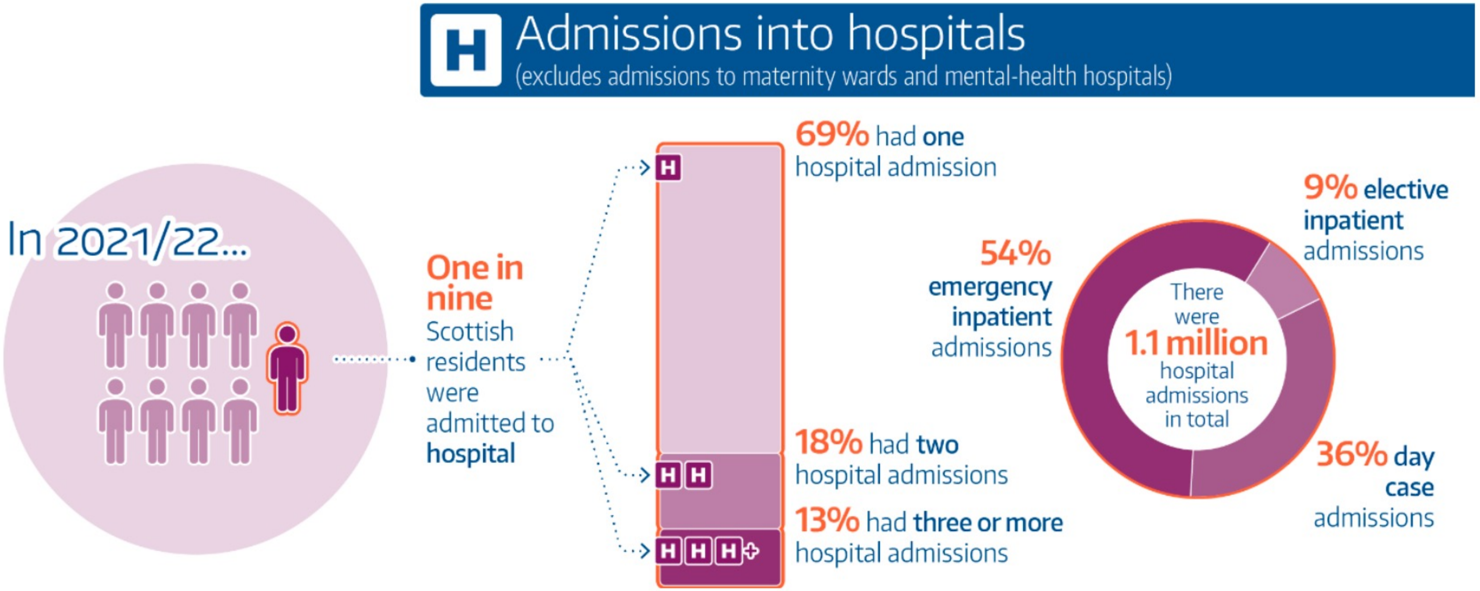
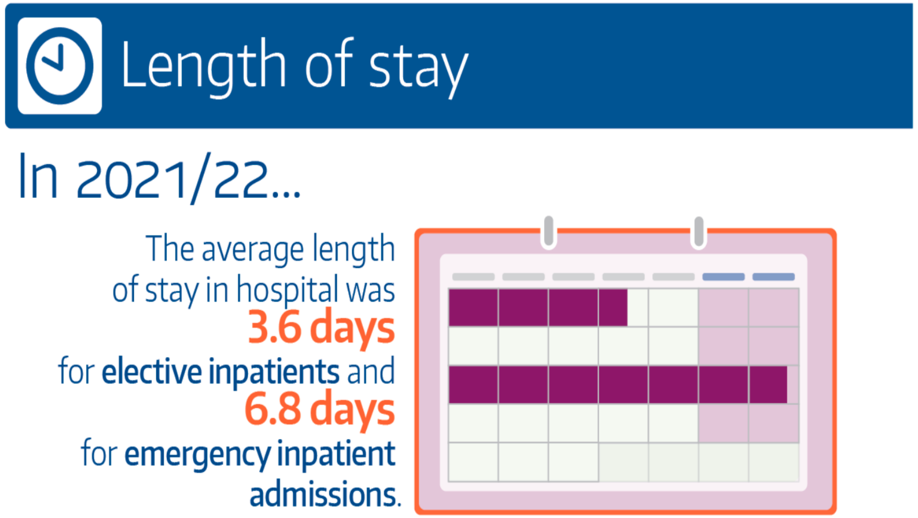
Predicting emergency admissions in Scotland

James Liley & Ioanna Thoma
Durham University & University of Edinburgh

22/06/2023
Digital Health Community Day
Nexus, University of Leeds

Hospital admissions in Scotland





Source: Public Health Scotland

Context and motivation

- Within a year (between 1 April 2021 and 31 March 2022) approximately 1 in 14 had at least one EA.
- Modern public health policies aim to reduce this burden through proactive strategies, e.g., by appropriate primary care intervention.
- Machine learning (ML) can support such interventions by identifying individuals at high risk of EA who may benefit from anticipatory care.
- Successful interventions can lead to better patient outcomes and reduced pressures on secondary care.

SPARRA: Scottish Patients at Risk of Readmission and Admission

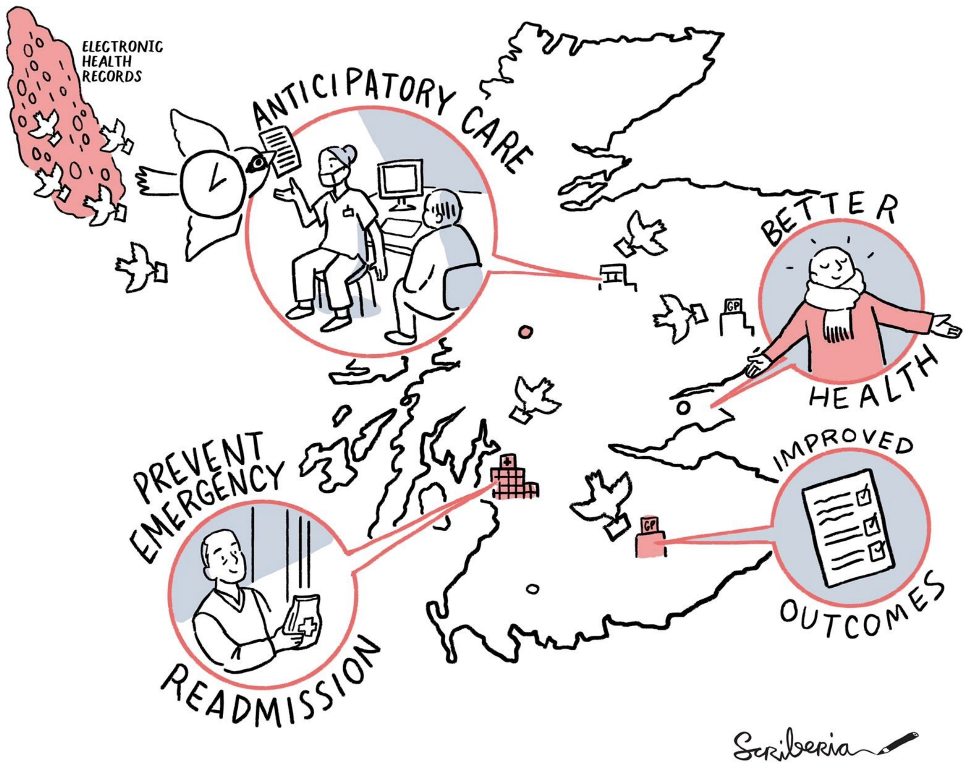
Predicts an individual's risk of being admitted to hospital as an emergency inpatient within the next year

1% = very unlikely to be admitted

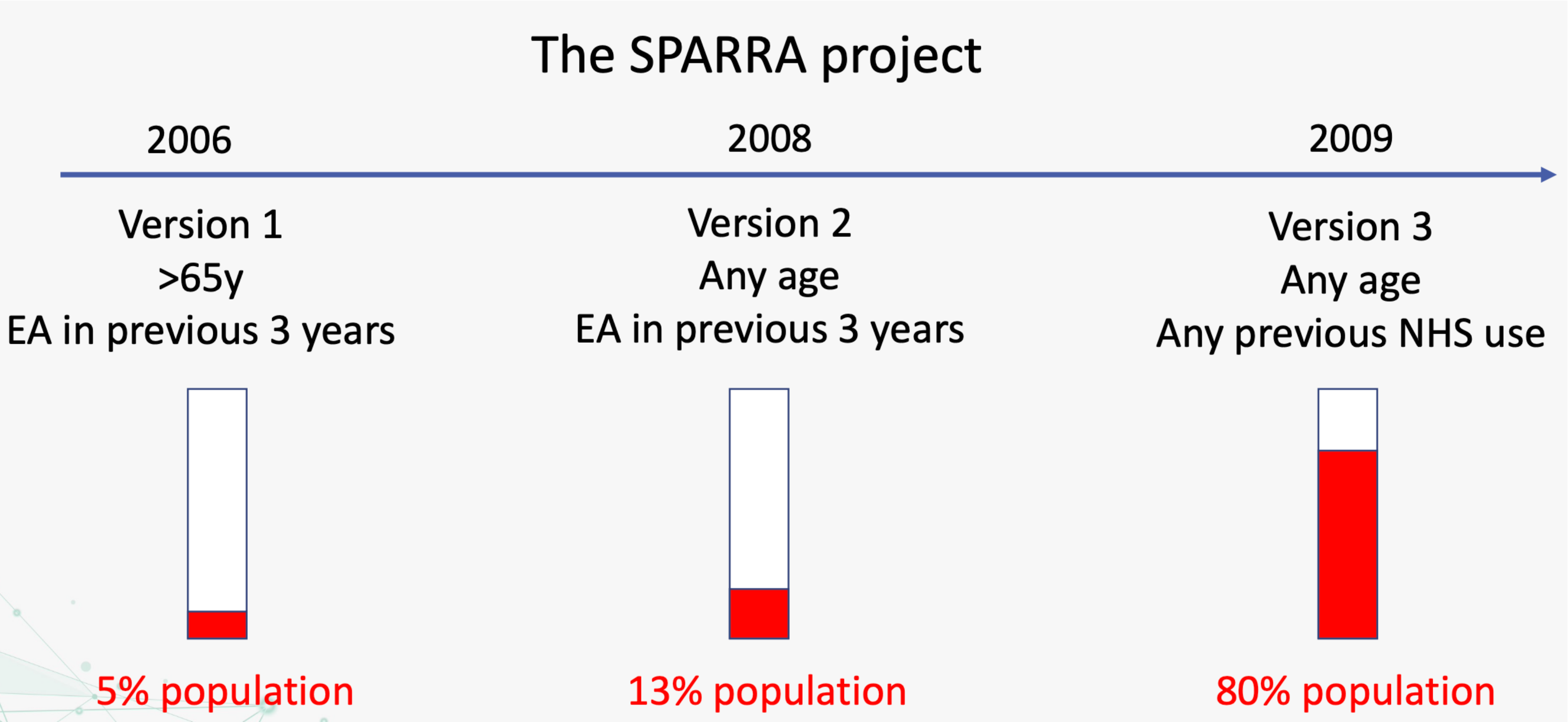
99% = very likely to be admitted

Individually used to plan anticipatory care plans

Collectively used to plan for future demand



Deployment history

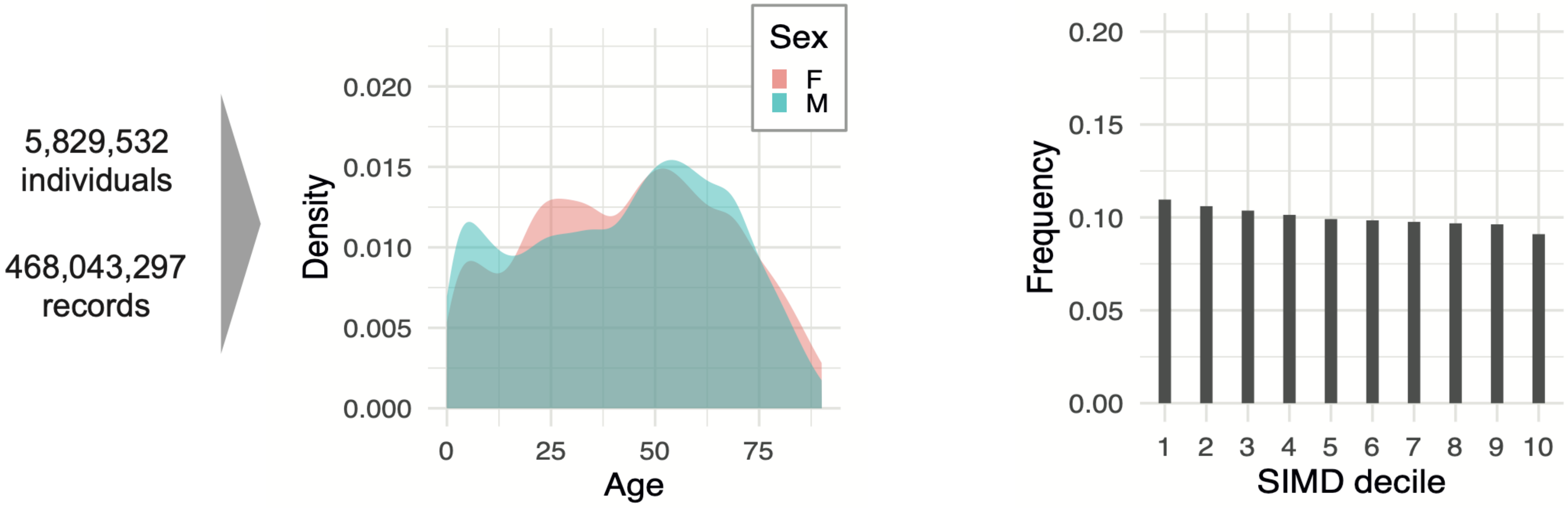


SPARRA Version 4: Any age - Any NHS Use
Exploit modern: Feature engineering, ML/AI, Interpretability, Reproducibility

Prediction based upon electronic health data for the past 1-3 years
Hospital visits / stays, Prescribing, Long term conditions (since 1981)

Input data

- Multiple national EHR databases for 5.8 million Scottish residents between 1 May 2013 and 30 April 2018
- After exclusions: 12.8 million samples of 4.8 million individuals
- The cohort is slightly older, has more females and is moderately more deprived than the general population



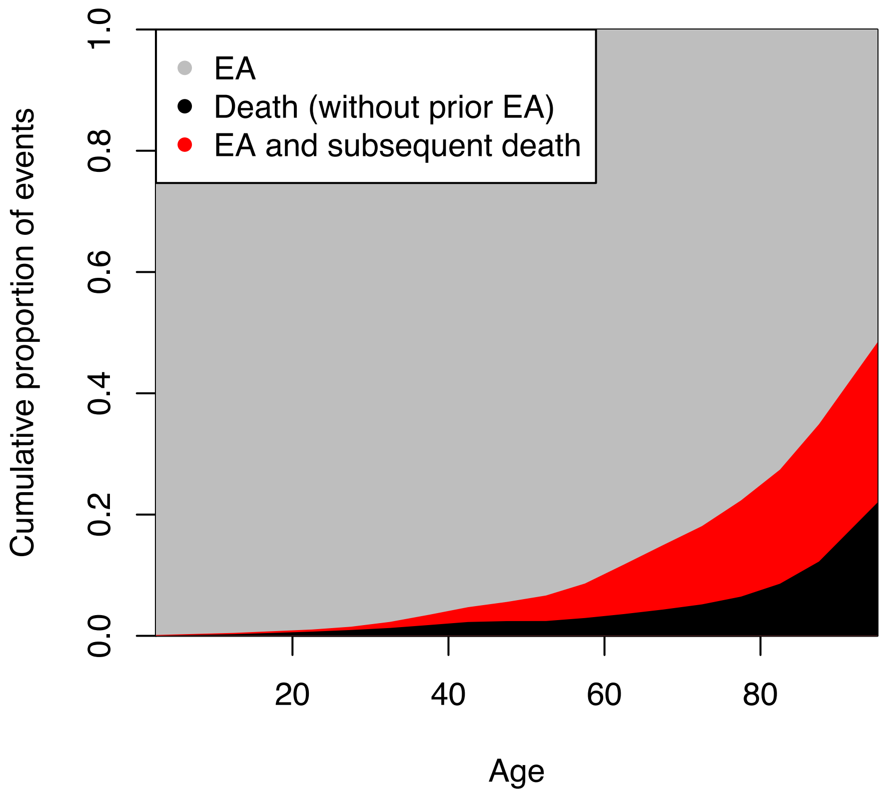
Outcomes

1,142,169 EA or death events (9%) =

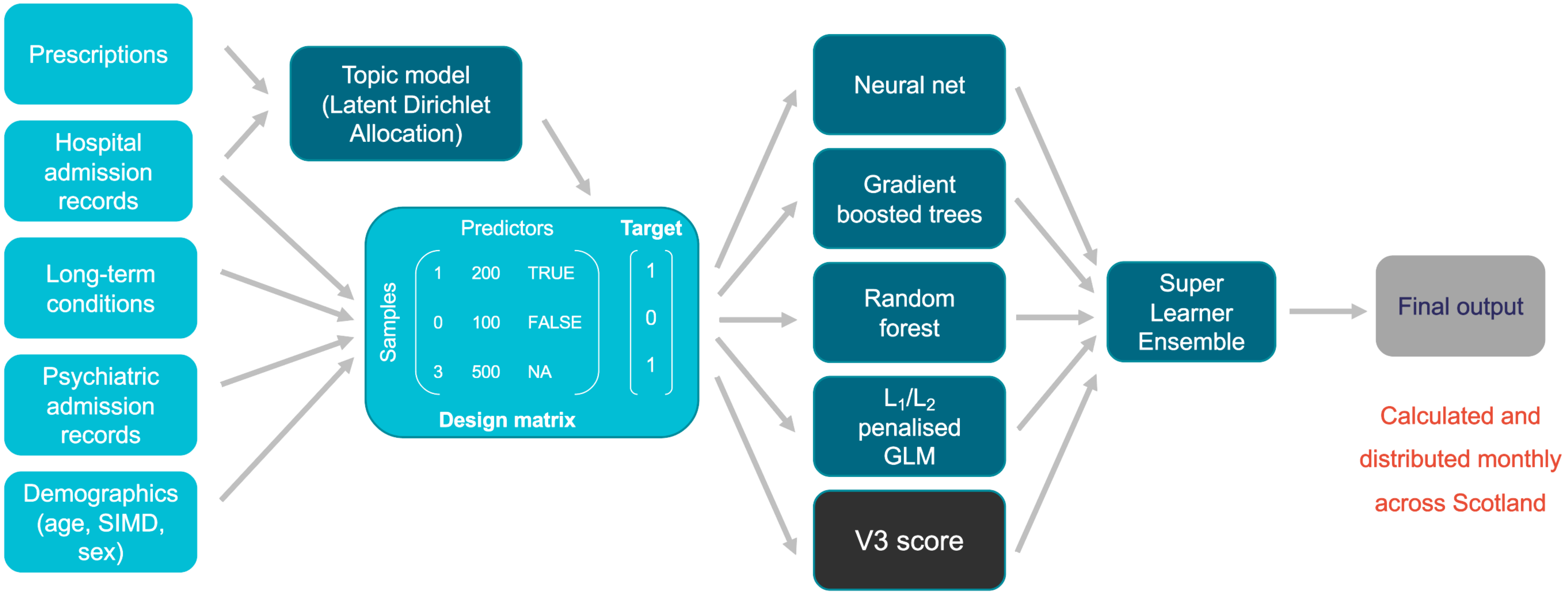
1,084,986 EAs recorded + 57,183 deaths recorded (without a prior EA within that year)

107,827 deaths after the EA

The proportion of deaths amongst the observed events increases with age, as expected.

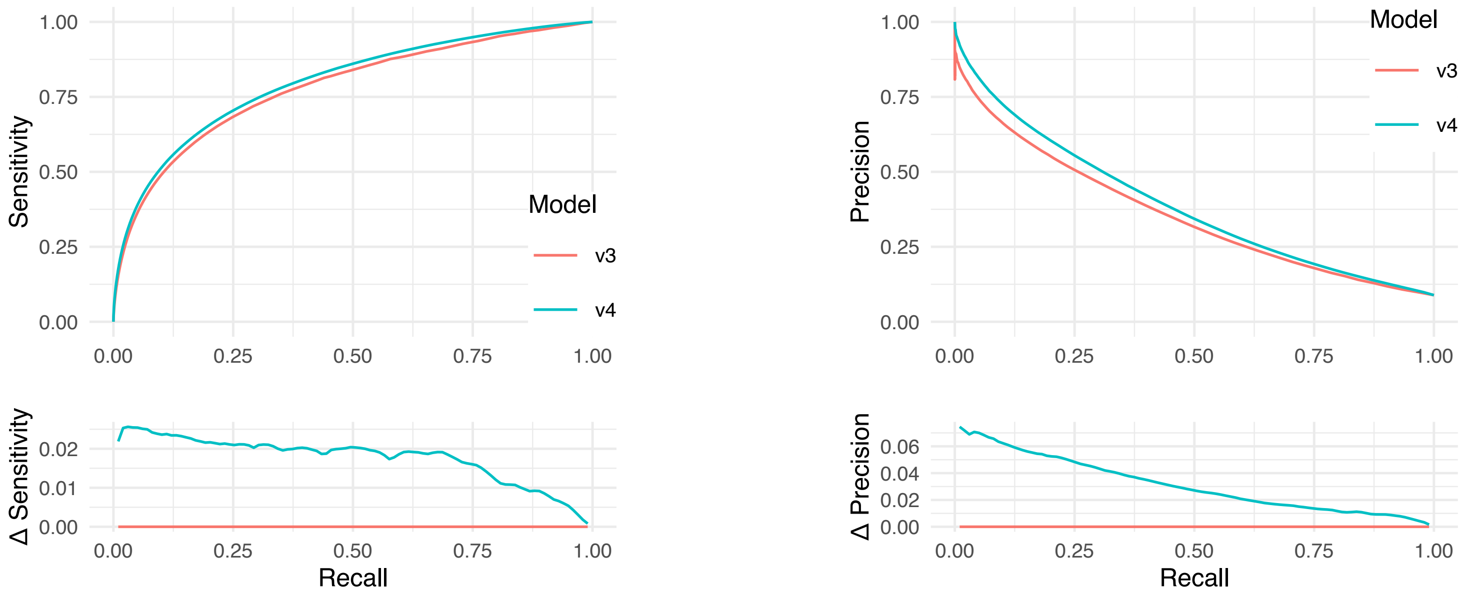


SPARRA version 4 fitting



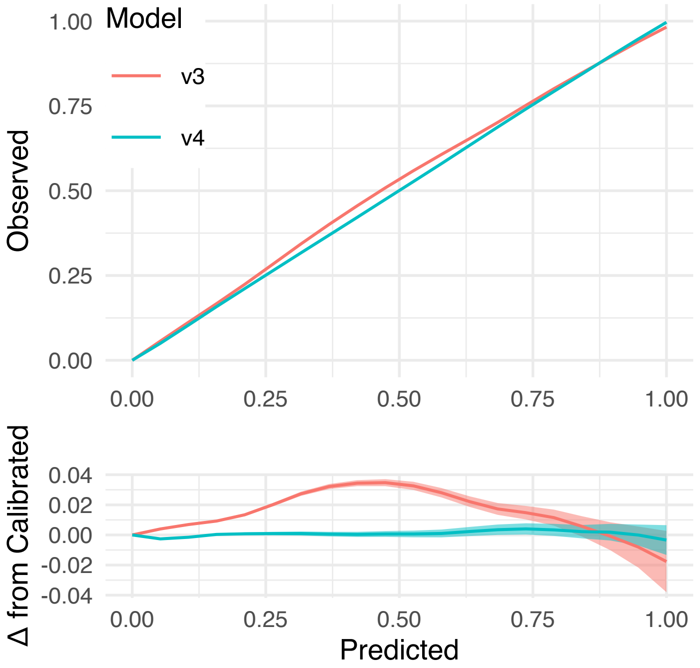
Predictive performance

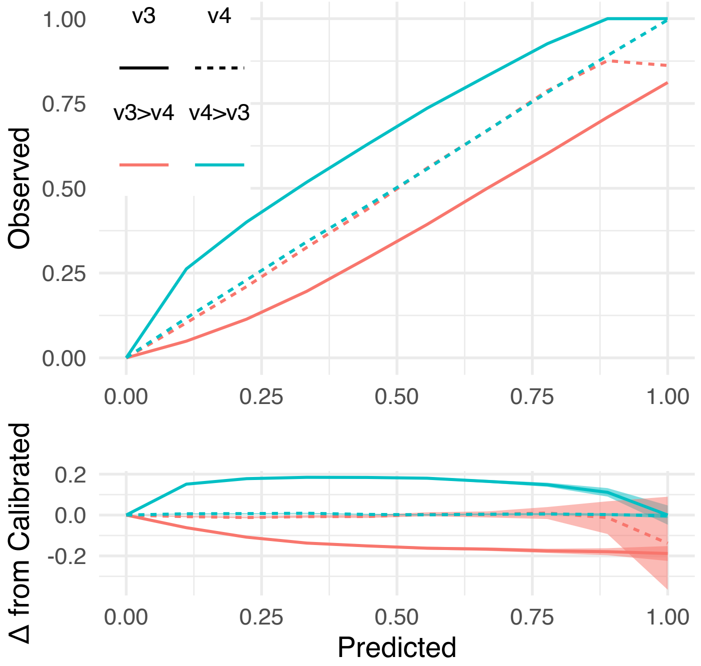
- Three-fold cross validation: SPARRAv4 was effective at predicting EA, and outperformed SPARRAv3 on the basis of AUROC and AUPRC



- SPARRAv4 was also better calibrated, particularly for medium risk individuals, who are less likely to be already known to GPs

Comparison of overall predictive performance between SPARRAv3 and SPARRAv4

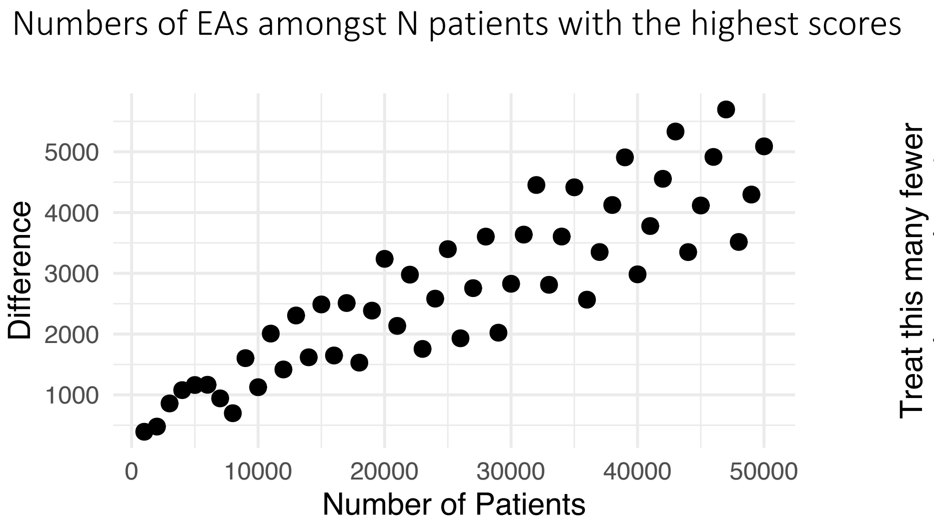


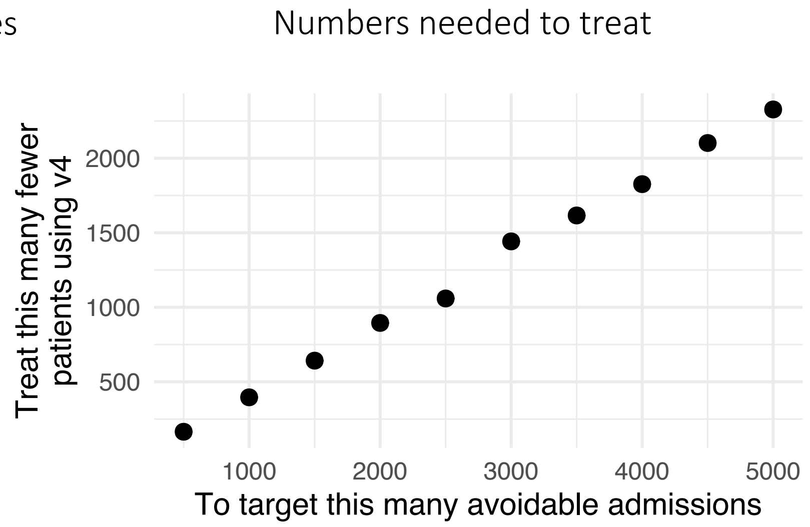


In individuals for whom v3 and v4 “disagreed”, *defined as* |v3-v4|> 0.1, v4 was better calibrated than v3

Impact

Amongst the 50,000 individuals judged to be at highest risk by SPARRAv3, around 4,000 fewer individuals were eventually admitted that were amongst the 50,000 individuals judged to be at highest risk by SPARRAv4

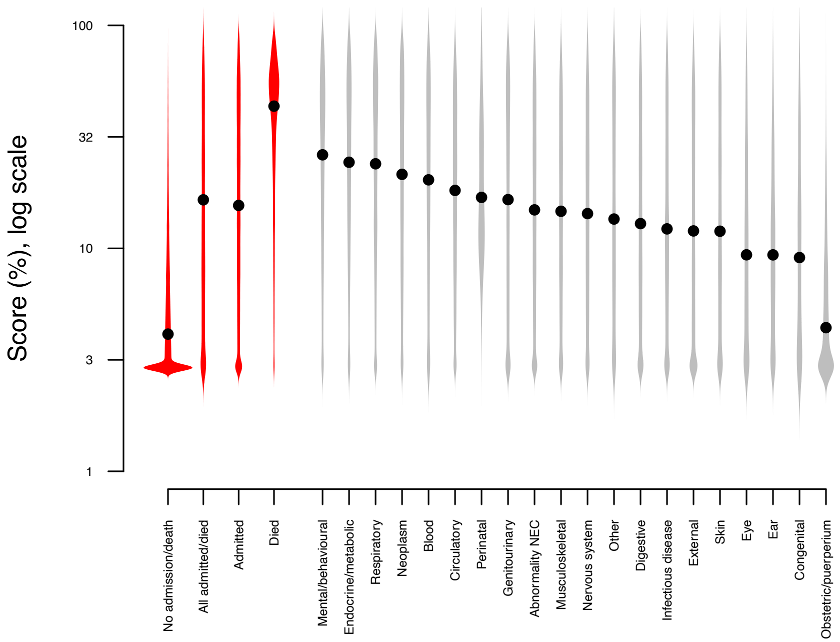




If we aim to pre-empt 3,000 avoidable admissions by intervention on the highest risk patients, then by using SPARRAv4, we would need to intervene on approximately 1,500 fewer patients compared to SPARRAv3.

Performance by admission type and imminence

- Certain medical classes of admission were predicted differentially well
- All-cause mortality associated with high SPARRAv4 scores



- SPARRAv4 tends to better predict imminent admissions
  - individuals with high risk scores were more likely to have an EA near the start of the 1 year outcome period

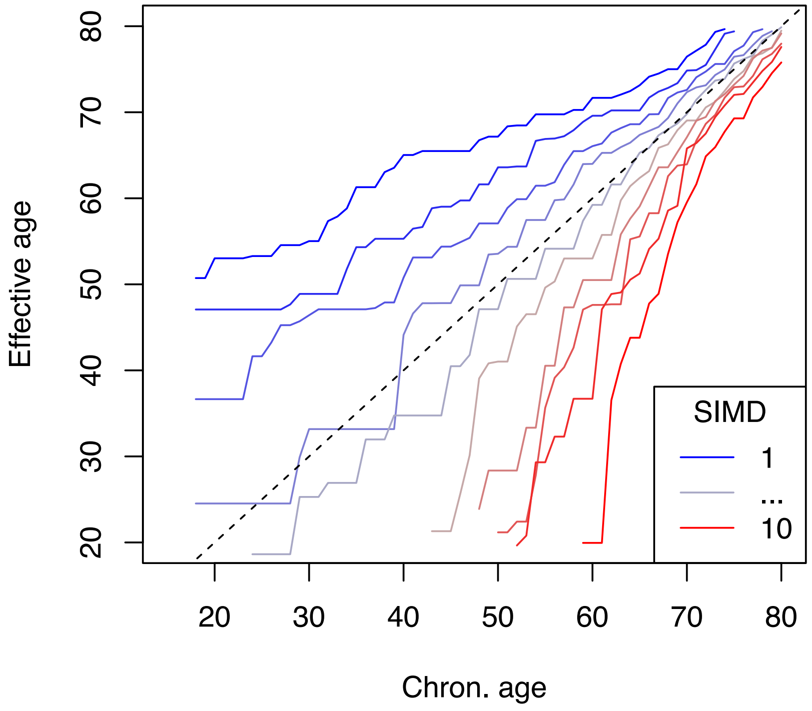
Model updating

- Since SPARRAv3 is already visible to GPs (who may intervene to reduce the risk of high-risk patients), v3 can alter observed risk in training data used for v4, with SPARRA becoming a `victim of its own success'
  - if some risk factor R confers high v3 scores prompting GP intervention (e.g., enhanced follow-up), then in the training data for v4, R may no longer confer increased risk -- potentially hazardous
- Should v4 replace v3, some individuals would therefore have their EA risk underestimated, potentially diverting important anticipatory care away from them
- Possible solution: deploy the maximum of v3 and v4, which averts this potential danger at the cost of a small decrease in calibration.

Liley J, et al. AISTATS 2021

Lessons learned

- Exemplar of a population-scale machine learning score to be deployed in a healthcare setting.
- Opportunities for improved patient outcomes and NHS cost savings: identifying moderate-risk patients not yet prioritised.
- Certain types of admissions can be predicted more efficiently.
- Individual scores tend to rise slowly over time, but very high scores are transient and decrease over time.
- Most features had non-linear effects: SIMD has a substantial effect on EA risk, with the difference between SIMD1 and SIMD10 equivalent to 20-40 additional years of age.



for an (age, SIMD) pair, the age at mean SIMD with the equivalent EA rate

A big team effort

github.com/jamesliley/SPARRAv4

University of Edinburgh
- Catalina Vallejos Meneses
- Ioanna Thoma
- Louis Chislett

Durham University
- James Liley
- Louis Aslett
- Sam Emerson

University of Warwick
- Sebastian Vollmer
- Sophie Staniszewska

The Alan Turing Institute
- Gergo Bohner
- Nathan Cunningham
- Franz Kiraly
- Ioanna Manolopoulou
- Bilal Mateen
- Katrina Payne
- Chris Holmes

Public Health Scotland
- Katie Borland
- David Carr
- Scott Heald
- Sam Oduro
- Jill Ireland
- Keith Moffat
- Rachel Porteous
- Stephen Riddell
- Simon Rogers (NSS)
- Beth Bruce (eDRIS)
- Lizzie Nicholson (eDRIS)

Partners/Funders

The Alan Turing Institute

Durham University

The University of Edinburgh

Public Health Scotland

WARWICK THE UNIVERSITY OF WARWICK

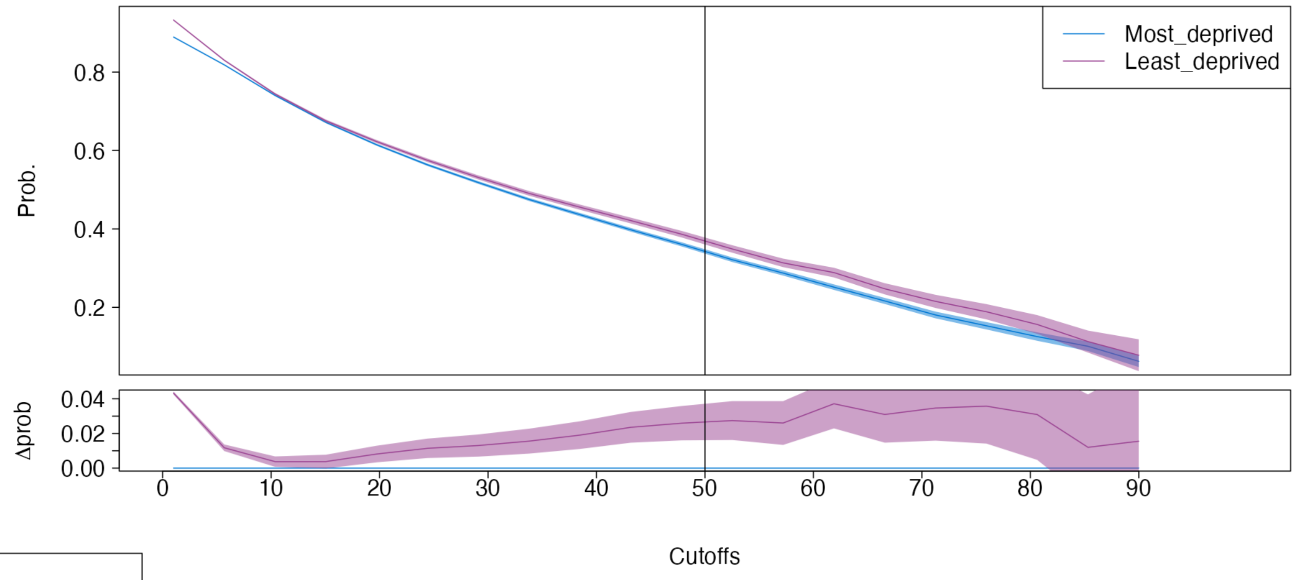
HDRUK Health Data Research UK

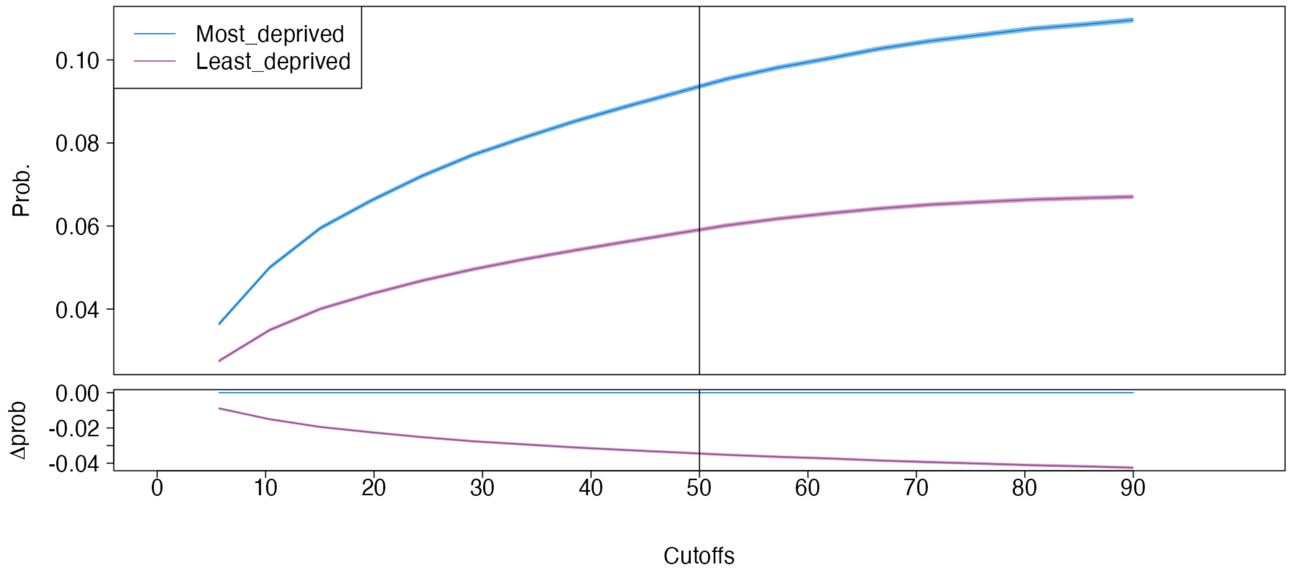
UKRI UK Research and Innovation

The Health Foundation

Health inequalities analysis - WIP

False discoveries





False omissions

Stability and performance attenuation

- We assess the durability of performance for a model trained once and employed to generate predictions at future times, confirming it does not deteriorate.
  - It is essential to update predictions despite the absence of model updates, since evolving patient covariates lead to the performance attenuation of any point-in-time prediction.
- Individual scores tend to rise slowly over time, but very high scores are transient and decrease over time.
- The static scores performed reasonably well even 2-3 years after baseline although discrimination and calibration were gradually lost