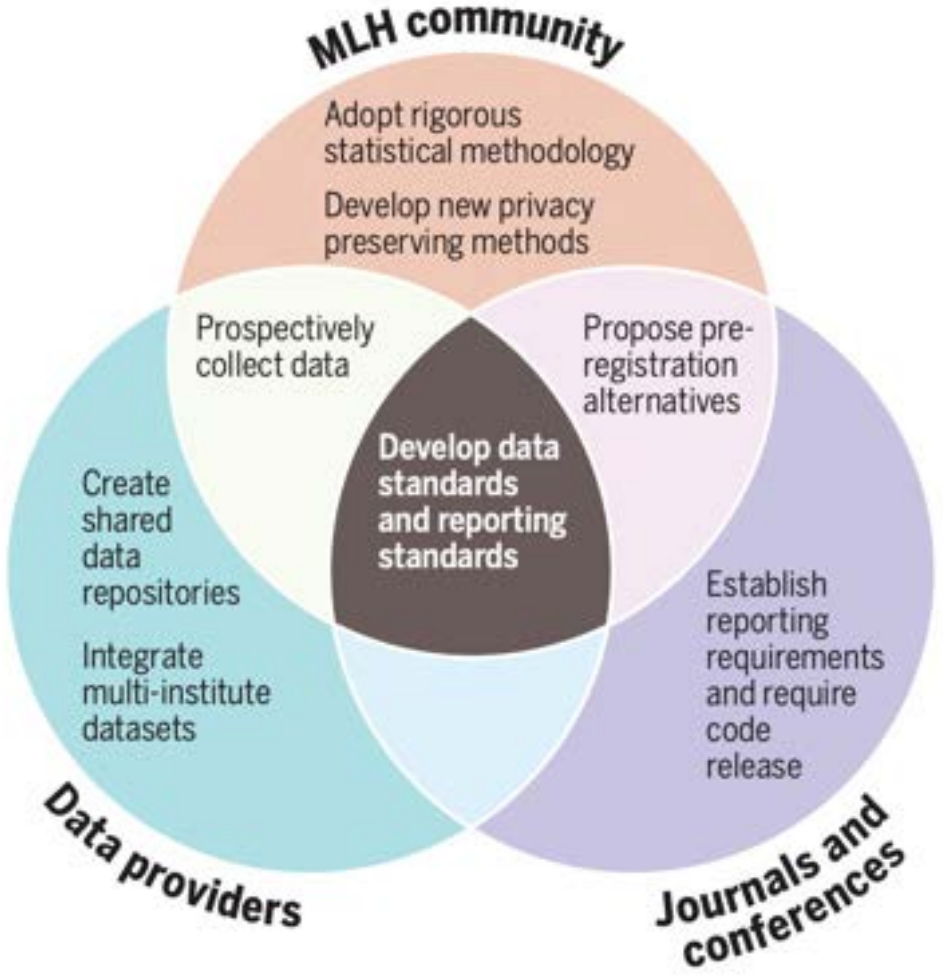


Fig. 2. Improving reproducibility in MLH research. Shown are eight recommendations for improving the reproducibility of machine learning in health research. These recommendations are subdivided according to which primary stakeholders directly drive these changes: the community of MLH researchers, health data providers (e.g., clinical care organizations), and journals and conferences (the primary publishers of machine learning research).

HDRUK
Health Data Research UK

McDermott et al., ***Sci. Transl. Med.*** 13, eabb1655 (2021) 24 March 2021

National project – Reproducible machine learning

HDRUK
Health Data Research UK

Can reproducible machine learning be embedded in UK health data science to support trustworthy clinical insights?

1) How should researchers report machine learning in health data science?

2) Can synthetic datasets be used to evaluate the stability of machine learning models in health data science?

3) What are the minimal requirements for reproducibility in restricted ‘safe-haven’ environments?

4) Can we initiate a culture of reproducible machine learning?

Main achievements in first year

Implementation project has provided flexibility to align with new opportunities around impactful driver projects

WP1 reporting - TRIPOD-AI

WP2 synthetic data – wider HDR UK activities

WP3 safe haven environments - SPARRA

WP4 training - HDR UK Summer School & UK Biobank

PPIE – one of the first efforts in this space

Need to avoid temptation to ‘boil the ocean’. Machine learning and reproducibility are large fields in their own right.

Meeting aim: *To share progress and identify new opportunities around reproducible machine learning in health data science.*

HDRUK
Health Data Research UK

Time	Item	Speakers
9:30	Why we need reproducible machine learning to support trustworthy clinical insight	**Aiden Doherty** (Oxford)
9:50	How should researchers report machine learning in health data science?	**Gary Collins** (Oxford)
10:05	What are the opportunities and challenges around the generation of synthetic healthcare datasets?	**Allan Tucker** (Brunel)
10:20	How can one conduct reproducible clinically-relevant machine learning in restricted 'safe-haven' environments?	**James Liley** (Edinburgh)
10.35	**Break** (15mins)	
10:50	Tools and open datasets to support training activities around reproducible machine learning: an activity monitoring use-case.	**Shing Chan** (Oxford)
11.05	What opportunities are there to embed reproducible machine learning across national training programmes?	**Thanasis Tsanas** (Edinburgh)
11.20	How can we learn from the public voice to positively contribute to reproducible machine learning in healthcare?	**Sophie Staniszewska** (Warwick)
11.35	How might outputs from this project be more discoverable by researchers across HDR UK?	**Susheel Varma** (HDR UK)
11.50	AOB and Closing Remarks	**Aiden Doherty**
12:00	**Meeting Ends**	

Improving the reporting of clinical prediction models developed using machine learning (TRIPOD-AI)

Gary Collins
Professor of Medical Statistics
Centre for Statistics in Medicine
University of Oxford

TRIPOD Statement (prediction models using regression)

- **Published in Jan 2015, in 11 journals**
- **Focus on models developed using regression**
  - However, guidance is relevant for ML
- **Explanation document (73 pages) focusses solely on examples from regression**
  - Touches on conduct
  - Opportunity to flag methodological issues
  - e.g., model presentation/availability
- **Needs to be tailored to the ML community**
  - e.g. examples, terminology, model presentation/availability
- **TRIPOD-AI launched [Collins & Moons, Lancet 2019]**

Annals of Internal Medicine RESEARCH AND REPORTING METHODS

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement

Gary S. Collins, PhD; Johannes B. Reitsma, MD, PhD; Douglas G. Altman, DSc; and Karel G.M. Moons, PhD

Prediction models are developed to aid health care providers in estimating the probability or risk that a specific disease or condition is present (diagnostic models) or that a specific event will occur in the future (prognostic models), to inform their decision making. However, the overwhelming evidence shows that the quality of reporting of prediction model studies is poor. Only with full and clear reporting of information on all aspects of a prediction model can risk of bias and potential usefulness of prediction models be adequately assessed. The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Initiative developed a set of recommendations for the reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes. This article describes how the TRIPOD Statement was developed. An extensive list of items based on a review of the literature was created, which was reduced after a Web-based survey and revised during a 3-day meeting in June 2011 with methodologists, health care professionals, and journal editors. The list was refined during several meetings of the steering group and in e-mail discussions with the wider group of TRIPOD contributors. The resulting TRIPOD Statement is a checklist of 22 items, deemed essential for transparent reporting of a prediction model study. The TRIPOD Statement aims to improve the transparency of the reporting of a prediction model study regardless of the study methods used. The TRIPOD Statement is best used in conjunction with the TRIPOD explanation and elaboration document. To aid the editorial process and readers of prediction model studies, it is recommended that authors include a completed checklist in their submission (also available at www.tripod-statement.org).

Ann Intern Med. 2015;162:55-63. doi:10.7326/M14-0697 www.annals.org

For author affiliations, see end of text.

For contributors to the TRIPOD Statement, see the Appendix (available at www.annals.org).

Annals of Internal Medicine RESEARCH AND REPORTING METHODS

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration

Karel G.M. Moons, PhD; Douglas G. Altman, DSc; Johannes B. Reitsma, MD, PhD; John P.A. Ioannidis, MD, DSc; Petra Macaskill, PhD; Ewout W. Steyerberg, PhD; Andrew J. Vickers, PhD; David F. Ransohoff, MD; and Gary S. Collins, PhD

The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) Statement includes a 22-item checklist, which aims to improve the reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes. The TRIPOD Statement aims to improve the transparency of the reporting of a prediction model study regardless of the study methods used. This explanation and elaboration document describes the rationale; clarifies the meaning of each item; and discusses why transparent reporting is important, with a view to assessing risk of bias and clinical usefulness of the prediction model. Each checklist item of the TRIPOD Statement is explained in detail and accompanied by published examples of good reporting. The document also provides a valuable reference of issues to consider when designing, conducting, and analyzing prediction model studies. To aid the editorial process and help peer reviewers and, ultimately, readers and systematic reviewers of prediction model studies, it is recommended that authors include a completed checklist in their submission. The TRIPOD checklist can also be downloaded from www.tripod-statement.org.

Ann Intern Med. 2015;162:W1-W73. doi:10.7326/M14-0698 www.annals.org

For author affiliations, see end of text.

For members of the TRIPOD Group, see the Appendix.

Importance of reporting


Research: increasing value, reducing waste 5

Reducing waste from incomplete or unusable reports of biomedical research

- Poor reporting limits (or prevents) the use of new research findings - avoidable waste

- Research reproducibility
  - …impossible without complete reporting of all relevant aspects of scientific design, conduct, measurements, data and analysis

- Critical appraisal (risk of bias assessment)
  - Crucial for assessing study design and methods

- Readers need a clear understanding of exactly what was done
  - Clinicians, researchers, systematic reviewers, policy makers
  - Full and transparent reporting important for determining relevance and applicability

- Consequences of poor reporting include reducing or distorting the evidence base, usability of findings, and potential outcomes for clinical practice and ultimately to patients

- Huge number of reviews evaluating reporting of publications

Why should we care?

Check for updates

FAST TRACK

Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal

Laure Wynants,[1,2] Ben Van Calster,[2,3] Gary S Collins,[4,5] Richard D Riley,[6] Georg Heinze,[7] Ewoud Schuit,[8,9] Marc M J Bonten,[8,10] Darren L Dahly,[11,12] Johanna A Damen,[8,9] Thomas P A Debray,[8,9] Valentijn M T de Jong,[8,9] Maarten De Vos,[2,13] Paula Dhiman,[4,5] Maria C Haller,[7,14] Michael O Harhay,[15,16] Liesbet Henckaerts,[17,18] Pauline Heus,[8,9] Michael Kammer,[7,19] Nina Kreuzberger,[20] Anna Lohmann,[21] Kim Luijken,[21] Jie Ma,[5] Glen P Martin,[22] David J McLernon,[23] Constanza L Andaur Navarro,[8,9] Johannes B Reitsma,[8,9] Jamie C Sergeant,[24,25] Chunhu Shi,[26] Nicole Skoetz,[19] Luc J M Smits,[1] Kym I E Snell,[6] Matthew Sperrin,[27] René Spijker,[8,9,28] Ewout W Steyerberg,[3] Toshihiko Takada,[8] Ioanna Tzoulaki,[29,30] Sander M J van Kuijk,[31] Bas C T van Bussel,[1,32] Iwan C C van der Horst,[32] Florien S van Royen,[8] Jan Y Verbakel,[33,34] Christine Wallisch,[7,35,36] Jack Wilkinson,[22] Robert Wolff,[37] Lotty Hooft,[8,9] Karel G M Moons,[8,9] Maarten van Smeden[8]

For numbered affiliations see end of the article.

Correspondence to: L Wynants laure.wynants@maastrichtuniversity.nl (ORCID 0000-0002-3037-122X)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2020;369:m1328 http://dx.doi.org/10.1136/bmj.m1328

ABSTRACT

OBJECTIVE

To review and appraise the validity and usefulness of published and preprint reports of prediction models for diagnosing coronavirus disease 2019 (covid-19) in patients with suspected infection, for prognosis of patients with covid-19, and for detecting people in the general population at increased risk of covid-19 infection or being admitted to hospital with the

DATA SOURCES

PubMed and Embase through Ovid, up to 1 July 2020, supplemented with arXiv, medRxiv, and bioRxiv up to 5 May 2020.

STUDY SELECTION

Studies that developed or validated a multivariable covid-19 related prediction model.

DATA EXTRACTION

At least two authors independently extracted data

Update 3 (1-July-2020)

- 169 studies describing 232 prediction models
  - 7 risk scores, 118 diagnostic; 107 prognostic
  - Mixture of modelling procedures

- Bottom line: 226 at high risk of bias; 6 at unclear risk of bias

- "This review indicates that almost all published prediction models are **poorly reported**, and at **high risk of bias** such that their reported predictive **performance is probably optimistic**. "

Reporting of machine learning models*

The completeness of reporting and adherence to the TRIPOD Statement of clinical prediction models using machine learning methods in oncology: a systematic review

Paula Dhiman[1,2], Jie Ma[1], Constanza Andaur Navarro[3], Beni Speich[1,4], Garrett Bullock[5], Shona Kirtley[1], Richard D Riley[6], Ben Van Calster[7], Karel GM Moons[3], Gary S Collins[1,2].

[1] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, OX3 7LD, UK

[2] NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom

[3] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

[4] Department of Clinical Research, Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, University of Basel, Basel, Switzerland

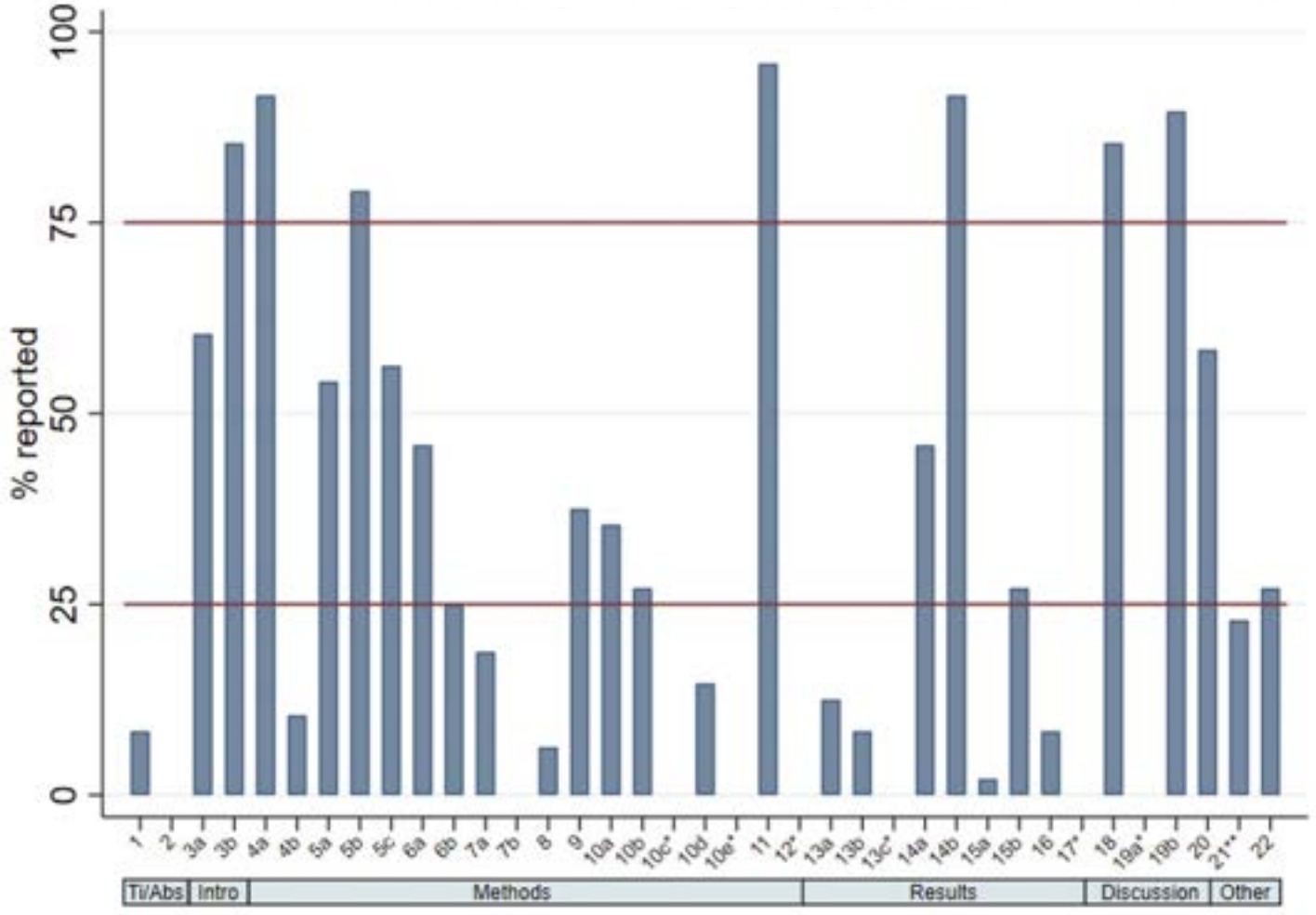
[5] Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, Oxford, UK

[6] Centre for Prognosis Research, School of Primary, Community and Social Care, Keele University, Staffordshire, UK, ST5 5BG

* Under review

Reporting deficiencies

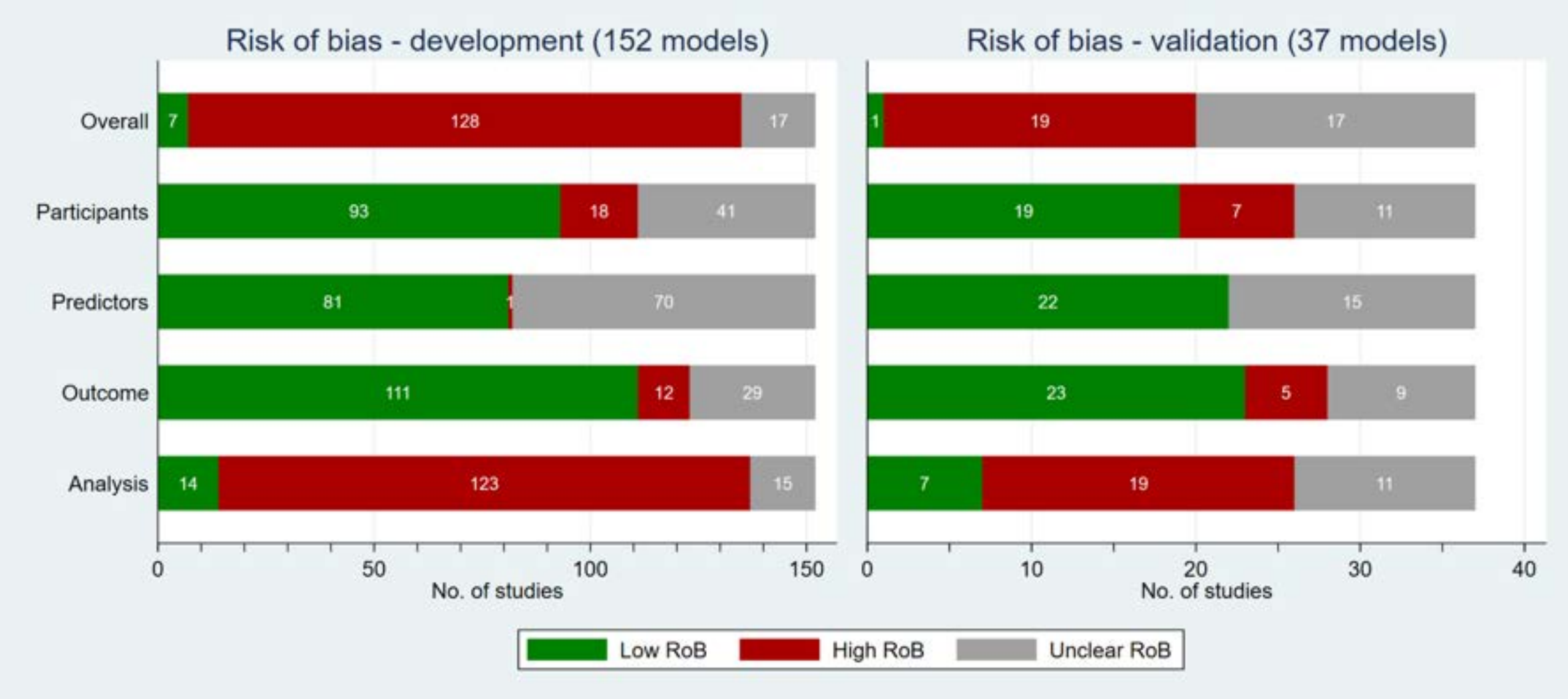
- Item 4b - study dates
- Item 8 - Sample size
- Item 10b - model building/internal validation
- Item 13b - characteristics of participants
- Item 15a - model availability
- Item 16 - performance measures with CIs



TRIPOD adherence

Lower adherence (≤15%)	Higher adherence (≥70%)
× Title	✓ Objectives
× Abstract	✓ Source of data
× Predictor assessment	✓ Eligibility criteria
× Sample size	✓ Reporting of risk groups
× Description/comparison of between development and validation data	✓ Unadjusted associations
× Participant flow and characteristics	✓ Limitations
× Presentation and interpretation of full model	✓ Overall interpretation of results
× Model performance	

Risk of bias – PROBAST



TRIPOD for machine learning/AI

THE LANCET

Access provided by University of Oxford

COMMENT | VOLUME 393, ISSUE 10181, P1577-1579, APRIL 20, 2019

Reporting of artificial intelligence prediction models

Gary S Collins ✉ · Karel G M Moons

Published: April 20, 2019 · DOI: https://doi.org/10.1016/S0140-6736(19)30037-6 · Check for updates

References

Article Info

Figures

Data-driven technologies that form the basis of the digital health-care revolution provide potentially important opportunities to deliver improvements in individual care and to advance innovation in medical research. Digital health technologies include mobile devices and health apps (m-health), e-health technology, and intelligent monitoring. Behind the digital health revolution are also methodological advancements using artificial intelligence and machine learning techniques. Artificial intelligence, which encompasses machine learning, is the scientific discipline that uses computer algorithms to learn from data, to help identify patterns in data, and make predictions. A key feature underpinning the excitement behind artificial intelligence and machine learning is their potential to analyse large and complex data structures to create prediction models that personalise and improve diagnosis, prognosis, monitoring, and administration of treatments, with the aim of improving individual health outcomes. Prediction models to support clinical decision making have existed for decades, and these include well known tools such as the Framingham Risk Score,[1] QRISK3,[2] Model for End-stage Liver Disease,[3] ABCD[2] score,[4] and the Nottingham Prognostic Index.[5] Health-care professionals, medical researchers, policy makers, guideline developers, patients, and members of the general public are all

https://osf.io/zyacb/

A protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction studies based on artificial intelligence

Gary S. Collins[1,2 *], Paula Dhiman[1,2], Constanza L. Andaur Navarro[3,4], Jie Ma[1], Lotty Hooft[3,4], Johannes B. Reitsma[3], Patricia Logullo[1], Andrew L. Beam[5,6], Lily Peng[7], Ben Van Calster[8,9,10], Maarten van Smeden[3], Richard D. Riley[11], Karel G.M. Moons[3,4 *]

[1] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, Oxford, OX3 7LD, United Kingdom

[2] NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, United Kingdom NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, United Kingdom

[3] Julius Center for Health Sciences & Primary Care, and Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

[4] Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, The Netherlands

[5] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States

[6] Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, United States

[7] Google Health, 3400 Hillview Ave, Palo Alto, CA 94304, United States

[8] KU Leuven, Department of Development and Regeneration, Leuven, Belgium

[9] Department of Biomedical Data Sciences, Leiden University Medical Centre (LUMC), Leiden, the Netherlands

[10] EPI-centre, KU Leuven, Leuven, Belgium

[11] Centre for Prognosis Research, School of Medicine, Keele University, Staffordshire, ST5 5BG, United Kingdom.

Consensus statement

- Delphi about to be launched to get consensus on items to include in the checklist
  - Interested in participating in the Delphi then contact me
- Anticipate TRIPOD-AI to be not too dissimilar to the original TRIPOD
- Biggest difference will be in the terminology, examples, and methods guidance
- Getting agreement on how to present/make the model available

Participant Information Sheet

Developing the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis for Artificial Intelligence (TRIPOD-AI)

What is the purpose of the survey?

The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) statements aims to improve the reporting of clinical prediction model studies, however, there are distinct challenges in artificial intelligence studies (such as transparency and description of the model, methodology, availability of data, description and replicability), which are not addressed by the current guidance. Given the rapid expansion of research in this area, there is an urgent need for guidance on the conduct and reporting of clinical prediction model studies involving artificial intelligence and machine learning. We are proposing to develop an extension to the TRIPOD statement to provide guidance for authors of studies involving artificial intelligence and machine learning (TRIPOD-AI).

Who is conducting the survey?

This project is led by the TRIPOD-AI group, which supported by a grant from HDR UK, Cancer Research (C49297/A27294), the NIHR Biomedical Research Centre, Oxford, and the Netherlands Organisation for Scientific Research.

The TRIPOD-AI group is led by Professor Gary Collins, (Professor of Medical Statistics, University of Oxford, United Kingdom) and Professor Carl Moons, (Professor Clinical Epidemiology, UMC Utrecht, the Netherlands).

Harmonisation of two languages

Statistics	Machine learning
Covariates	Features
Outcome variable	Target
Model	Network, graphs
Parameters	Weights
Model for discrete var.	Classifier
Model for continuous var.	Regression
Log-likelihood	Loss
Multinomial regression	Softmax
Measurement error	Noise
Subject/observation	Sample/instance
Dummy coding	One-hot encoding
Measurement invariance	Concept drift

Statistics	Machine learning
Prediction	Supervised learning
Latent variable modeling	Unsupervised learning
Fitting	Learning
Prediction error	Error
Sensitivity	Recall
Positive predictive value	Precision
Contingency table	Confusion matrix
Measurement error model	Noise-aware ML
Structural equation model	Gaussian Bayesian network
Gold standard	Ground truth
Derivation–validation	Training–test
Experiment	A/B test

Slide taken from Maarten van Smeden, Biometrischen Kolloquium, 2020

COMPUTER SCIENCE

Artificial intelligence faces reproducibility crisis

Unpublished code and sensitivity to training conditions make many claims hard to verify

The most basic problem is that researchers often don't share their source code. At the AAAI meeting, Odd Erik Gundersen, a computer scientist at the Norwegian University of Science and Technology in Trondheim, reported the results of a survey of 400 algorithms presented in papers at two top AI conferences in the past few years. He found that only 6% of the presenters shared the algorithm's code. Only a third shared the data they tested their algorithms on, and just half shared "pseudocode"—a limited summary of an algorithm. (In many cases, code is also absent from AI papers published in journals, including *Science* and *Nature*.)

Researchers say there are many reasons for the missing details: The code might be a work in progress, owned by a company, or held tightly by a researcher eager to stay ahead of the competition. It might be dependent on other code, itself unpublished. Or it might be that the code is simply lost, on a crashed disk or stolen laptop—what Rougier calls the "my dog ate my program" problem.

Reporting, Model availability + independent evaluation

e.g.,

- Make it available on a repository (e.g., GitHub)
- Grant access to get predictions for your data set
- Gain access to the code by setting-up non-disclosure agreements

Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist

Here we present the MI-CLAIM checklist, a tool intended to improve transparent reporting of AI algorithms in medicine.

Beau Norgeot, Giorgio Quer, Brett K. Beaulieu-Jones, Ali Torkamani, Raquel Dias, Milena Gianfrancesco, Rima Arnaout, Isaac S. Kohane, Suchi Saria, Eric Topol, Ziad Obermeyer, Bin Yu and Atul J. Butte

Reproducibility (Part 6): choose appropriate tier of transparency
Tier 1: complete sharing of the code
Tier 2: allow a third party to evaluate the code for accuracy/fairness; share the results of this evaluation
Tier 3: release of a virtual machine (binary) for running the code on new data without sharing its details
Tier 4: no sharing

Matters arising

Transparency and reproducibility in artificial intelligence

Table 2 | Frameworks to share code, software dependencies and deep-learning models

Resource	URL
Code	
BitBucket	https://bitbucket.org
GitHub	https://github.com
GitLab	https://about.gitlab.com
Software dependencies	
Conda	https://conda.io
Code Ocean	https://codeocean.com
Gigantum	https://gigantum.com
Colaboratory	https://colab.research.google.com
Deep-learning models	
TensorFlow Hub	https://www.tensorflow.org/hub
ModelHub	http://modelhub.ai
ModelDepot	https://modeldepot.io
Model Zoo	https://modelzoo.co
Deep-learning frameworks	
TensorFlow	https://www.tensorflow.org/
Caffe	https://caffe.berkeleyvision.org/
PyTorch	https://pytorch.org/

(anticipated) TRIPOD-AI Timelines

- 1st round of the Delphi survey to be launched shortly
- 2nd round of the Delphi survey to be launched in May
- Post Delphi checklist by June/July
- (online) Consensus meeting in the Autumn

*more information can be found on the OSF (https://osf.io/zyacb/)

Complementary initiatives

- PROBAST-AI (1st around of Delphi survey about to be launched)
  - Risk of bias tool for prediction models using AI)
- STARD-AI (1st round of Delphi survey closed)
  - Reporting of diagnostic test accuracy studies using AI
- QUADAS-AI (Delphi in preparation)
  - Risk of bias tool for diagnostic test accuracy studies using AI
- DECIDE-AI (1st round of Delphi survey will shortly be closed)
  - Reporting of studies looking at early clinical evaluation, human factors evaluation, preparatory steps towards large-scale clinical trials
- SPIRIT-AI/CONSORT-AI (already published in BMJ/Nature Med,…)
  - Protocols and results of clinical trials of AI interventions

Opportunities and challenges around the generation of synthetic healthcare data

Dr Allan Tucker
Intelligent Data Analytics Group,
Department of Computer Science, Brunel University London.

IDA Research

CPRD

NHSX

Medicines & Healthcare products Regulatory Agency

Brunel University London

BBC | Sign in | News | Sport | Weather | iPlayer | Sounds

NEWS

Home | Coronavirus | US Election | UK | World | Business | Politics | Tech | Science

Health

Care.data: How did it go so wrong?

Nick Triggle
Health correspondent

19 February 2014

digitalhealth
news networks intelligence

News | Features | Covid-19 | Jobs | CCIO Network | Health CIO | Intelligence | Events | Networks

AI and Analytics | Clinical Software | Cyber Security | Digital Patient | Infrastructure | Interoperability

NEWS

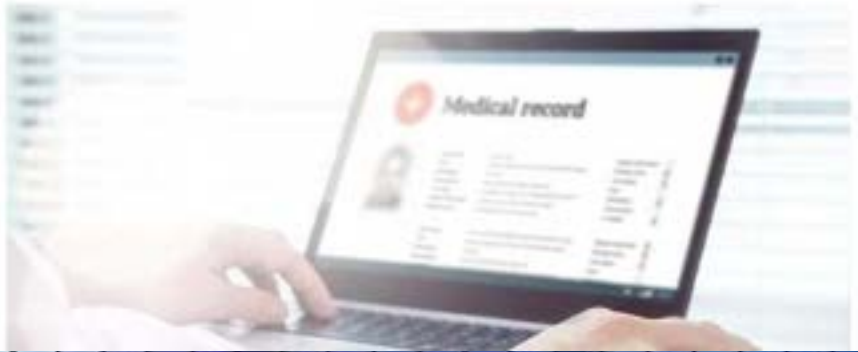
Data sharing | Great North Care Record | Joe McDonald | LHCREs | Matthew Gould | NHS Digital | NHS England | NHSX | Patient data | Sensyne Health | Simon Stevens | System C

NHS 'risks repeat of care.data in talks to commercialise medical records'

Andrea Downey

19 December 2019

Share this...



Background

- Adding Noise / Jitter
- K-anonymisation: removing or generalising data
  - but what is an acceptable K?
  - not good on high dimensions
- Differential Privacy: removing an individual has no effect on analysis
  - but repeated requests for aggregate data enabling id of individuals
  - Impact of rare cases (elderly individuals / rare disease / personalised medicine)

Generative Adversarial Networks

Data Science Campus

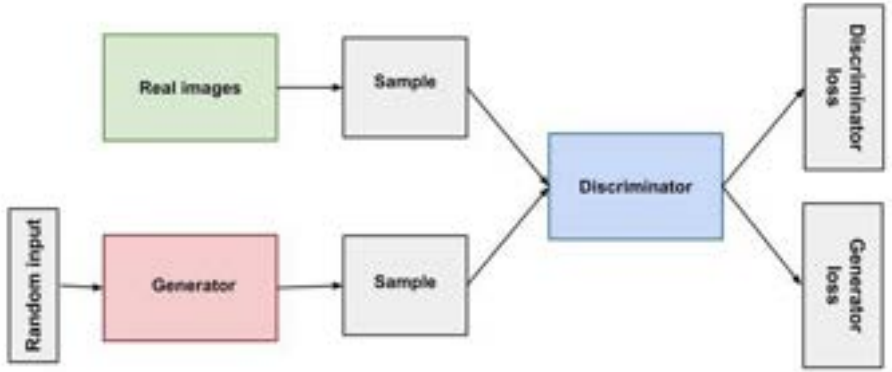
Search site

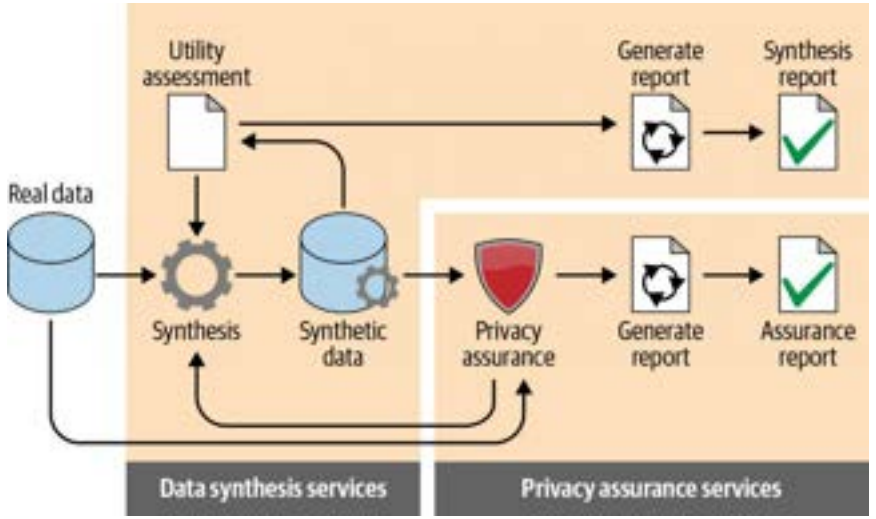
Home | About us | Partnerships | Projects | Learning and development | Team | Contact us

DATA SCIENCE FOR PUBLIC GOOD

Generative adversarial networks (GANs) for synthetic dataset generation with binary classes

Chaitanya Joshi | February 21, 2019



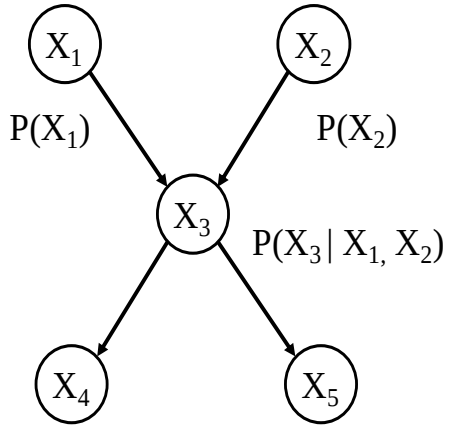


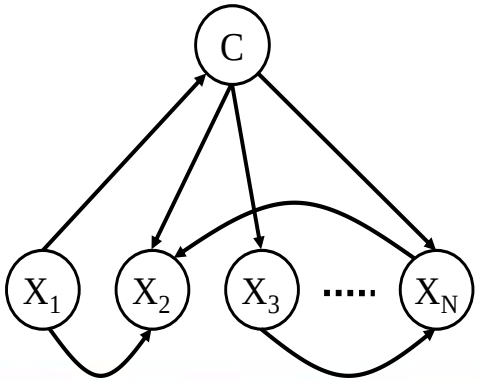
Brunel University London

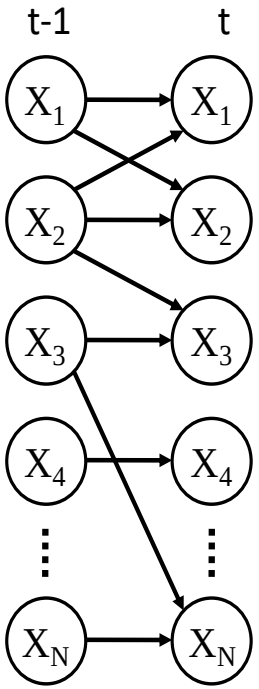
Bias and Transparency

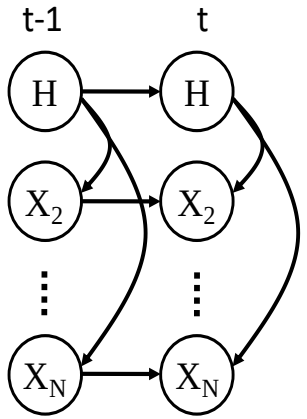
- Impacts of “Black Box” models
- Is the model / data / society biased?
  - Inductive bias of generative model learning algorithms hasn’t attracted enough attention
  - Data analysis is always “Secondary”, collected for different reasons
  - Covid is highlighting the biases we have in diagnosing / treating different groups
- What dependencies are represented in the model?
- Can we confirm known biological relationships?

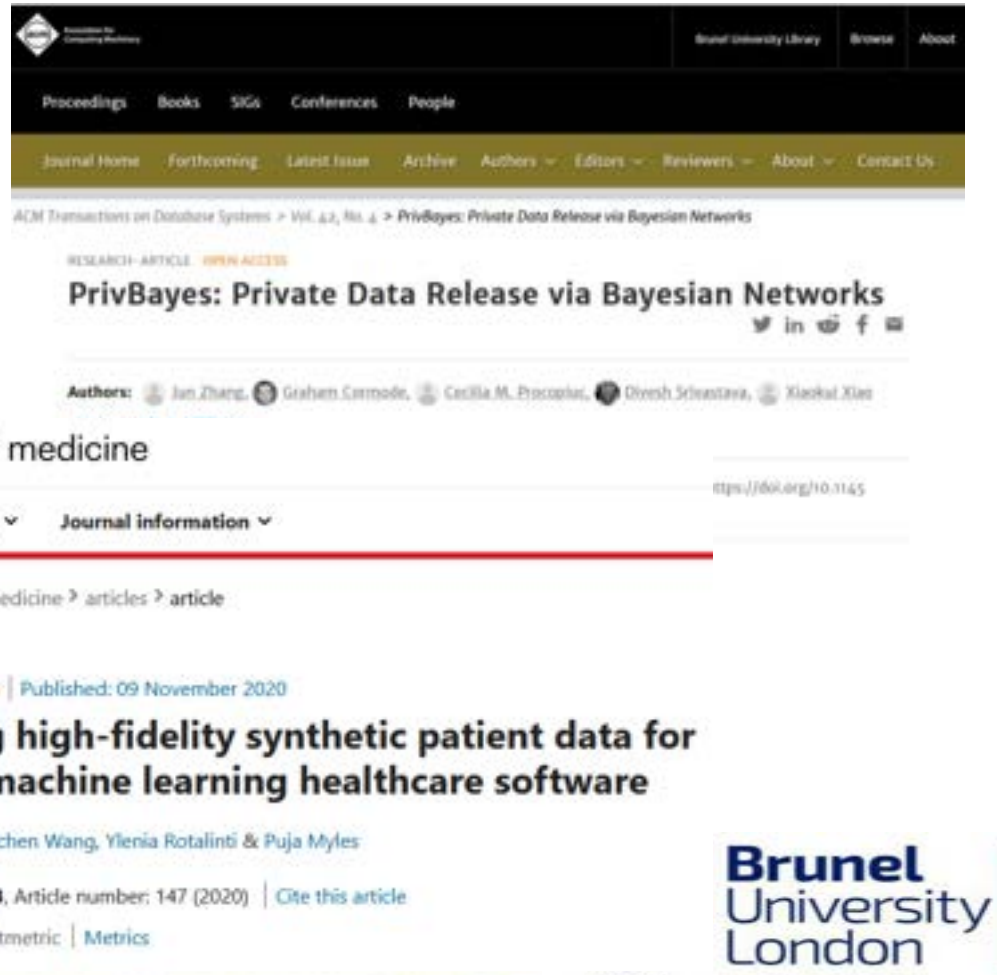
Probabilistic Graphical Networks (flexible conditional data generation)





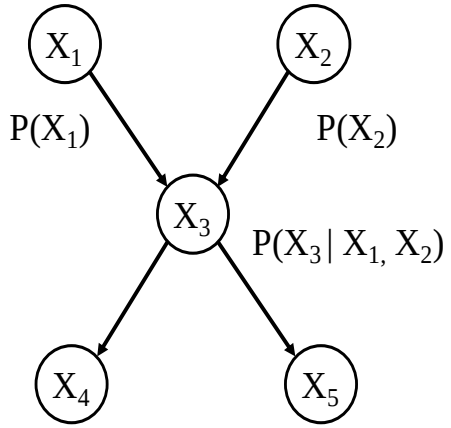


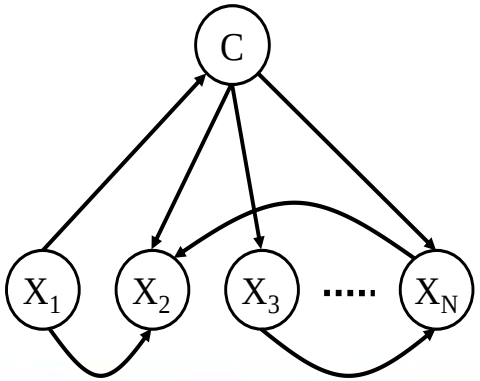


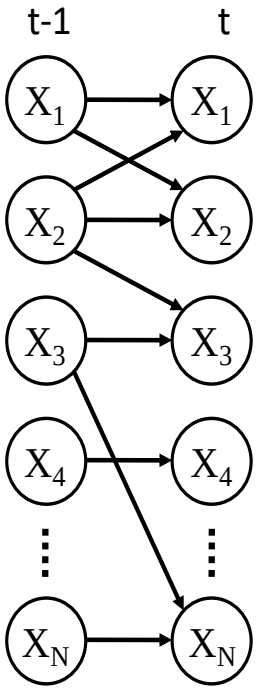


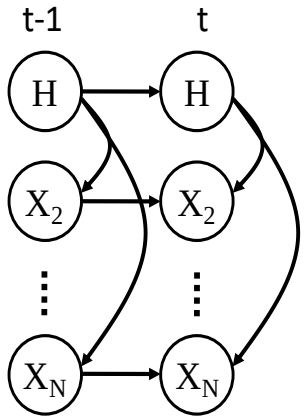
Brunel University London

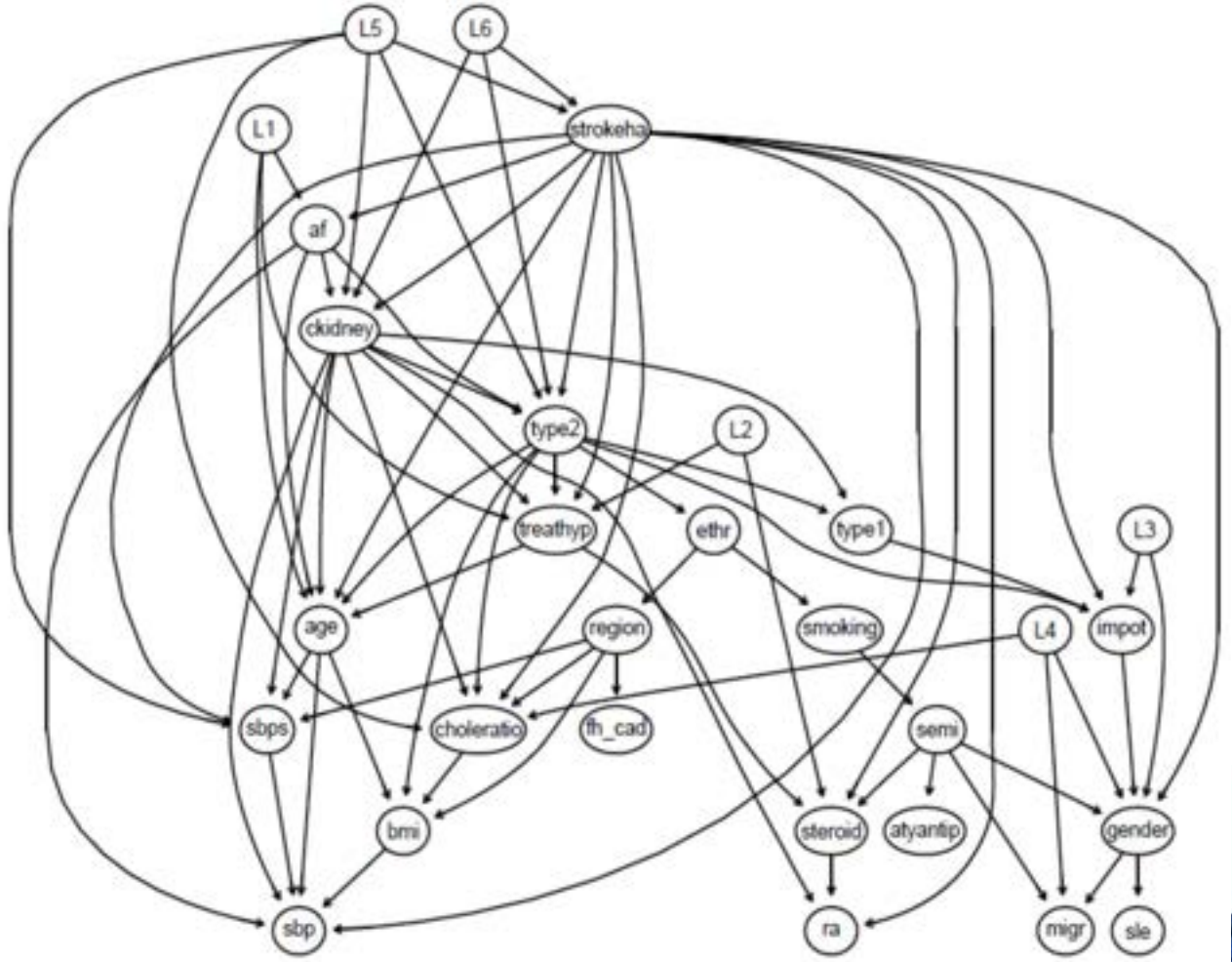
Probabilistic Graphical Networks (flexible conditional data generation)





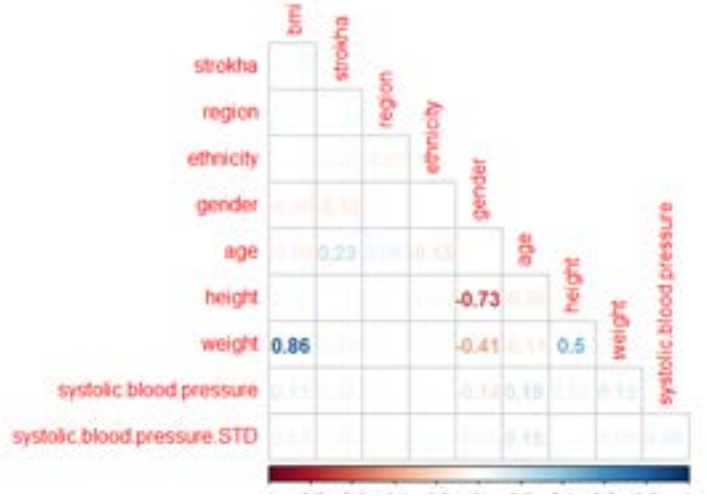




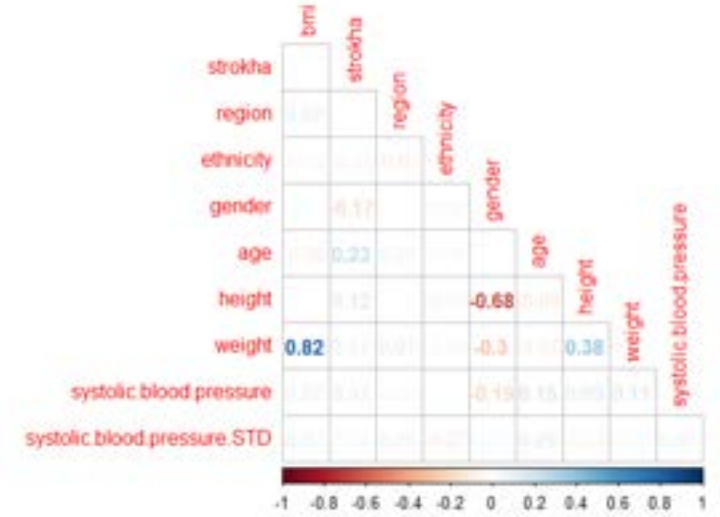


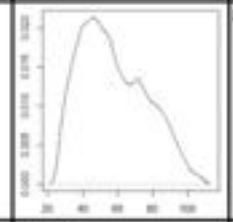
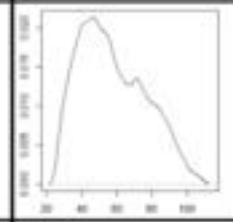
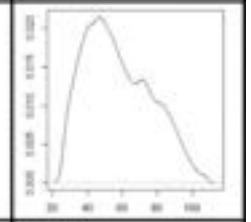
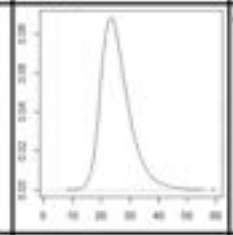
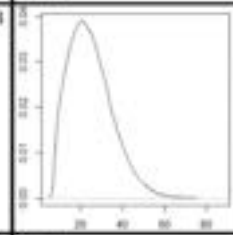
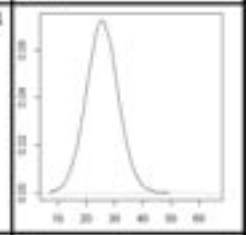
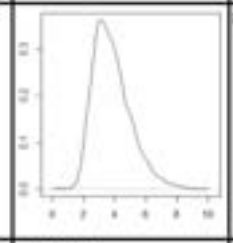
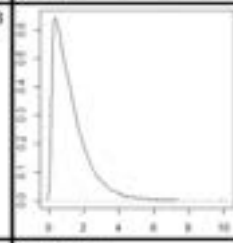
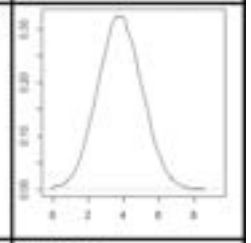
Fidelity – CPRD

ground truth



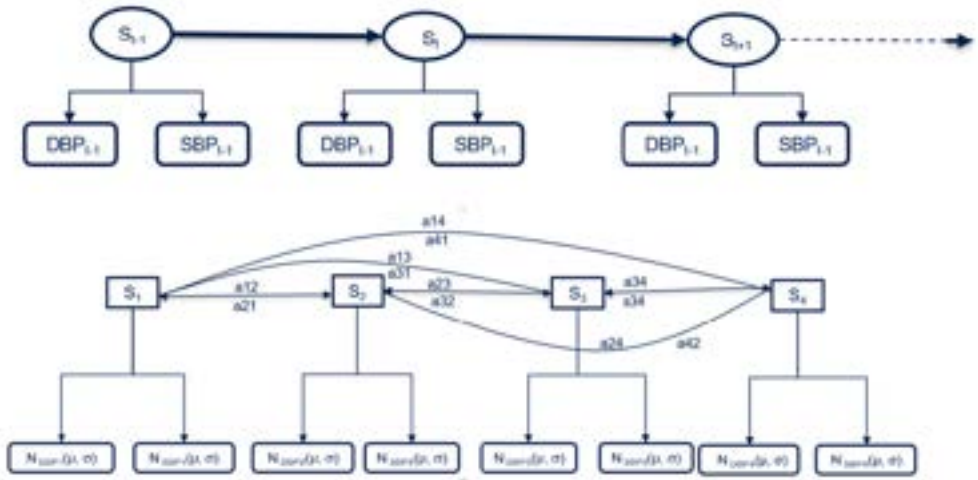
synthetic data

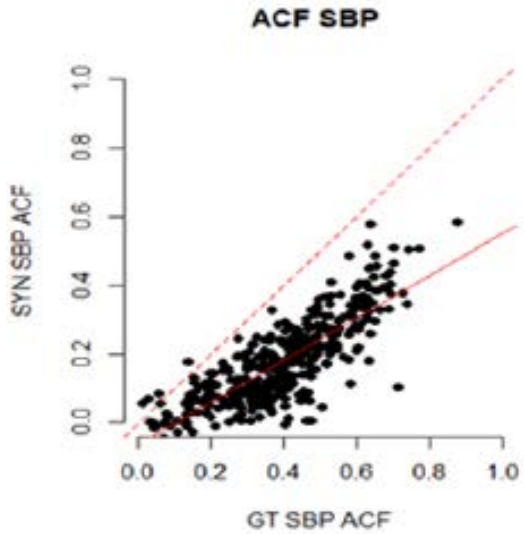


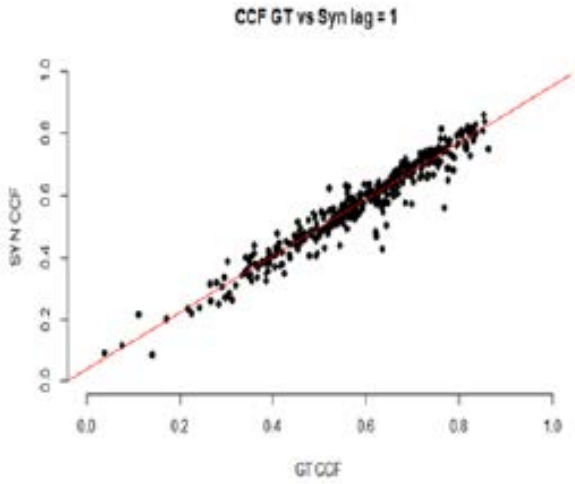
	a) Ground Truth – including missing data		b) Synthetic Data - Modelling GT in a) with missing nodes		c) Synthetic Data - Modelling GT in a) with latent variables	
Factor	Distinct / Missing	Distribution	Distinct / Missing	Distribution	Distinct / Missing	Distribution
age	84 / 0		84 / 0		84 / 0	
strokeha	2 / 0	F:93.6% T:6.4%	2 / 0	F:93.8% T:6.2%	2 / 0	F:93.6% T:6.4%
bmi	580 / 15.73		437 / 7.08		437 / 0.01	
choleratio	573 / 88.59		80 / 41.93		80 / 0.08	
smoking	5 / 0	1: (76.7%) 1: (14.2%) 2: (3.4%) 3: (3.9%) 4: (1.8%)	5 / 0	1: (76.8%) 1: (14.1%) 2: (3.3%) 3: (4.0%) 4: (1.8%)	5 / 0	1: (76.8%) 1: (14.1%) 2: (3.3%) 3: (4.0%) 4: (1.8%)

Brunel University London

Fidelity –MIMIC

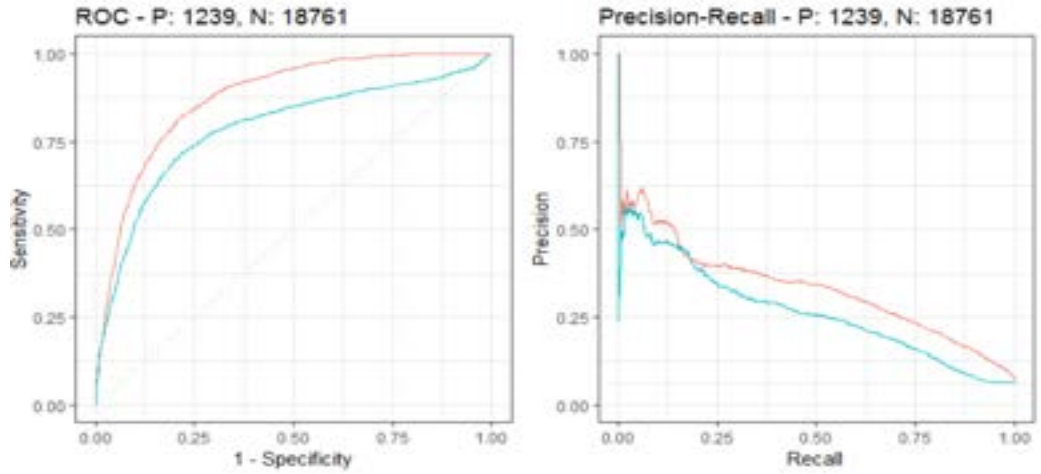


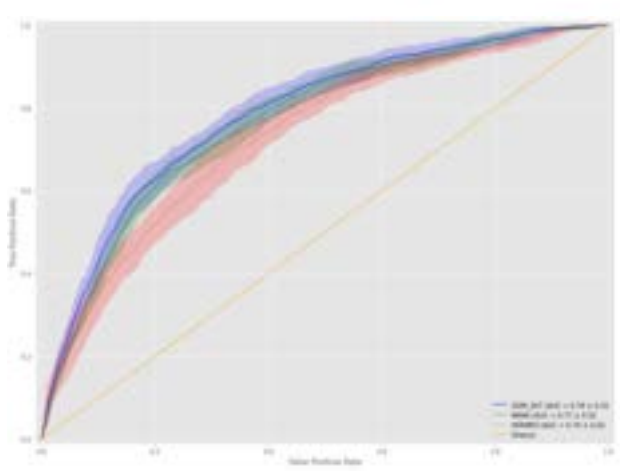


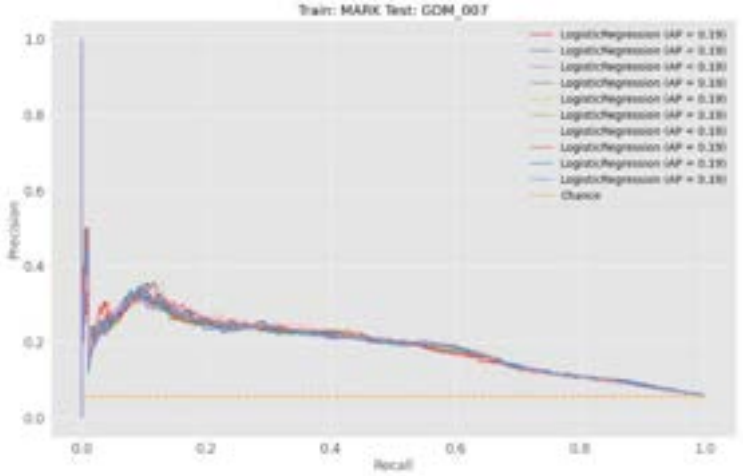


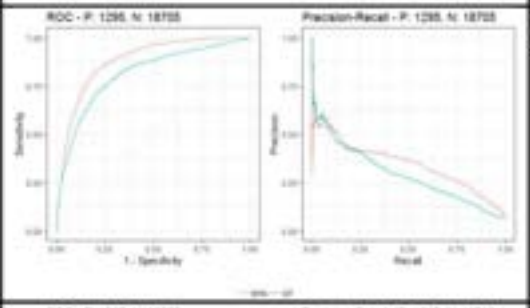
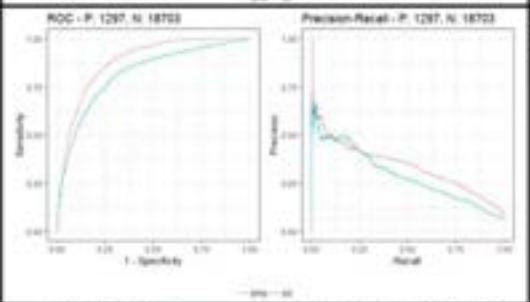
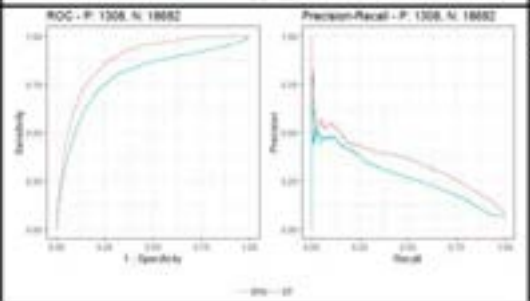
Brunel University London

Validation – MIMIC / Sensyne Health / CPRD







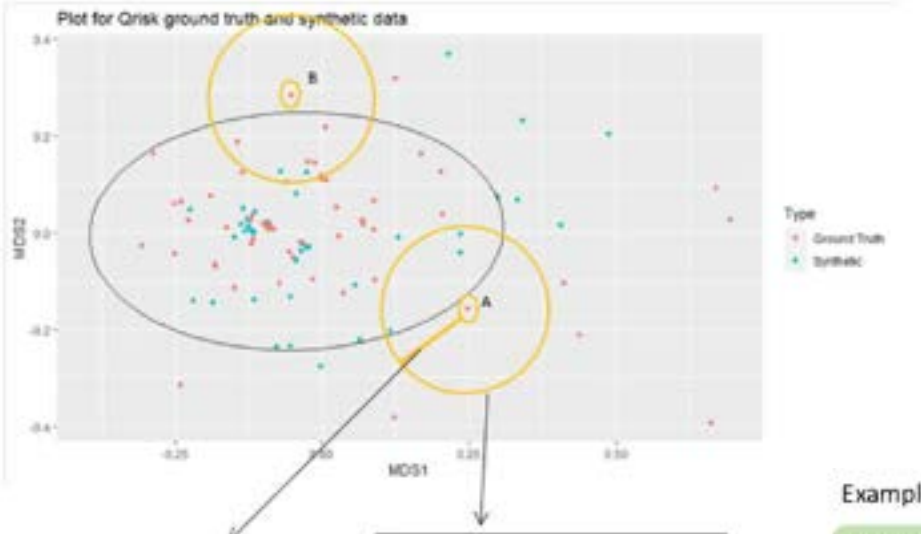
Iteration	ROC & PR curves (P=Positive cases, N = Negative cases)	Granger causality p value (α=0.05): PR, ROC	AUC GT: PR, ROC	AUC SYN: PR, ROC
1		<0.001, <0.001	0.293, 0.812	0.346, 0.881
2		<0.001, <0.001	0.292, 0.818	0.341, 0.881
3		<0.001, <0.001	0.274, 0.797	0.349, 0.880

Privacy Tests

- Explore effect of sample size and number of variables on risk of “clones” / “inliers” / “outliers”.
- Results are based on 10 iterations of resampling without replacement.
- Risk of inliers / outliers (single real datapoints close to a synthetic datapoint) always very small.
- Actual number of outliers generated stays relatively stable (between 10 and 70).

GT population size	R_{clone}	R_{in}, Pr = 0.001	R_{out}, Pt = 0.999
100,000	0.016	462 (0.4620%)	25 (0.0250%)
200,000	0.013	770 (0.3850%)	34 (0.0170%)
300,000	0.014	613 (0.2043%)	24 (0.0080%)
400,000	0.012	553 (0.1383%)	53 (0.0133%)
500,000	0.016	529 (0.1058%)	19 (0.0038%)
600,000	0.009	254 (0.0423%)	45 (0.0075%)
700,000	0.008	534 (0.0763%)	24 (0.0034%)
800,000	0.011	581 (0.0726%)	13 (0.0016%)
900,000	0.012	518 (0.0576%)	33 (0.0037%)
1,000,000	0.012	30 (0.0030%)	45 (0.0045%)
2,000,000	0.01	78 (0.0039%)	29 (0.0015%)

Simulating attack - Matching to Real Patients



Distance threshold

3 Synthetic NN are close to the GT patient A and 2 Synthetic patients are close to GT patient B within the distance radius.

GT Attributes	Num Of GT Outliers	Num Of GT Patients in 10000 Sample	Num of *Single Synth* NN	Proportion of GT Outliers with Single Synth NN
age, smoking, region, gender, ethnicity, ckidney	396.3	6248	113.8	1.82%
age, smoking, region, gender, ethnicity, bmi, sbps, ckidney	377	6250	77.7	1.24%
age, smoking, region, gender, ethnicity, bmi, choleratio, sbp, sbps, ckidney	363.1	6289	48.22	0.76%
age, smoking, region, gender, ethnicity, bmi, sbps, ckidney, sle, atyantip, type1, streroid	363.4	6180	43.5	0.70%

Example output:

Total Synthetic dataset

Part of Ground truth dataset

"age", "bmi", "sbps , "smoking", "region", "gender"

analysis

ID	Type	age	bmi	sbps	smok	gen	reg
1	GT	88	23.9	16.57	3	F	West Mid
2	Syn	86	23.3	16.65	3	F	West Mid

Add the rest of the Ground truth dataset to check how the other attributes' values compare

+

ID	Type	stroke	steroid	migr	type1	type2	choleratio
1	GT	1	0	0	0	0	2.02
2	Syn	0	0	0	0	0	3.4

Brunel University London

GOV.UK

Search

Departments Worldwide How government works Get involved
Consultations Statistics News and communications

Coronavirus (COVID-19) | Guidance and support

Home > Medical devices regulation and safety

Press release

New synthetic datasets to assist COVID-19 and cardiovascular research

Ground-breaking innovation to support medical technologies

Published 29 July 2020
From: Medicines and Healthcare products Regulatory Agency

Related content

Drug Safety Update: monthly PDF newsletter

Safe use of emollient skin creams to treat dry skin conditions

CBMS 2019

ECMLPKDD
Würzburg 16.–20.09.2019

npj | digital medicine

Computational Intelligence
AN INTERNATIONAL JOURNAL

Brunel University London

Data Privacy vs Utility

- Synthetic data appears useful in validating models
- Classes of generative models offer promise
  - Large scale datasets take up resources
  - Flexible in handling data types (continuous, categorical, time etc.)
- But always a trade-off between utility and privacy:
  - Aggregated data vs personalised interventions
  - Rare disease
  - Marginalising over many features
- Dealing with bias is important …

Thanks

- Barbara Draghi
- Namir Oues
- Juan de Benedetti
- Zhenchen Wang
- Ylenia Rotalinti
- Puja Myles





IDA Research

NHSx

CPRD

Medicines & Healthcare products Regulatory Agency

Brunel University London

Reproducibility in population-wide emergency risk prediction

Scottish Patients At Risk of Re-admission and Admission
'SPARRA', version 4

James Liley
31/03/2021

Overview

1. The SPARRA model

a). Motivation and aim

b). Model overview

c). Findings

2. Data safe havens and challenges

a) ‘Hard’ obstacles to reproducibility

b) ‘Soft’ obstacles to reproducibility

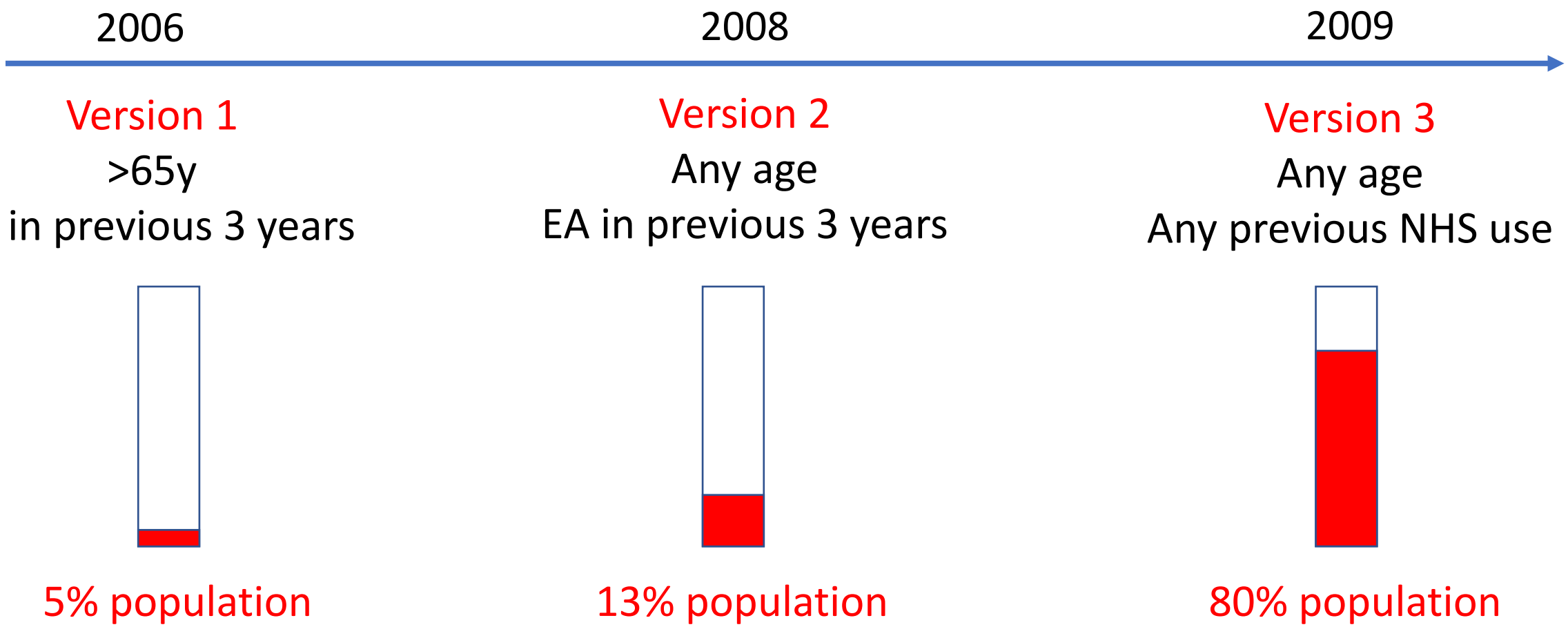
3. Practices to improve reproducibility

Public Health Scotland

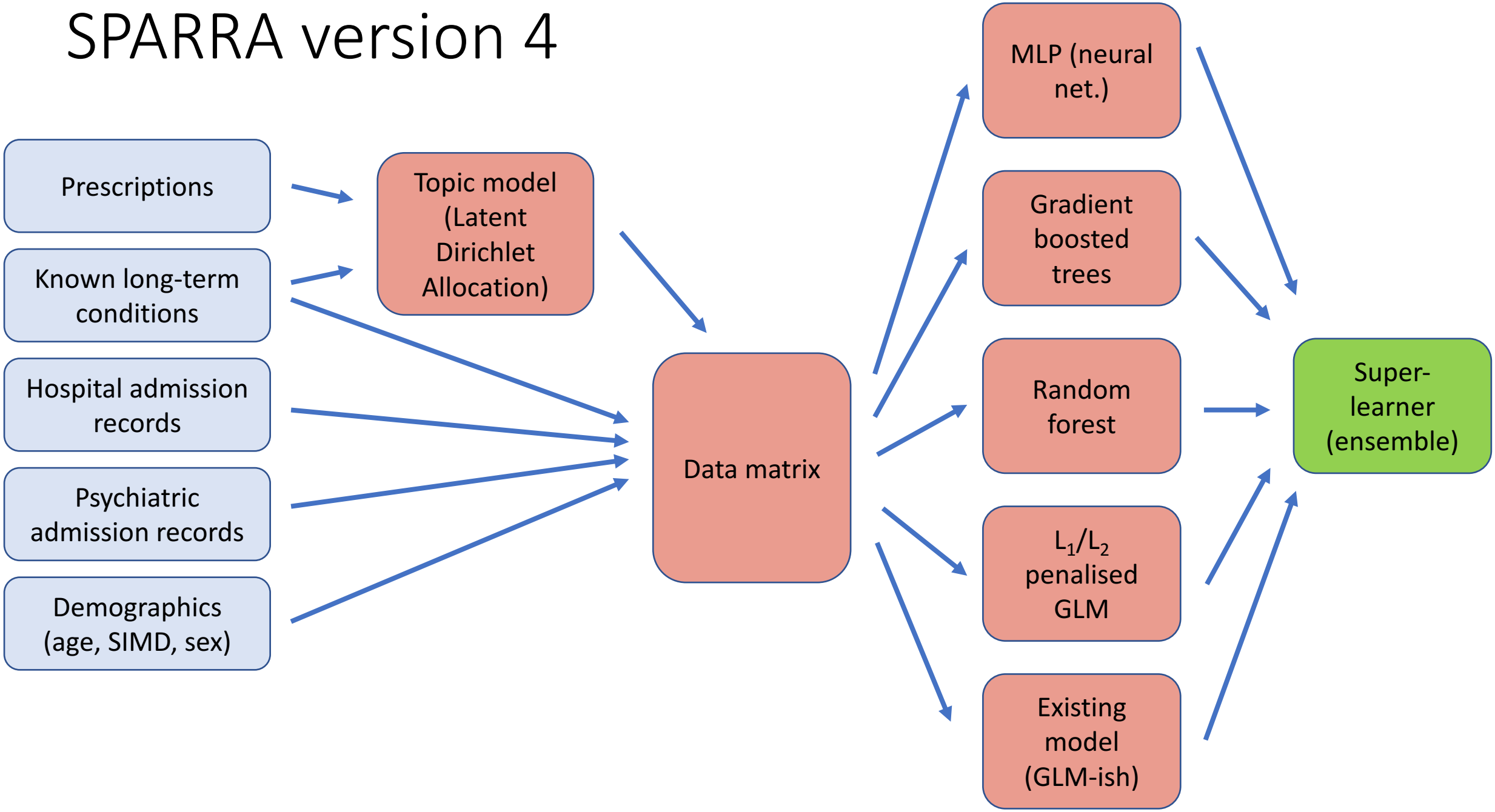
‘The NHS should work with other public services and with patients and carers to provide continuous, anticipatory care to ensure that, as far as possible, health care crises are prevented from happening’

A National Framework for Service Change in the NHS in Scotland, NHS Scotland, 2005

SPARRA: estimate probability that a patient will have an emergency hospital admission in the coming year on the basis of records of previous interactions with the healthcare system.



SPARRA version 4



Updating considerations

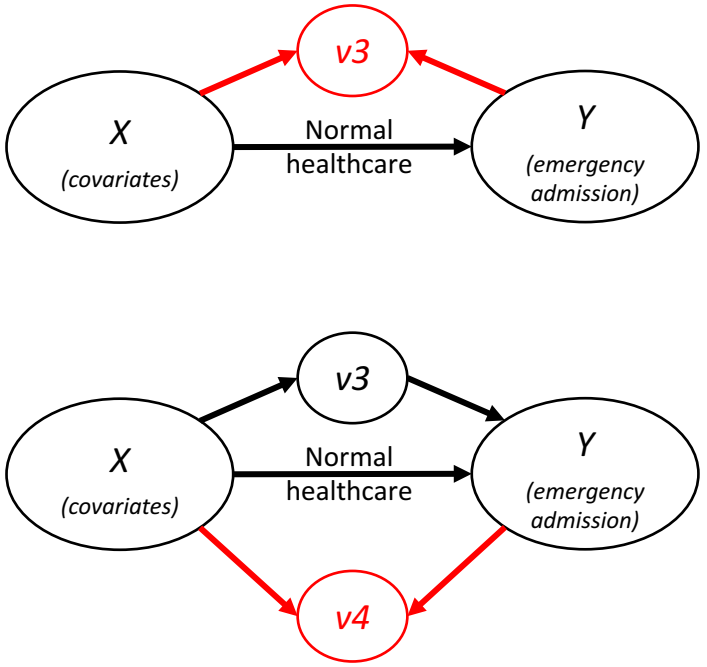
Does an existing model overestimate risk because the model is out-of-date, or because it is driving intervention?

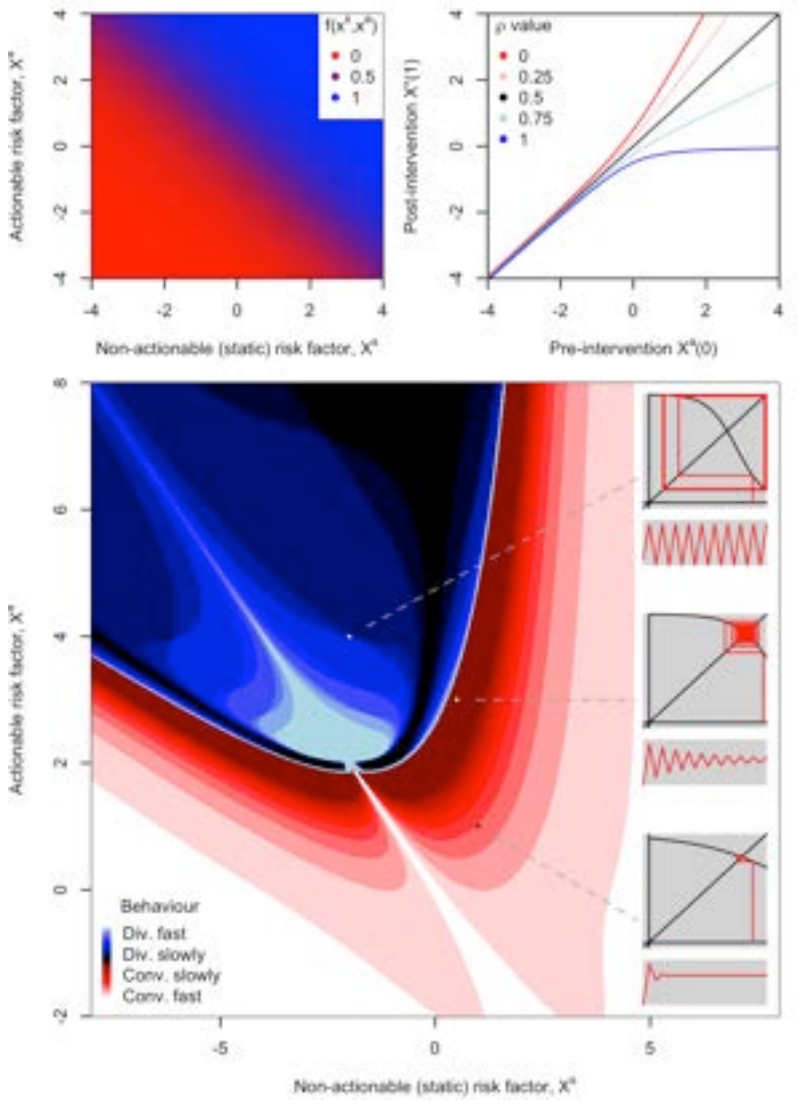
Replacing an existing score means the new score is used in a different system to the one in which it was fitted.

‘Naive’ updating leads to problems

Proposed solution: act on v3, *then* act on v4

Approximate by deploying maximum of v3 and v4 scores





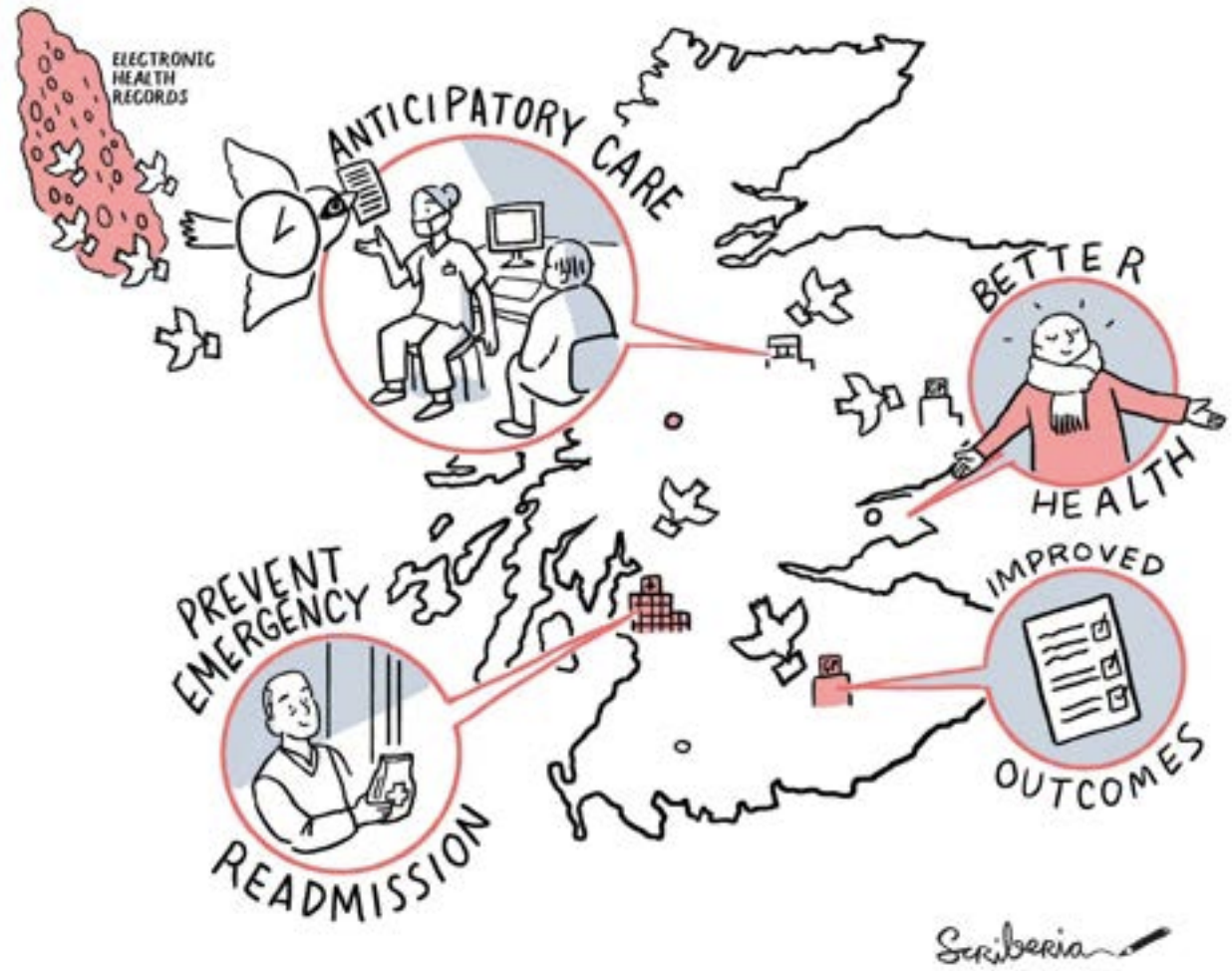
Liley, Emerson, Mateen, Vallejos, Aslett, Vollmer. AISTATS 2021

Deployment

SPARRA scores are deployed to GPs Scotland-wide, and may be used to guide intervention or public health action.

We need to make sure that the score going to GPs is the same as the score we are writing about in the paper.

This project was paid for by public funds and influences Scotland's healthcare strategies. It should be as open as possible.



Difficulties with reproducibility

Data safe havens (DSH) – 'Five Safes'

1. Restriction the personnel accessing the DSH
2. Clear project delineations
3. No communications to/from DSH
4. Ensure individuals in data are not identifiable.
5. Disclosure control on data/code exports

Other considerations

1. Population-wide scale
2. Complex data to results pipeline
3. Complex legal protections

Implications (DSH)

Raw data cannot be removed from DSH (4)
Other researchers cannot access DSH to perform checks (1)
No internet access on DSH (3)
Highly dependent on inflexible machine architecture
All code must be readable if it is to be exported (3,5)
Work progresses slowly (1-5)
Data export is slow (3,5)

Implications (other)

Generally not feasible to re-run entire pipeline whenever an error is discovered (1)
Many people to keep 'in the loop' (2)
SPARRA v3 algorithm/coefficients embargoed (3)
Difficult to attain permission to publish code (3)

https://blog.ons.gov.uk/2017/01/27/the-five-safes-data-privacy-at-ons/

Hard obstacle 1: full reproducibility is impossible

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Gold standard reproducibility: verify that algorithm gives the same results given the same input

Gold standard impractical as
- raw data cannot be exported
- model parameters cannot (generally) be exported
- verifying researchers cannot access DSH.

Technically can be reproduced on other safe havens; this is not usually practical.

Most important: works in deployment machine

Management

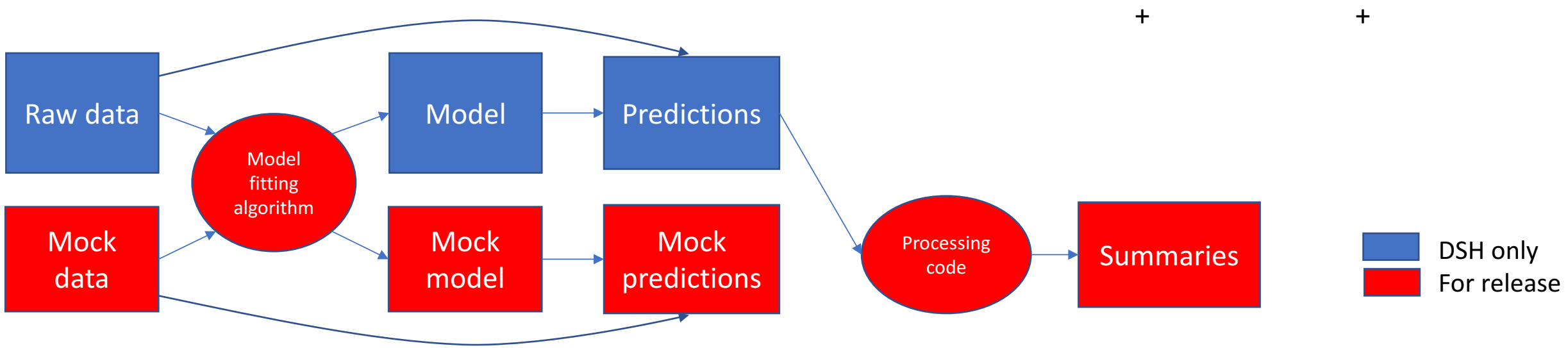
Mockup dataset with similar global properties to real data

Export of full model fitting algorithm, verified on mock data

Export of all processing code

Complete unit test for algorithm

Extensive export of summaries: figure + data + code



Hard obstacle 2: computation

Software considerations

1. No internet access on DSH; no github/stackExchange

2. Software updates difficult/slow.

Hardware considerations

1. Limited computational capacity

2. Rigid machine structure

3. No control over machine

Management (software)

1. Code often must be written from scratch. Protocol:
 - all code commented and written in a consistent style
 - all code independently checked and ran by another team member

2. Containerised deployment (Docker, MS Azure)

Management (hardware)

1. Careful algorithmic design; garbage collection, deleting data after use

2. Develop code to run on one machine or two

3. Multiple save checkpoints.

Soft obstacle 1: increased personnel

Personnel difficulties

1. Different teams of people generating and curating raw data (Public Health Scotland), analysing data and fitting models (us), and deploying the model (PHS)

2. No control over computation: crashes, data removal, etc. Pipelines cannot usually be ran end-to-end.

Management

1. Close liaison with PHS
 - Weekly/biweekly meetings
 - ‘Knowledge transfer’
 - Multidisciplinary teams

2. We cannot fix crashes/memory issues ourselves; so
 - pre-empt and avoid issues
 - pipelines split into independent sub-pipelines.

Soft obstacle 2: everything is more difficult!

General difficulties

1. Getting data on and off the DSH is very slow
2. Fixing any problem with computation is slow
3. Sharing preliminary data with colleagues is difficult
4. Checking raw data is difficult
5. Beholden to internet connectivity
6. Beholden to general workload of Public Health Scotland
7. Cannot copy-paste into data safe haven
8. No '#' key on data safe haven
9. Need to test new code extensively
10. No root access on data safe haven
11. Cannot delete temporary R files, so accidentally crashed sessions irreversibly use up memory
12. Vulnerable to nationwide security breaches
12. Lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorum ipsum
12. Lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorum ipsum lorem ipsum lorem ipsum
12. Lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorum ipsum lorem ipsum
12. Lorem ipsum lorem ipsum lorem ipsum lorem ipsum lorum ipsum

Management

Put reproducibility concerns at the forefront of workflow

Summary

Data safe haven work at population scale leads to difficulties in ensuring reproducibility, some of which are unavoidable.

Mock-up data is useful for verification of algorithms

For real-world applications, many people of different backgrounds need to be able to reproduce the results and use the methods. Ongoing ‘knowledge transfer’ and communication are important.

Working in data safe havens is slower and more difficult than ML researchers may be accustomed to. Additional time has to be allowed for completion of projects so that reproducibility can be maintained.

Acknowledgments

University of Edinburgh/Alan Turing Institute (ATI)

Catalina Vallejos

Vallejos and Sanguinetti groups

Turing Institute

Gergo Bohner

Nathan Cunningham

Franz Kiraly

Ioanna Manolopoulou

Bilal Mateen

Katrina Payne

Sebastian Vollmer

University of Durham/Turing Institute

Louis Aslett

Sam Emerson

Public Health Scotland

Katie Borland

David Carr

Scott Heald

Sam Oduro

Jill Ireland

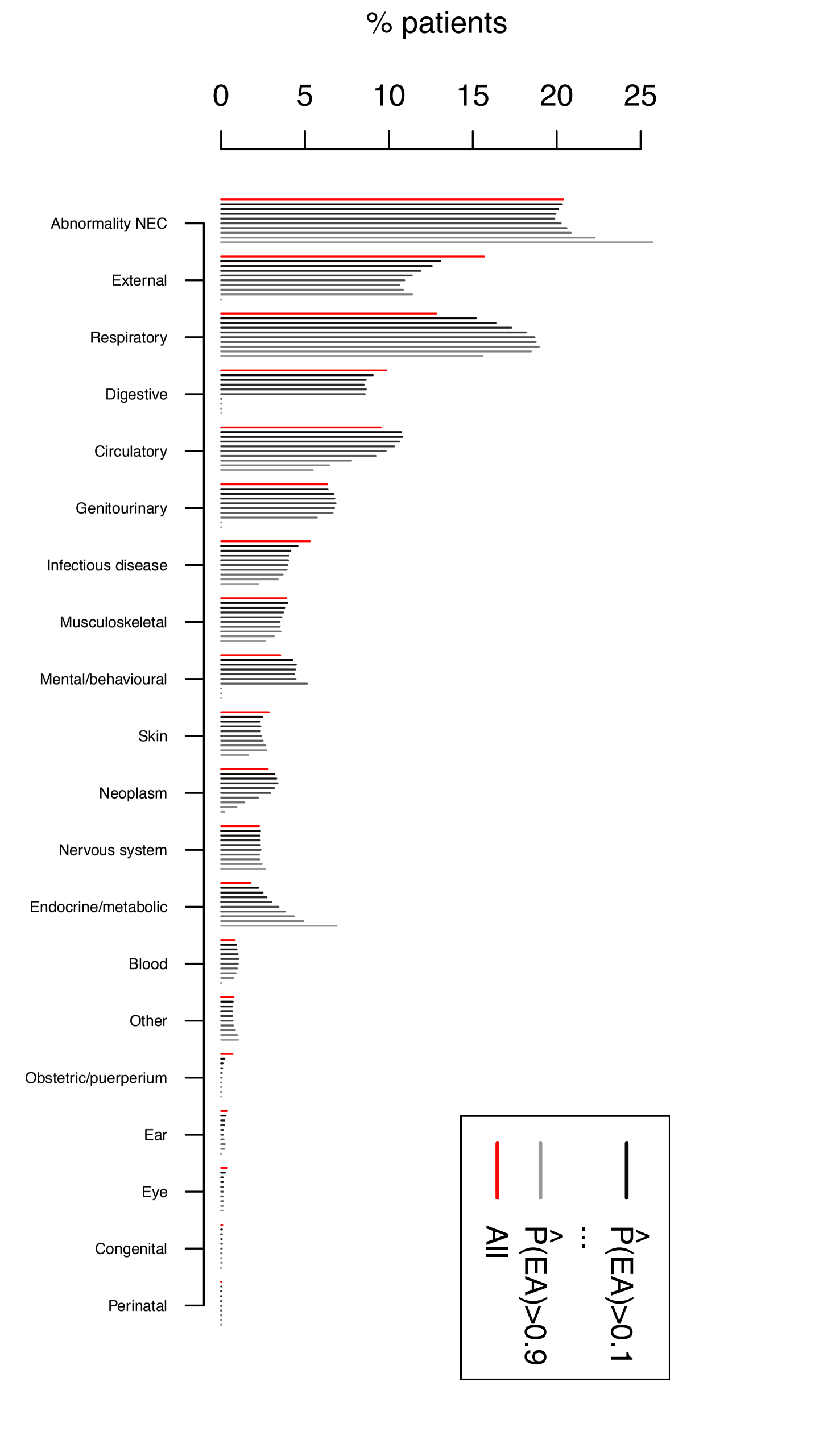
Keith Moffat

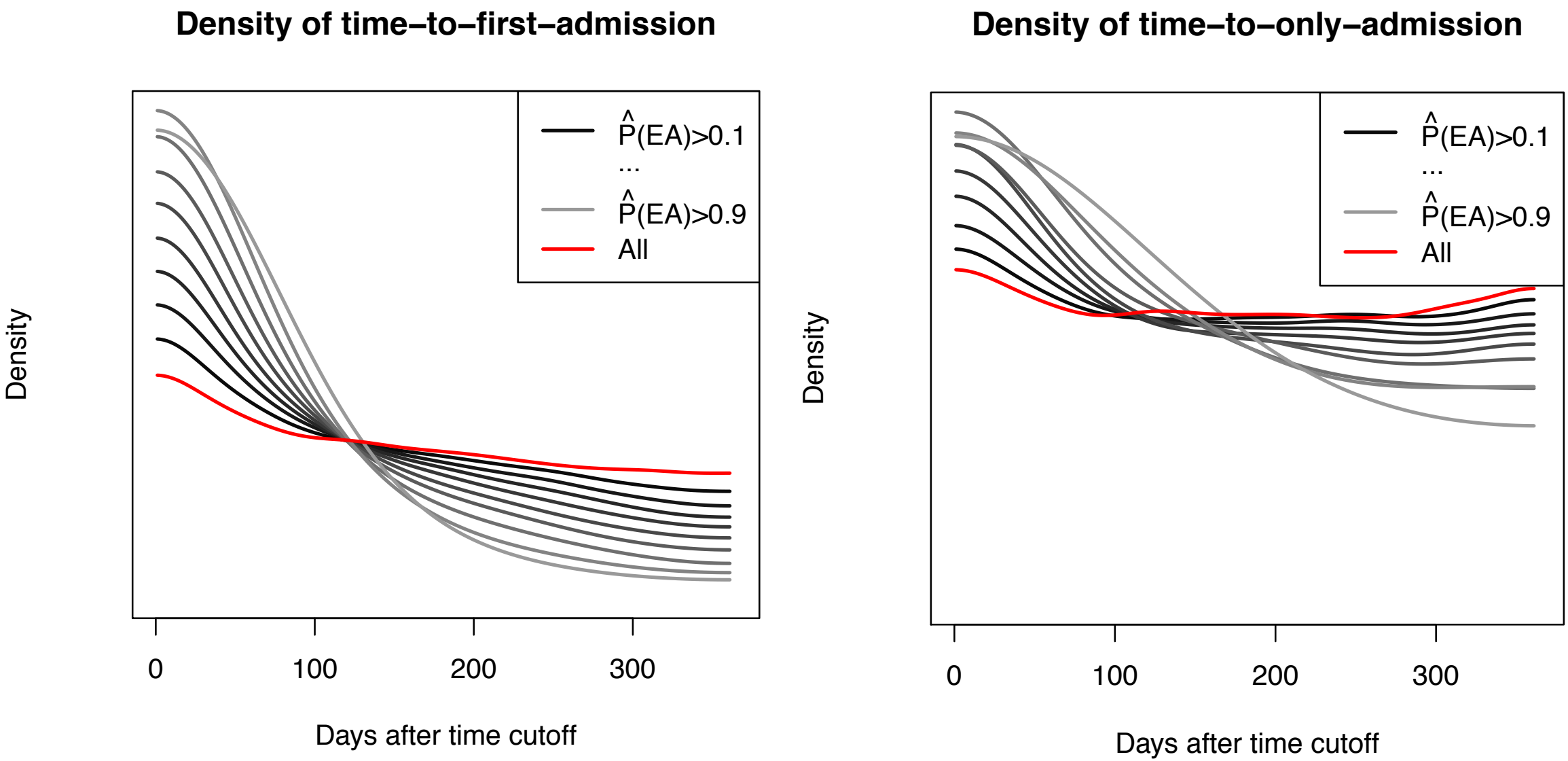
Rachel Porteous

Stephen Riddell

Beth Bruce (eDRIS)

Lizzie Nicholson (eDRIS)


% patients
0
5
10
15
20
25
Abnormality NEC
External
Respiratory
Digestive
Circulatory
Genitourinary
Infectious disease
Musculoskeletal
Mental/behavioural
Skin
Neoplasm
Nervous system
Endocrine/metabolic
Blood
Other
Obstetric/puerperium
Ear
Eye
Congenital
Perinatal
All
P^(EA)>0.9
P^(EA)>0.1


Density of time–to–first–admission
Density
P̂(EA)>0.1
...
P̂(EA)>0.9
All
0
100
200
300
Days after time cutoff
Density of time–to–only–admission
Density
P̂(EA)>0.1
...
P̂(EA)>0.9
All
0
100
200
300
Days after time cutoff

Tools and open datasets to support training activities around reproducible machine learning: An activity monitoring use-case

Work Package II

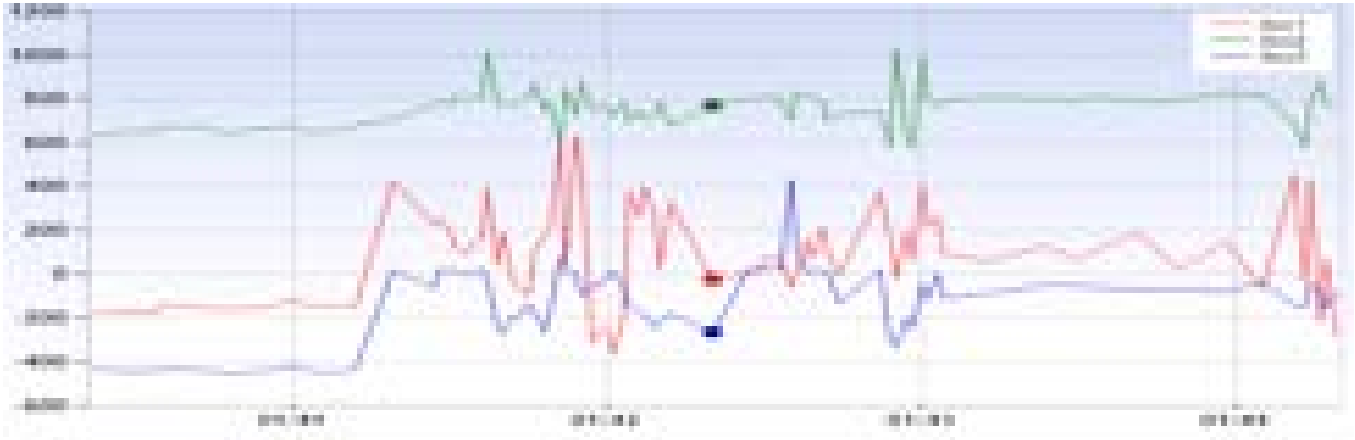
Shing Chan and Aiden Doherty

Big Data Institute & Nuffield Department of Population Health
University of Oxford

March 31, 2021

Accelerometers

Measures movement (more specifically, acceleration)



Applications

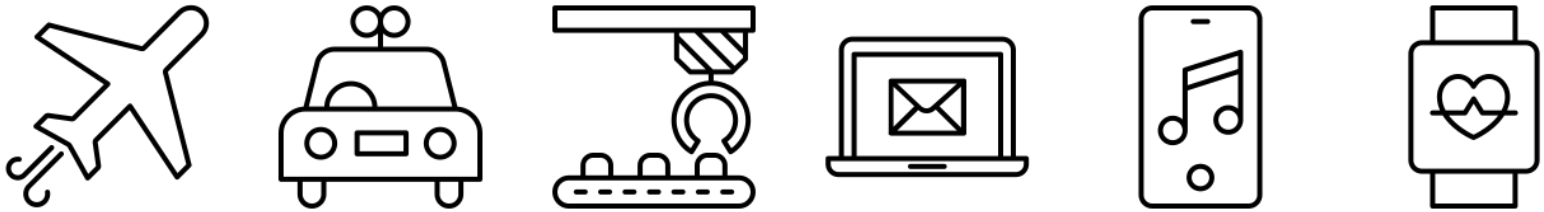


Image Credits: Made, Made x Made Icons, The Noun Project

Wearable technology

Image Credits: Macrovector, Shutterstock; Business Insider; Xiaomi

New opportunities for health care

Activity information remains under-exploited in health research

Meassuring activity is challenging

- How to measure

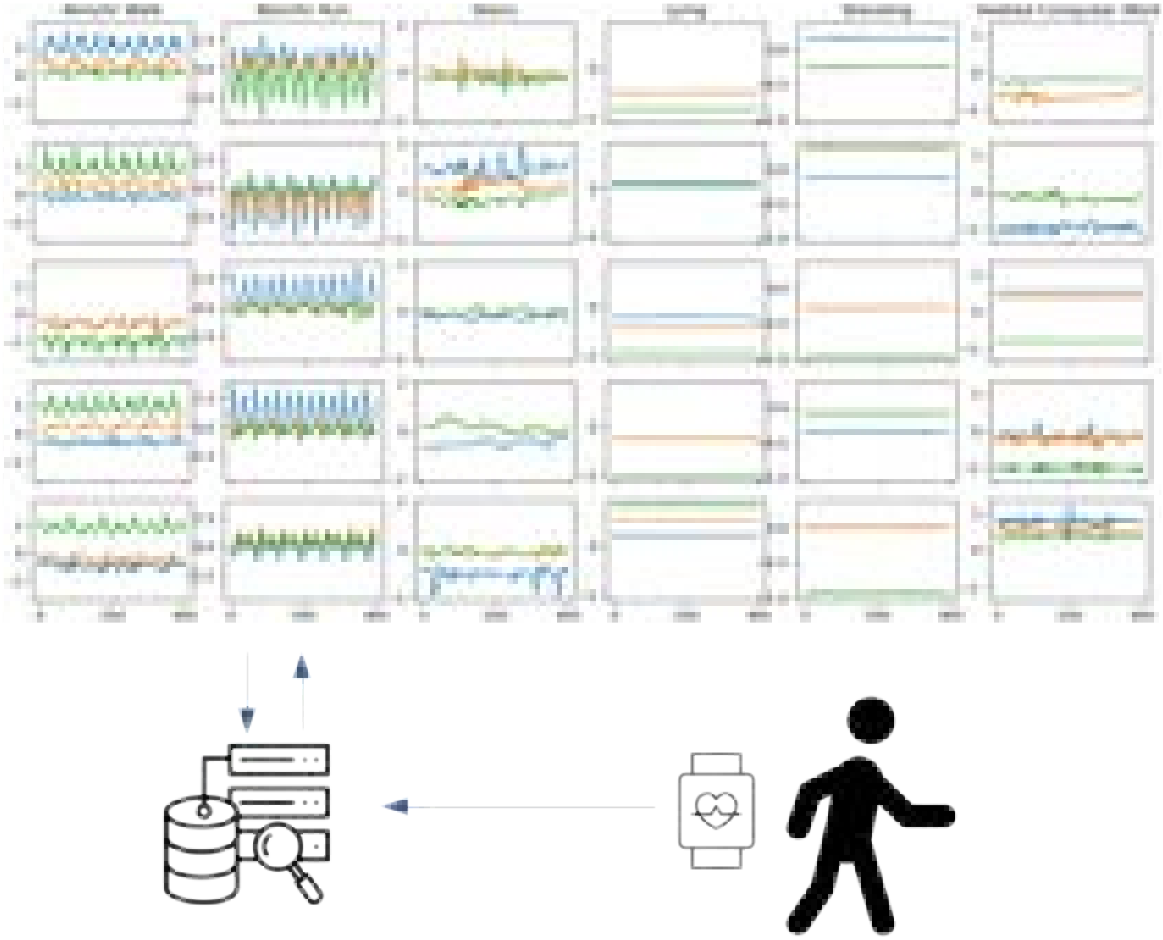
- What to measure

New directions: Wearables

- Objective, hi-res, free-living
- Efficient/scalable (e.g. UKBiobank)

Image Credits: ProSymbols, unlimicon, purplestudio, Yeoul Kwon, The Noun Project; Garmin

Activity recognition



Large and representative (i.e. free-living) data is key

Image Credits: Made x Made, ProSymbols, The Noun Project

Existing openly available wrist-worn accelerometer datasets

Most are small and collected in artificial settings

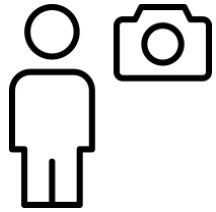
Dataset	Size	Annotations	Free-living
PAMAP2	9 ppl $\times$ 1 hrs	18 activities (walk, run, sit, ...)	No
MHEALTH	10 ppl $\times$ 15 mins	12 activities (walk, run, sit, ...)	No
ADL	16 ppl $\times$ 1 hrs	14 activities (brush teeth, comb hair, drink glass, ...)	No
UC Berkley WARD	20 ppl $\times$ 1 hrs	13 activities (walk, sit, stand, jog, stairs, ...)	No
CMU-MMAC	43 ppl $\times$ 5 mins	kitchen/cooking activities (pizza, salad, brownie, ...)	No
Opportunity	12 ppl $\times$ 1 hrs	morning activities (prepare breakfast, cleanup, ...)	Semi
Capture-24	**158 ppl $\times$ 24 hrs**	**$>$200 free-living annotations**	**Yes**

Capture-24: An activity monitoring dataset

- Wrist-worn accelerometer (Axivity3, triaxial, 100Hz)
- 158 subjects $\times$ 24 hours
- Oxford area, 2015
- Free-living
- More than 200 different annotations
- Compatible with the UKBiobank

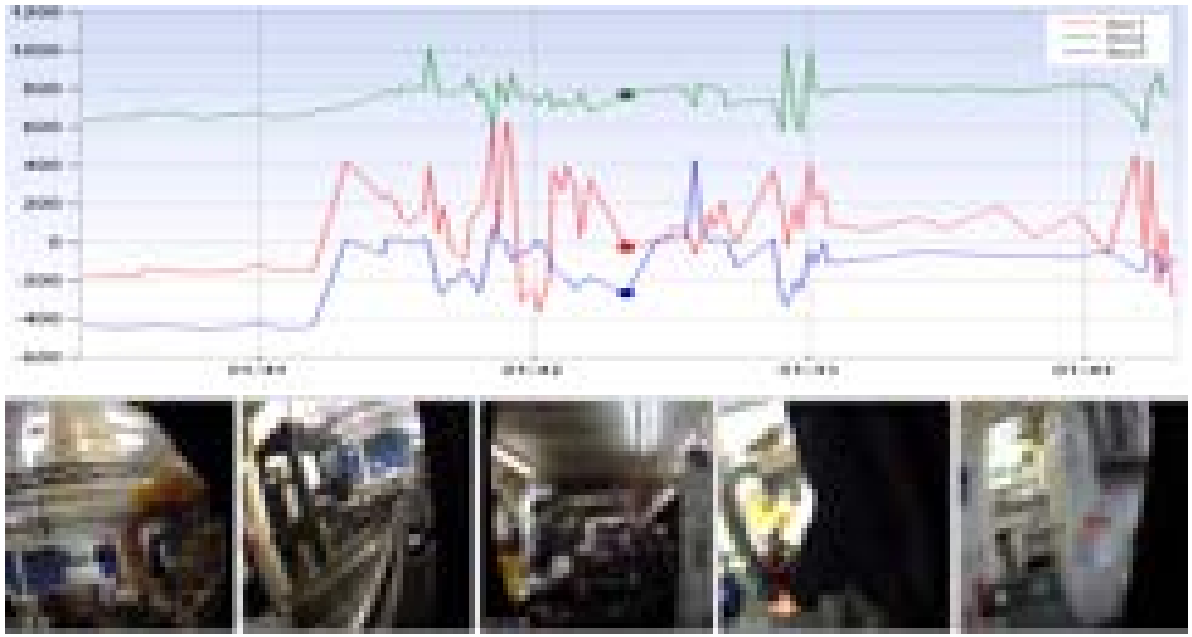
Capture-24: Collection

Accelerometer watch



Body camera

Sleep diary

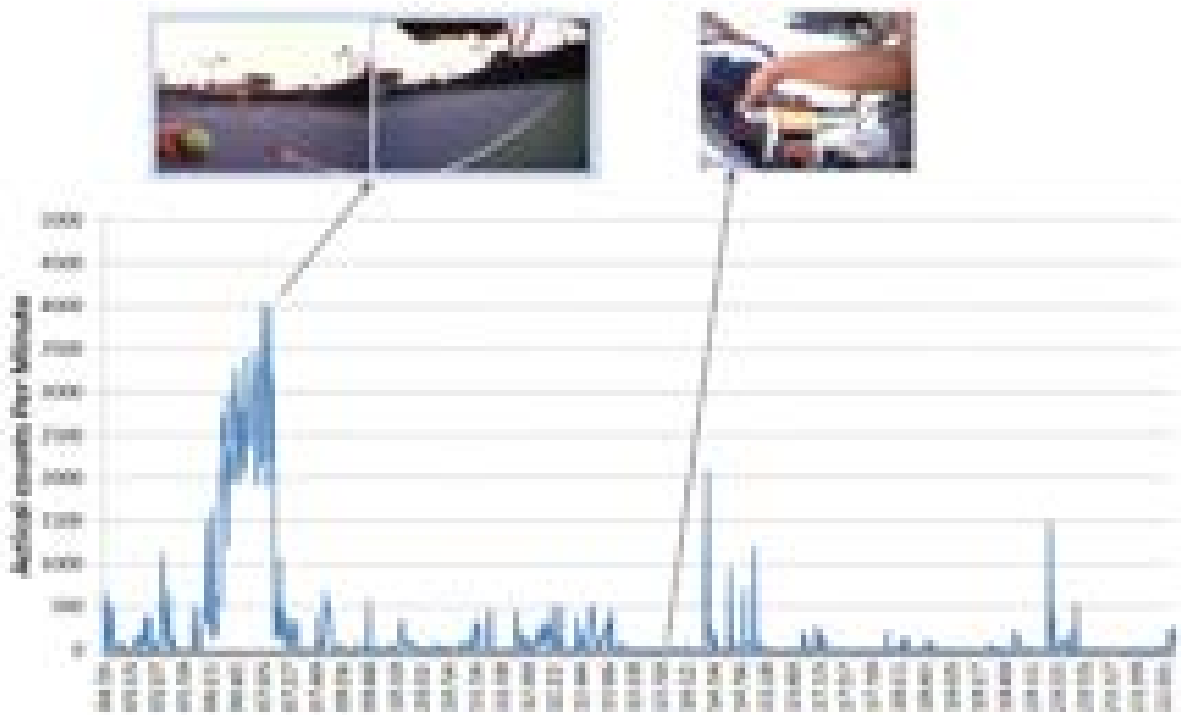


Accelerometer + Body camera

Image Credits: Made x Made, Andrejs Kirma, IconPai, The Noun Project

Capture-24: Annotation

Accelerometer timestamps marked using camera images



Capture-24: Annotations

More than 200 different annotations

1. home activity;miscellaneous;sitting;7010 lying and watching television with TV on as the primary activity
2. leisure;recreation;outdoor;5175 walking/running playing with child(ren)
3. home activity;miscellaneous;sitting;7021 sitting without observable activities
4. leisure;miscellaneous;walking;17133 walking upstairs
5. leisure;miscellaneous;walking;17070 descending stairs
6. home activity;miscellaneous;sitting;9060 sitting reading or using a mobile phone/smartphone/tablet or talking on the phone/computer (skype chatting)
7. home activity;miscellaneous;standing;5146 standing packing/unpaking household items ocasional lifting
8. leisure;miscellaneous;21070 (generic) walking/standing combination indoor
9. leisure;miscellaneous;21016 sitting child care only active periods
10. leisure;miscellaneous;21010 sitting non-desk work (with or without eating at the same time)
11. leisure;miscellaneous;17031 loading /unloading a car implied walking

...

Capture-24: Data Cleaning and Anonymisation

- Camera's low temporal resolution
- Sleep diary checked against accelerometry

- Rare annotations omitted or modified
- Study time perturbed (random offset), dates omitted
- Ages converted into age bands

Image credits: QualityIcons, The Noun Project

Capture-24: Successful use-cases

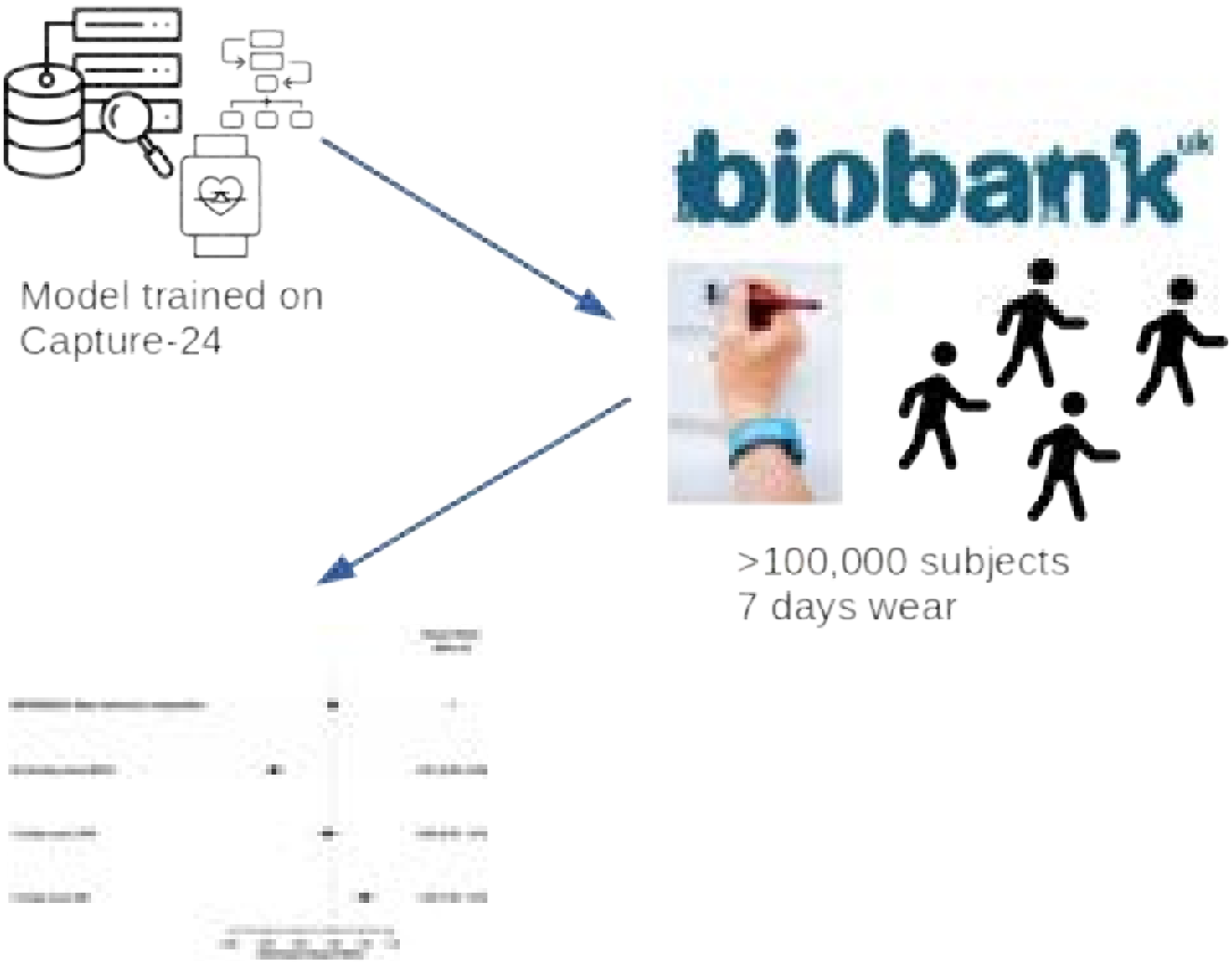
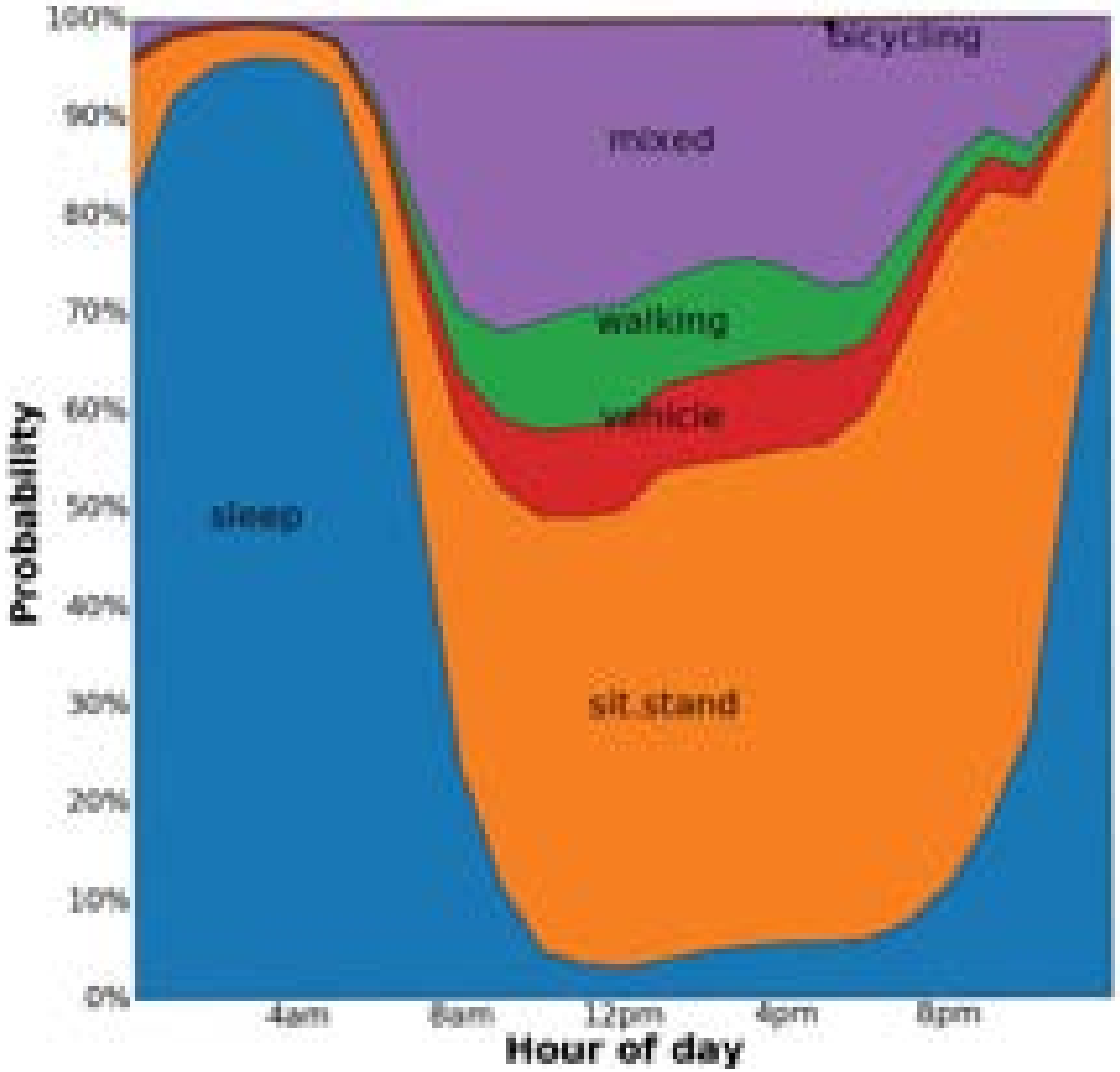


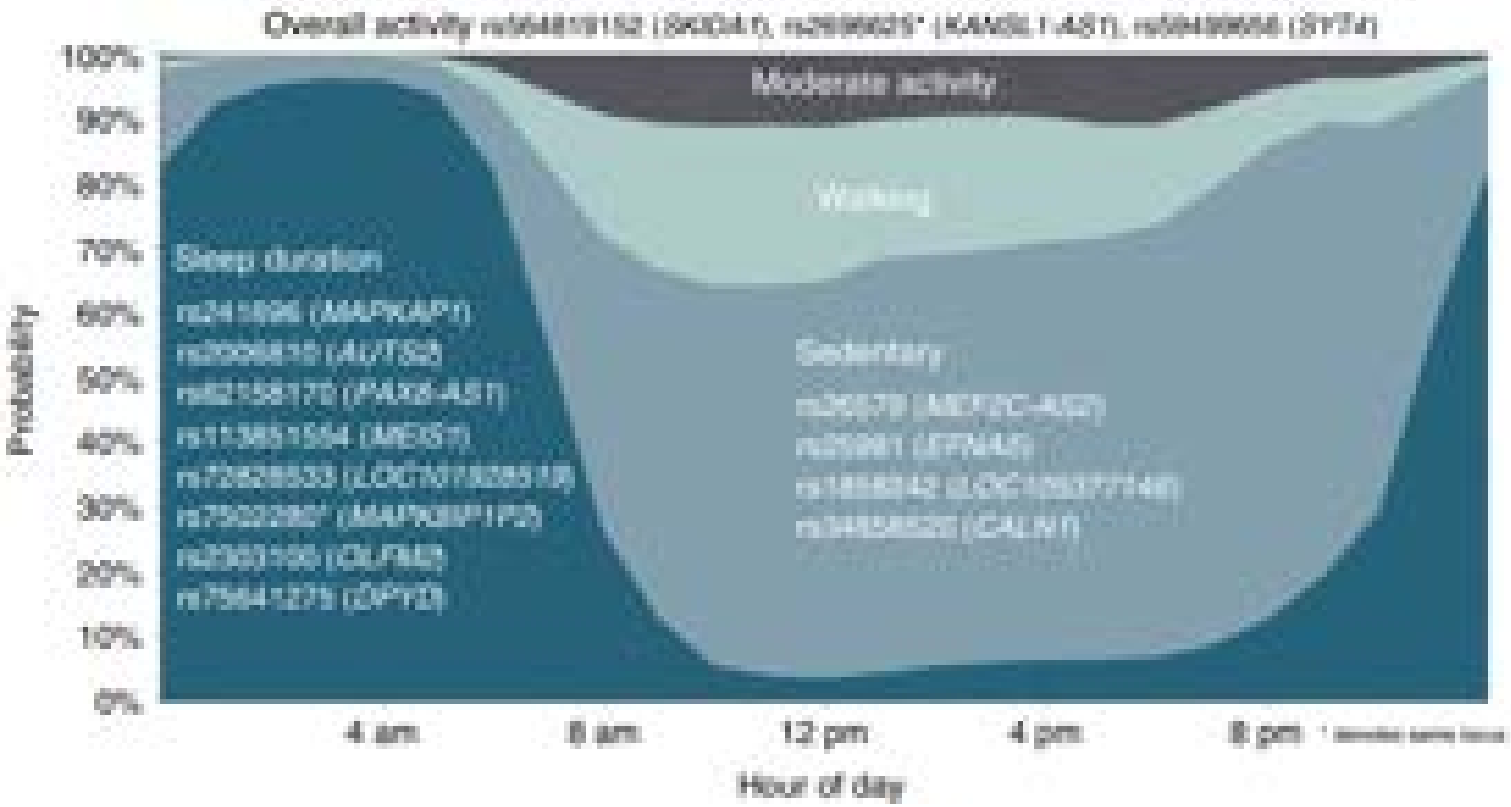
Image credits: Made x Made, ProSymbols, QualityIcons, The Noun Project

Capture-24: Successful use-cases



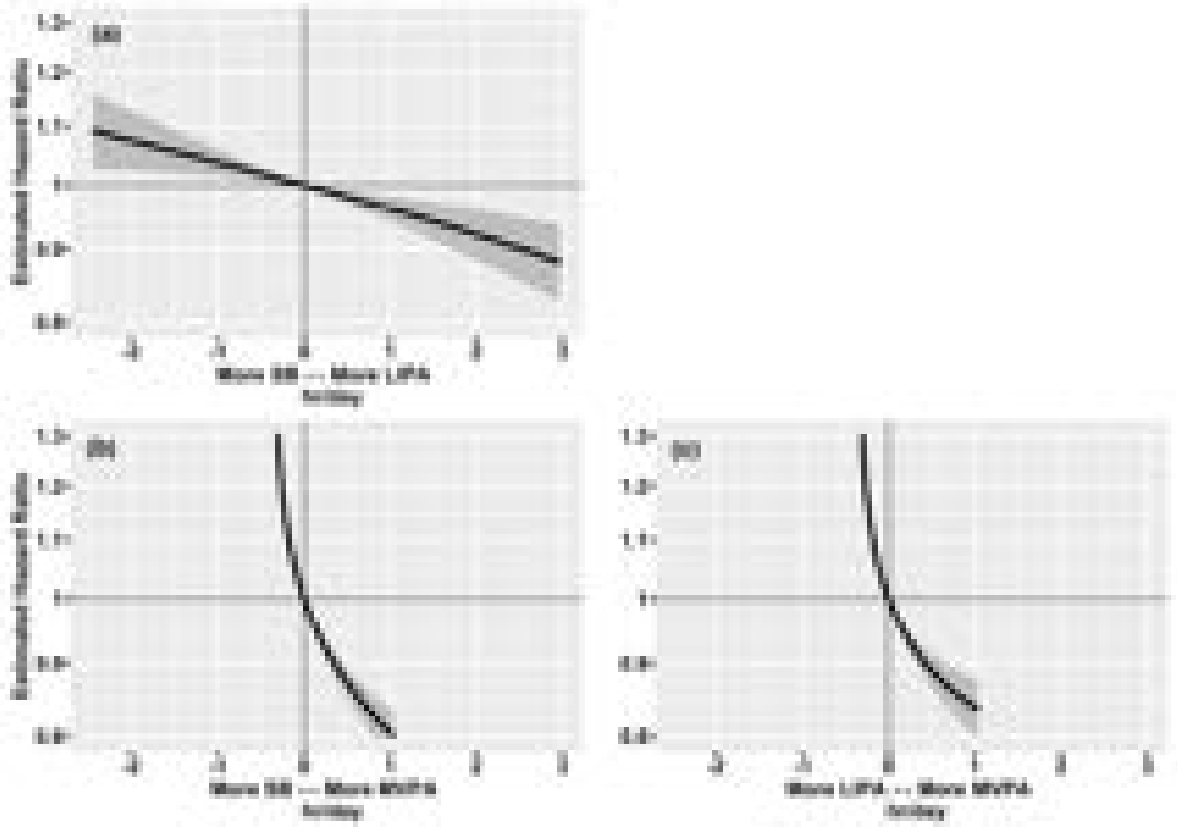
Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants – Willetts et al. 2018

Capture-24: Successful use-cases



GWAS identifies 14 loci for device-measured physical activity and sleep duration – Doherty et al. 2018

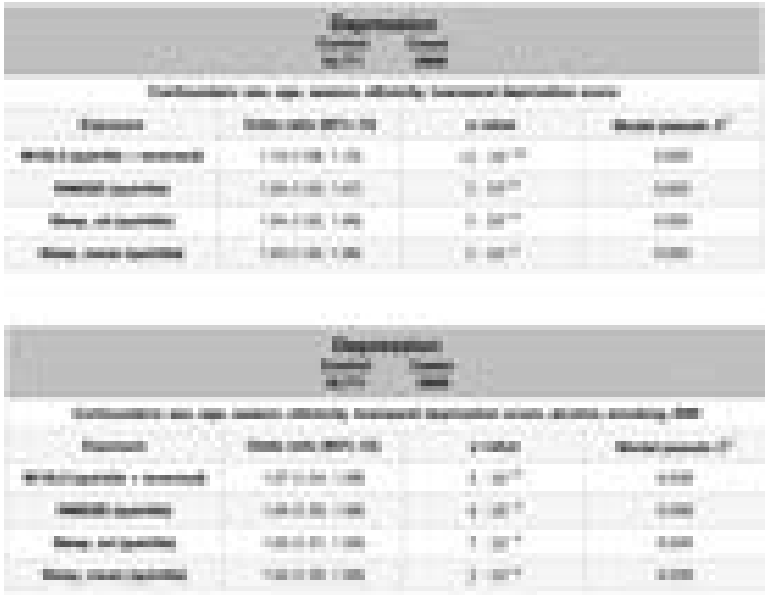
Capture-24: Successful use-cases



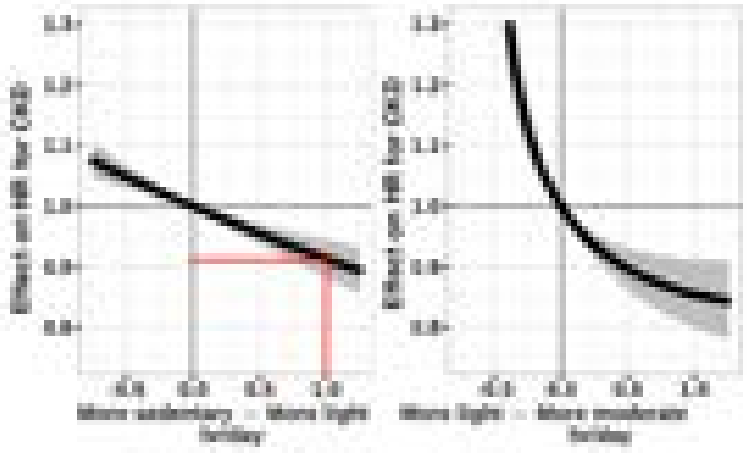
Reallocating time from device-measured sleep, sedentary behaviour or light physical activity to moderate-to-vigorous physical activity is associated with lower cardiovascular disease risk – Walmsley et al. 2020

Capture-24: Successful use-cases

Oxford Intensive 2-weeks Workshop Results

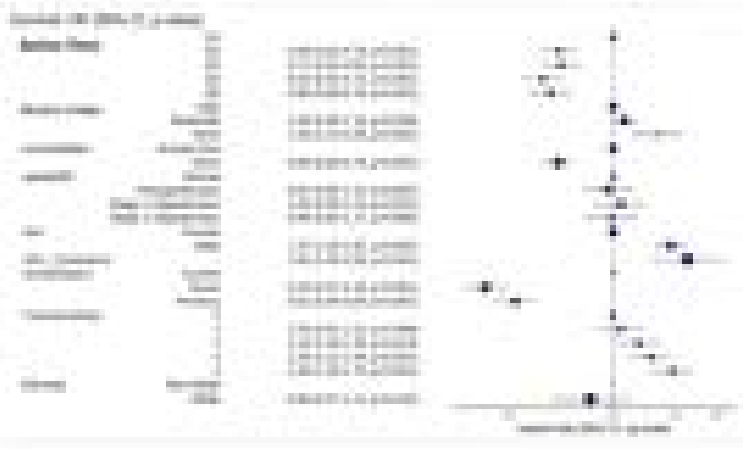


Mental disorders



Chronic kidney disease

Type II Diabetes



All-cause mortality

Capture-24: Resources & Links

Resources

- Accelerometer processing & analysis tool: `https://github.com/activityMonitoring/biobankAccelerometerAnalysis`
- Workshop materials: Activity recognition on the Capture-24: `https://github.com/activityMonitoring/week1_cdt_data_challenge`

Contact Us

aiden.doherty@ndph.ox.ac.uk
shing.chan@ndph.ox.ac.uk

NHS Digital Academy

Opportunities to embed reproducible machine learning across national training programmes

Dr Athanasios Tsanas ('Thanasis') atsanas@ed.ac.uk
Associate Professor in Data Science
Usher Institute, Medical School
University of Edinburgh

DARTH

THE UNIVERSITY of EDINBURGH | usher institute

NHS Digital Academy

Background in a nutshell

- Wachter review for UK healthcare
- 2017 Secretary of State commitment
- Train 300 aspiring NHS leaders
- Phase1: 4/2018 – 9/2021
- Programme run by Imperial + Edinburgh

NHS DA programme

- Leadership



- 6 Modules
- Module 5: Actionable Data Analytics and Clinical Decision Support


Latest developments

HDRUK
Health Data Research UK

- NHS DA renewal (5 years)?
- HDRUK new partner
- Module 5 material widely accessible through HDRUK
- HDRUK teaching platform

Professor Sophie Staniszewska

The contribution of Public Voice in Machine Learning in Health Care

WARWICK
MEDICAL SCHOOL

Patient and public involvement in research

By public involvement we mean research being carried out ‘with’ or ‘by’ members of the public rather than ‘to’, ‘about’ or ‘for’ them as defined by NIHR INVOLVE

The impact of public involvement in NIHR health and social care research is defined as:

“The changes, benefits and learning gained from the insights and experiences of patients, carers and the public when working in partnership with researchers and others involved in NIHR initiatives”

(NIHR INVOLVE 2019)

Why involve the public in research?

Makes research more relevant, focused on questions of importance to patients and the public

Enhances quality of research eg. ensuring a trial measures the right outcomes

A moral/ethical imperative

"Nothing about me without me"

Democratic accountability to the taxpayers

ML: Fairness, accountability and transparency

Staniszewska S, Brett J, Simera I, Seers, K, Mockford, C, Goodlad S Altman DG Moher D Barber R Denegri S Entwistle A Littlejohns P Morris C Suleman R Thomas V Tysall C (2017) GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. BMJ 2017; 358 doi: https://doi.org/10.1136/bmj.j3453

Co-production

Sharing power

Including all perspectives and skills

Respecting and valuing the knowledge of all

Reciprocity

Build and maintain relationships

Joint understanding and consensus and clarity over roles and responsibilities

Hickey, G., Brearley, S.,Coldham, T., Denegri, S., Green, G., Staniszewska, S., Tembo, D., Torok, K., and Turner, K. (2018) *Guidance on co-producing a research project.* Southampton: INVOLVE.

Progress so far for our project

Advertised for public contributors with responses from 50 people

Interviewed potential participants in March 2021 with academic lead from Warwick, PPI lead from Oxford and public contributor from Warwick

Appointed a panel of 12 people with a range of experience and perspectives who are keen to work with the research team

Strategic decision to recruit public contributors to maximise the diversity of public voice

Co-production in action:

The MEMVIE Study – An example of the potential of co-production in a complex area

Staniszewska, S., Hill, E.M., Grant, R. *et al.* Developing a Framework for Public Involvement in Mathematical and Economic Modelling: Bringing New Dynamism to Vaccination Policy Recommendations. *Patient* (2021). https://doi.org/10.1007/s40271-020-00476-x

WARWICK
MEDICAL SCHOOL

Mathematical and Economic Modelling for Vaccination and Immunisation Evaluation (MEMVIE)

- Public involvement less common in complex areas such as mathematical and economic modelling
- Modelling is important – provides decision makers with best available evidence to reach a decision
- Involving patients and public – enhance models, improve confidence and accelerate decision making
- Key outcome of MEMVIE: Identified a framework to identify the nature and type of public involvement to guide future models, identifies values, identifies conditions for implementation and provides a detailed long form version that considers involvement at each stage of the modelling process and a short form summary

What did it involve?

- 21 meetings over 5 years
- Each lasted 2-3 h. Email contact in-between with the group commenting on documents
- Deliberative knowledge space and Think Aloud techniques encouraged ideas and thoughts to emerge
- Public contributors were able to challenge the data, the basis for the collection of data and the interpretation of that data, thinking outside of the box in a safe space where modellers could rework their thinking
- The meetings enabled thematic development over time as the Reference Group contributors worked with the academic contributors on continuous iterations of the emerging framework

Academic contributor

WARWICK
MEDICAL SCHOOL

“When I joined midway through the duration of the MEMVIE project, I had not had any previous exposure to public involvement as part of the research process. I found it extremely beneficial to have an additional forum to describe our modelling process, discuss model assumptions and examine data. ***From my perspective, being given the opportunity to convey the work to public members through reasoned discourse, ensured justification of modelling aspects, aiding model integrity and validity.*** *In addition, public involvement generated broader discussion surrounding data curation and data collection (such as questionnaire content), producing recommendations that can be used to inform future developments."*

Conclusion

Public voice should be an integral part of how we think about Machine Learning

Consider moving from the passive (is this ok with the public?) to co-producing thinking with the public (which concepts to guide us, how do we create new knowledge together, can we develop frameworks for voice embedded in our methods)

Co-production holds great promise for facilitating the exploration of public voice in machine learning through deliberative knowledge spaces where ideas and concepts can be explored

Capture and publish your involvement to contribute to the evidence base to guide practice

Contact

Sophie.Staniszewska@warwick.ac.uk

Professor Sophie Staniszewska

Division of Health Sciences

Warwick Medical School

University of Warwick