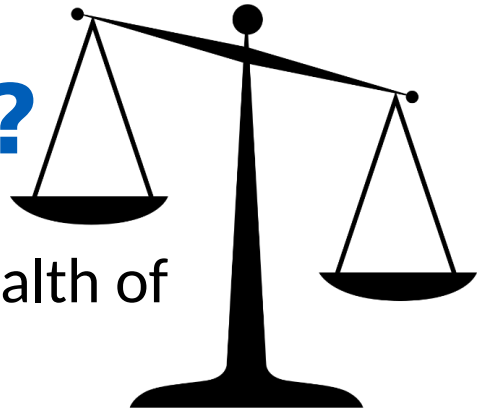
Diabetes prevalence management and health inequalities

Digital Analysis and Research Team (DART)
Data Science Internship

Author: Stephen Richer
Supervisor: Paul Carroll

What are Health Inequalities?



- Health ***inequality*** generically refers to differences in the health of individuals or groups.
  - Red-Green colour blindness is more common in men.

- Health ***inequity*** are specific inequalities deemed to be *unfair* or *unjust*.
  - Higher incidences of lung-cancer in areas of socio-economic deprivation may be associated with targeted tobacco advertising in lower income neighbourhoods.

'Health equity means that everyone has a fair and just opportunity to be as healthy as possible. This requires removing obstacles to health such as poverty, discrimination and their consequences, including powerlessness and lack of access to good jobs with fair pay, quality education and housing, safe environments and health care' Braveman et al., 2017

What is Diabetes?

Type 1

- Failure of pancreas in insulin production – sudden onset with no clear cause.

Type 2 (most common)

- Insulin resistance – failure of cells to respond to insulin.

Risk Factors

- Being overweight
- Sedentary lifestyle
- Family history of diabetes
- Age 45 or older
- Certain Ethnic Groups

'Diabetes is a metabolic disorder characterised by chronic hyperglycaemia (high blood sugar)' (WHO, 1999)

Project Outcomes

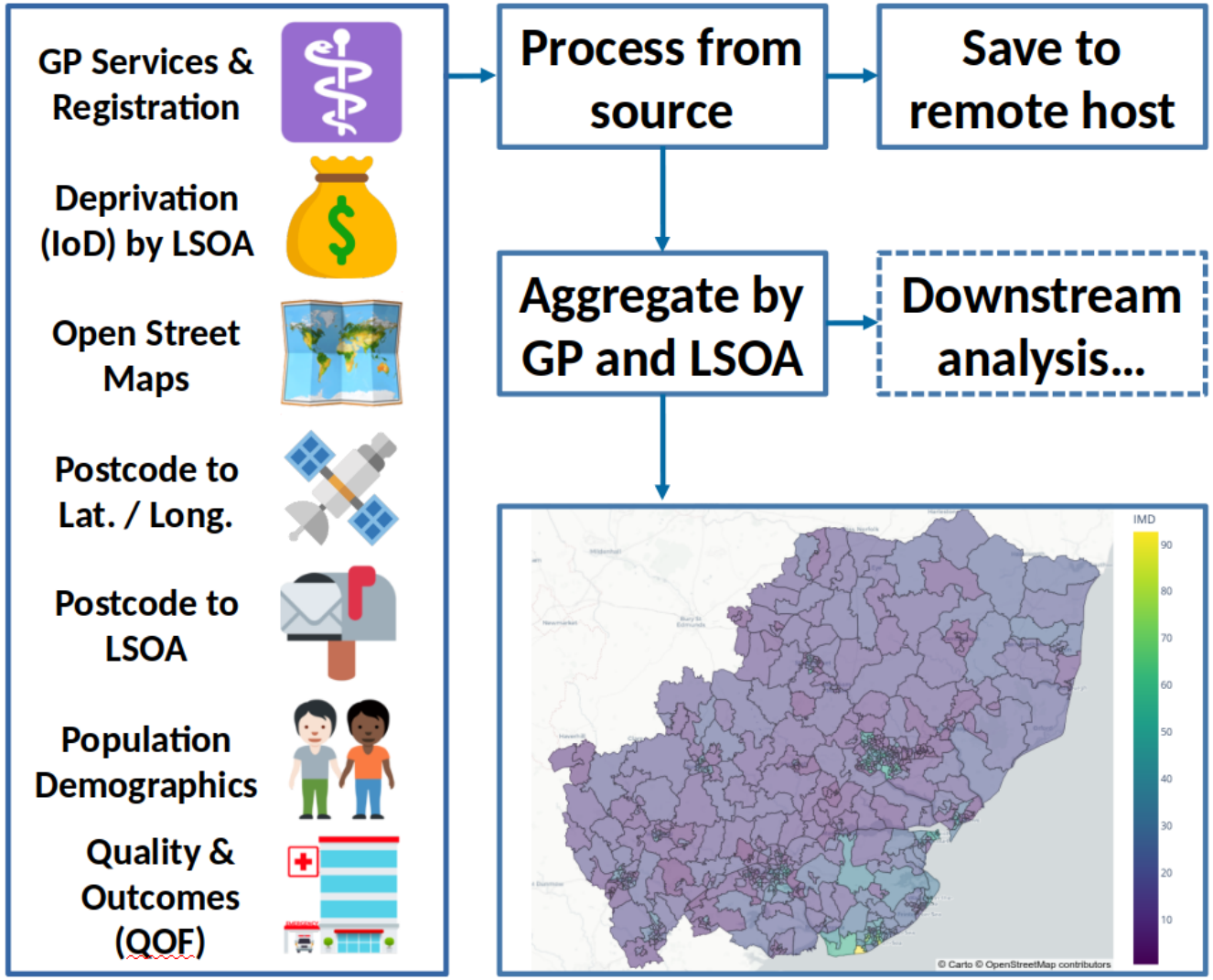
1) Inequality Analysis: National Level and at Ipswich Diabetes Centre
Exploring trends in Diabetes prevalence and inequality of service accessibility.

2) Estimate 'Did Not Attend' Probability (DNApredict)
Supervised binary classifier to estimate DNA risk.

3) Morbidity Network Analysis (MultiNet)
Python package to generate and analyse simple graphs from morbidity data.

1) Inequality Analysis:
National Level and at Ipswich Diabetes Center

ESNEFT Tools



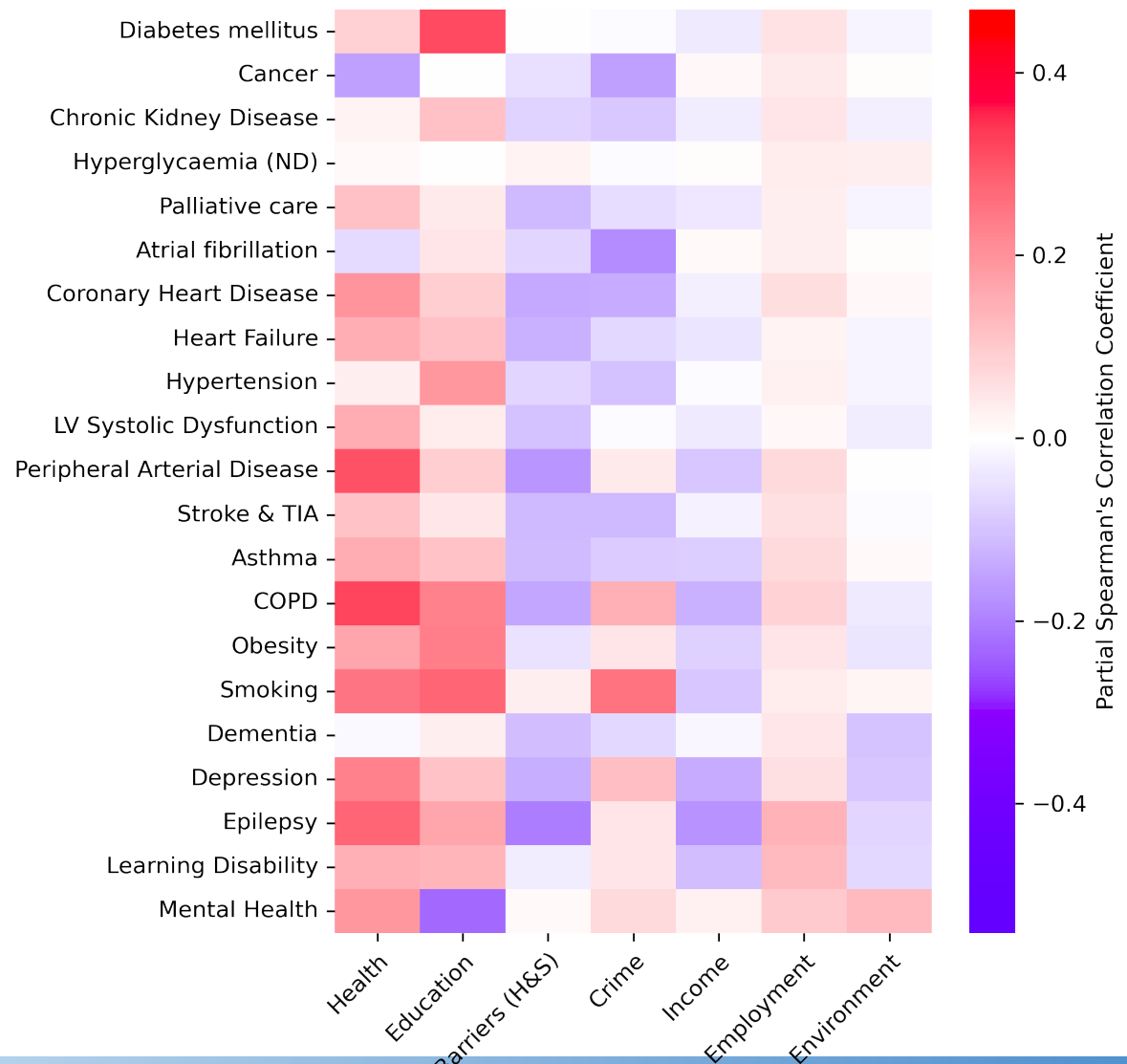
ESNEFT Tools implements a simple interface for accessing and manipulating a wide variety of datasets relating to demographic and population inequalities.

Available at:
https://github.com/nhsx/p24-diabetes-inequal

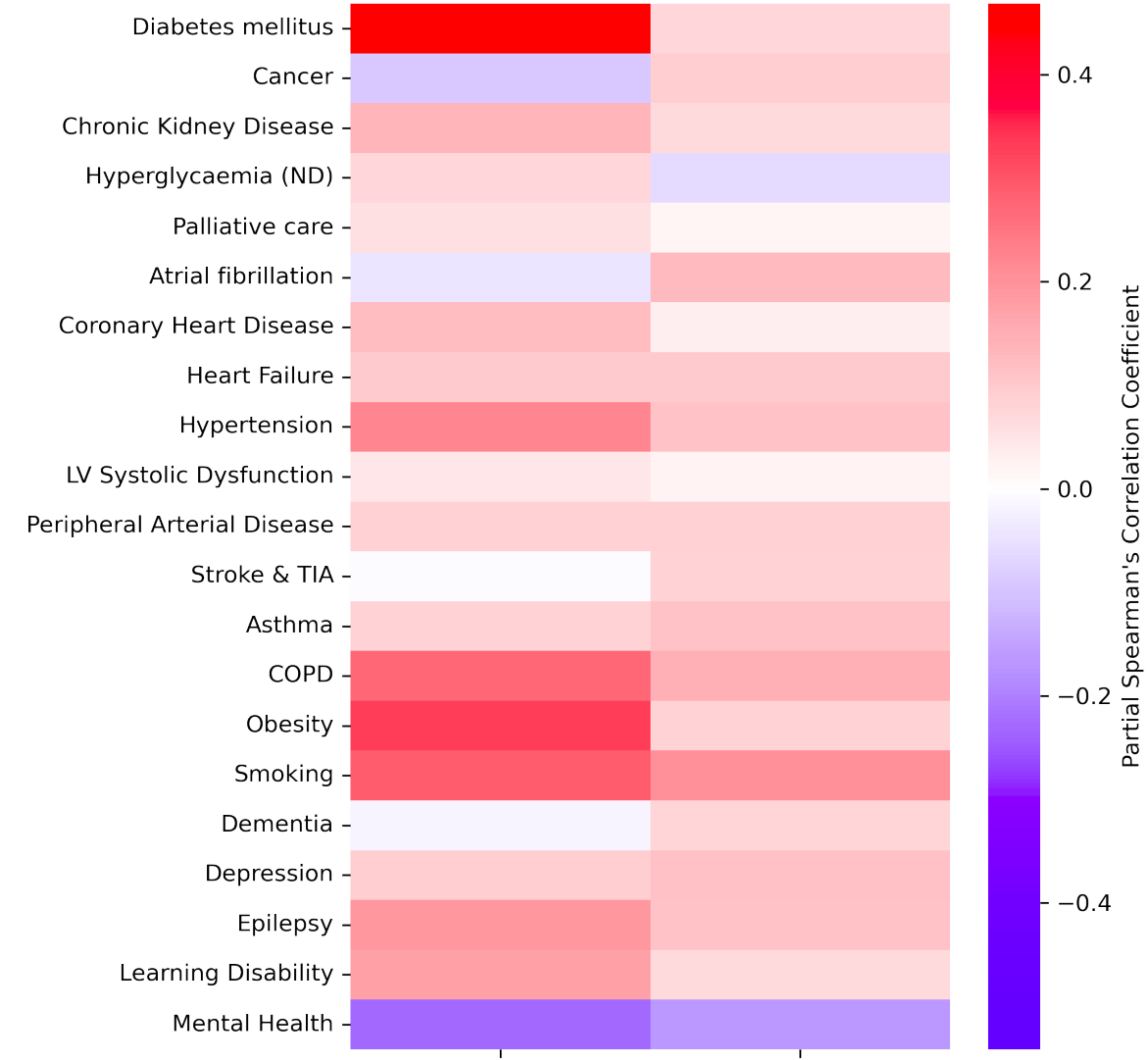
Investigating sub-domains of deprivation associated with Disease Prevalence

- The Index of Multiple Deprivation (IMD) is comprised of 7 sub-domains and each sub-domain is itself built from underlying indicators.
- Analysing the underlying indicators can provide more meaningful insights into which specific factors of deprivation are associated with the feature of interest (e.g. diabetes prevalence)
- Analysis method:
  - ESNEFT Tools extracts GP-level disease prevalence data from the QoF and maps it to LSOA level.
  - A partial Spearman correlation was performed between each disease prevalence and the deprivation score.
    - Partial correlation can control from the confounding influence of the other deprivation domains and demographic factors (e.g. age).

- Diabetes prevalence was most strongly associated with the Education sub-domain.
- Within Education, Diabetes was most strongly associated with Adult Skills and specifically Adult Education.



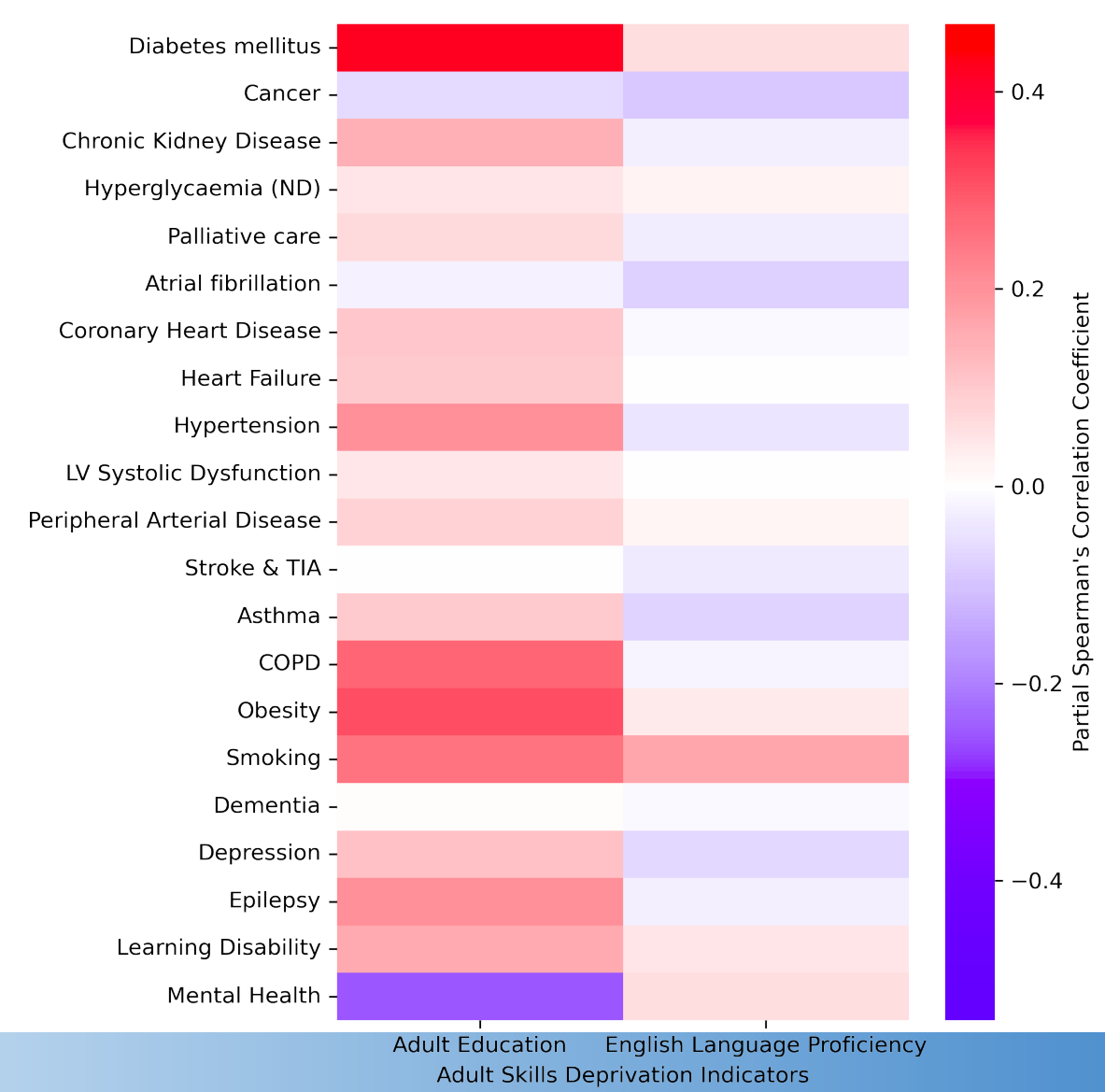
- Figure shows heat map of partial correlation coefficient between disease prevalence and the 7 sub-domains of deprivation.
- Health deprivation is most strongly associated with disease prevalence.
  - This is expected as the sub-domain is defined by the health of the population.
- Education deprivation is second most strongly associated.
  - Diabetes, in particular, is highly associated with Education deprivation.
- To investigate further, the analysis was repeated on the underlying indicators of Education deprivation: Adult Skills and Youth Skills.



- Adult Skills was identified as the predominant underlying indicator most strongly associated with Diabetes prevalence

- Adult Skills itself is defined by two indicators:
  - Adult Education: The proportion of the working age population with low or no qualifications.
  - English Language Proficiency: The proportion of the working age population who cannot speak English or cannot speak it well.

- As before, to investigate further, the analysis was repeated on these underlying indicators of Adult Skills deprivation.



- Adult Education was identified as the predominant underlying indicator most strongly associated with Diabetes prevalence
- English Language Proficiency shows minimal association with disease prevalance.

Investigating associations with deprivation and service usage and accessibility at the Ipswich Diabetes Centre

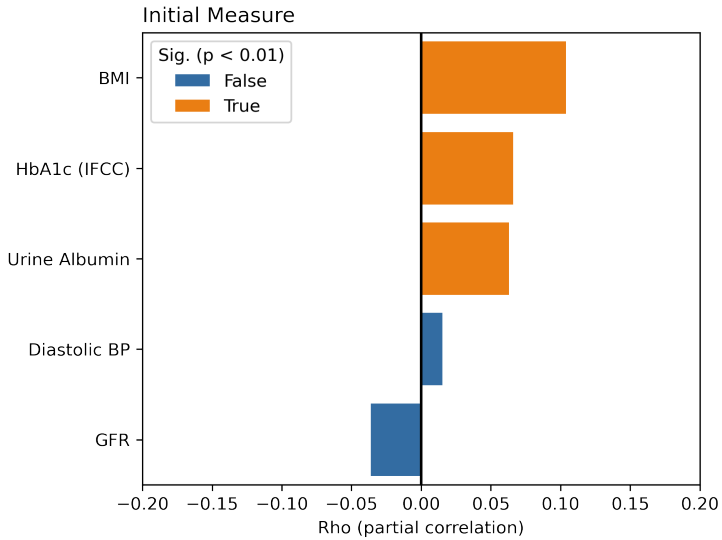
- 268,912 anonymised records, corresponding to a full history of 8,521 patients, were obtained from the Ipswich Diabetes Centre from between approximately 2014 and 2022.
- Patient demographic data were obtained by joining a separate ESNEFT outpatient dataset by anonymised patient identifier – 5,713 patients successfully associated.
- Patient was assigned a deprivation score according to the LSOA associated with their postcode of residence.
- The analysis sought to explore difference in service utilisation associated with deprivation scores.
  - Analysis revealed that patients from deprivation LSOA's had worse health metrics upon initial referral
  - No other deprivation associated differences were detected following referral.
- Findings suggest that referral of patients from deprived backgrounds may be delayed, perhaps due to delayed diagnosis or earlier disease onset.

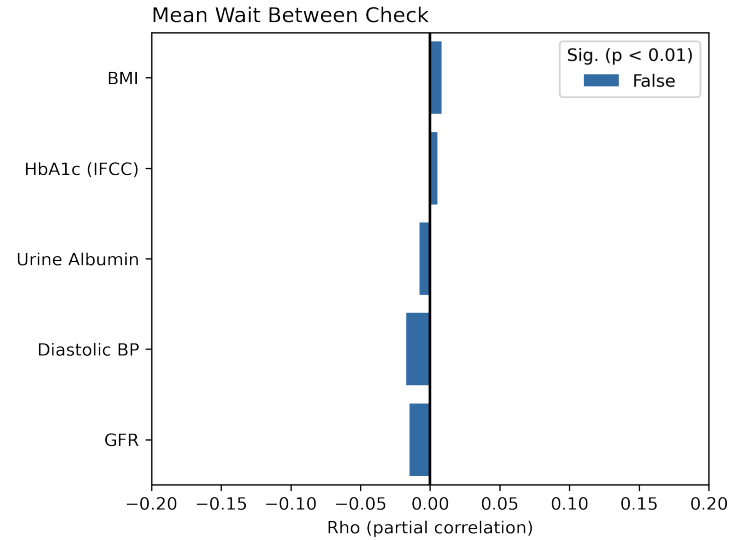
Example of Diabetes Centre Dataset

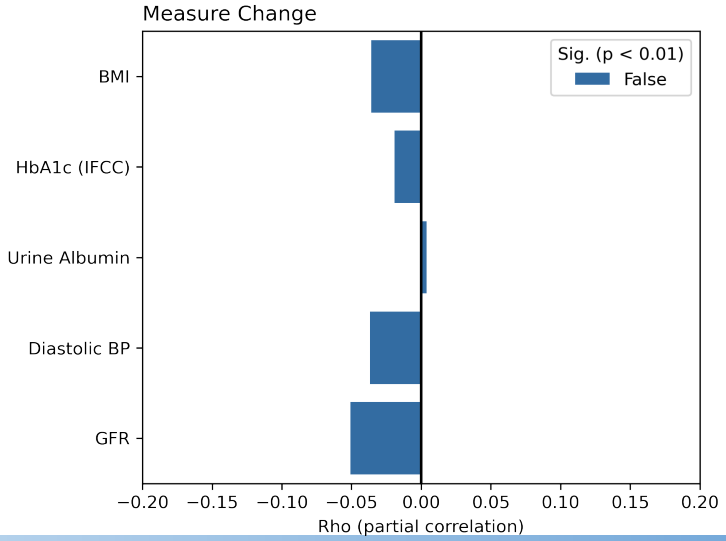
Table 2.1: **Diabetes Centre Data with Patient Demographics**

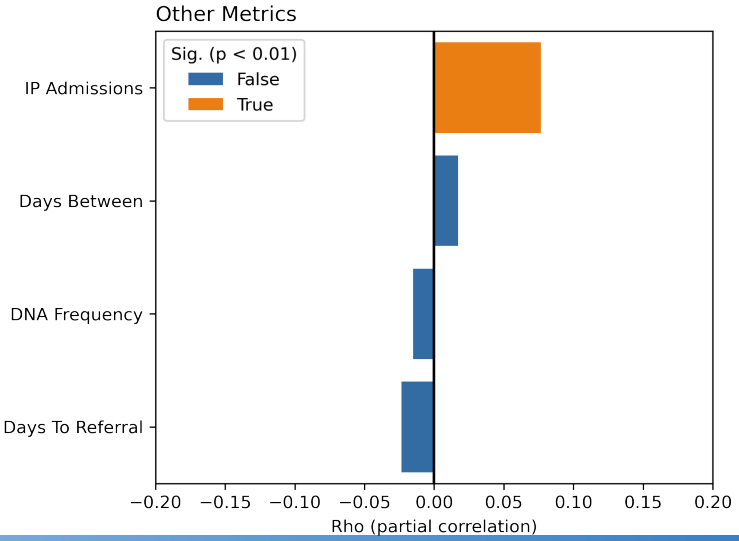
ID	Date	CTV3Desc	Value	Sex	Age	Ethnicity	Deprivation
1	12 Oct 2020	HbA1c	82.0	Male	48	White	34.6
1	21 Dec 2020	HbA1c	79.0	Male	48	White	34.6
1	14 Jan 2021	BMI	28.6	Male	48	White	34.6
1	18 Feb 2021	Did Not Attend	-	Male	48	White	34.6
2	01 Feb 1986	Diagnosis Date	-	Female	66	Mixed	21.5
2	03 Mar 2021	HbA1c	49.0	Female	66	Mixed	21.5
2	03 Mar 2021	BMI	33.1	Female	66	Mixed	21.5

Note: This data is fictitious for illustrative purposes. CTV3Code and Recording Unit columns have been removed from this example. Deprivation represents the raw score of the Adult Skills sub-domain of Education. Demographic data was not part of the original Diabetes Centre data and was joined from a separate outpatient dataset.

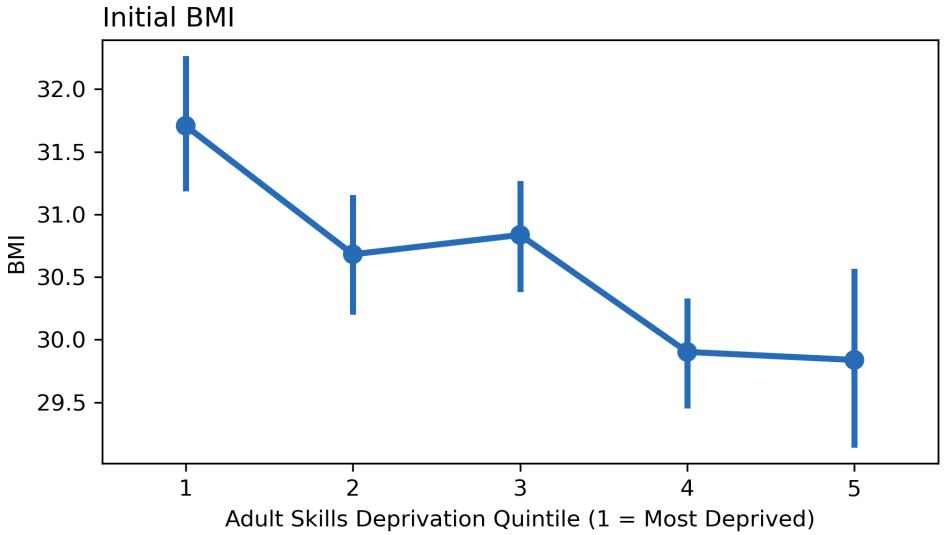


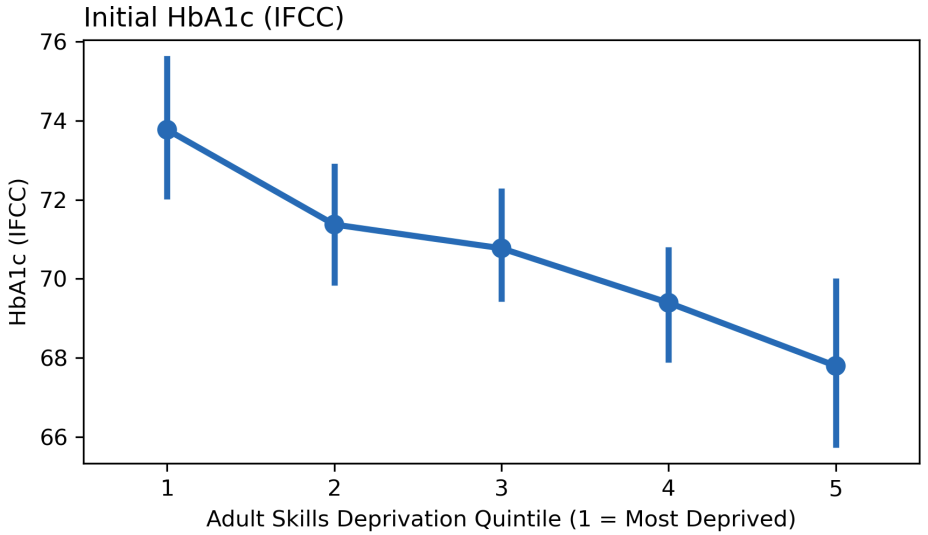


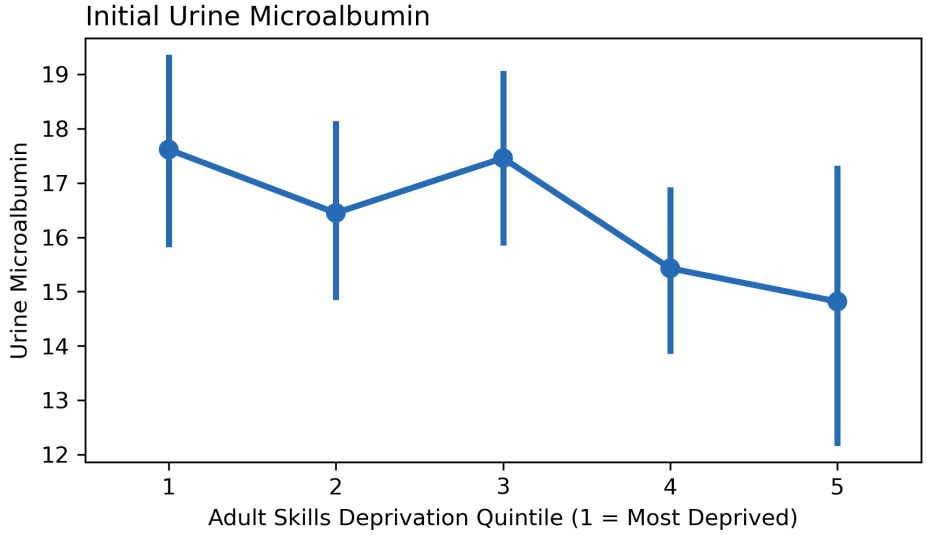


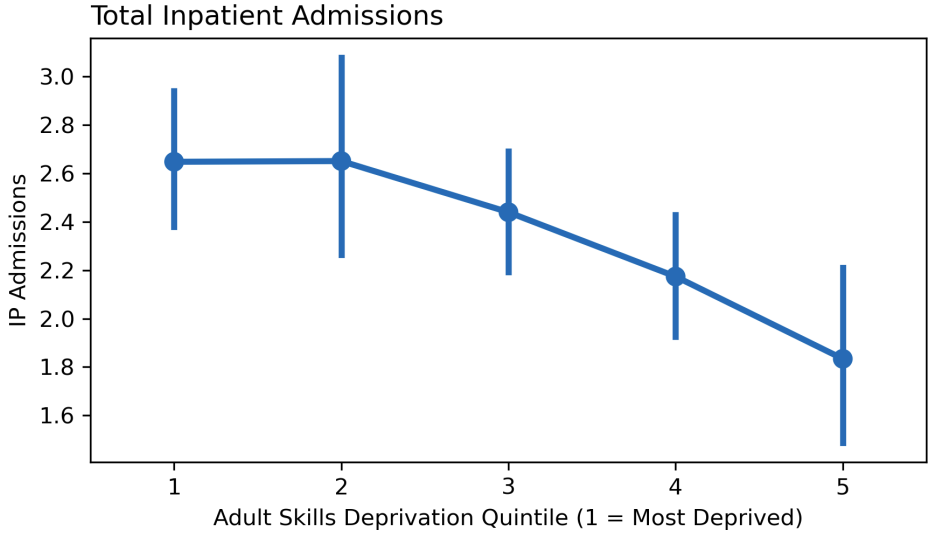


- Partial correlation analysis revealed a positive correlation with deprivation and initial health metrics of BMI, HbA1C and Urine Microalbumin.
- These metrics were all higher (worse) in patients from deprived LSOAs.
- Patients from deprived LSOAs were also more likely to have a more IP admissions.
- Next steps would be to identify the referral routes of patient sub-groups. Are patients from deprived backgrounds more likely to be referred following IP admission?


Initial BMI
BMI
32.0
31.5
31.0
30.5
30.0
29.5
1
2
3
4
5
Adult Skills Deprivation Quintile (1 = Most Deprived)


Initial HbA1c (IFCC)
HbA1c (IFCC)
76
74
72
70
68
66
1
2
3
4
5
Adult Skills Deprivation Quintile (1 = Most Deprived)


Initial Urine Microalbumin
Urine Microalbumin
19
18
17
16
15
14
13
12
1
2
3
4
5
Adult Skills Deprivation Quintile (1 = Most Deprived)


Total Inpatient Admissions
IP Admissions
3.0
2.8
2.6
2.4
2.2
2.0
1.8
1.6
1
2
3
4
5
Adult Skills Deprivation Quintile (1 = Most Deprived)

Caveats of Deprivation Analysis

- Deprivation indexes represent aggregate summaries of a specific area – they cannot be used to infer patient deprivation.
  - Deprived patients may live in affluent areas, affluent patients may live in deprived areas.
  - Cannot distinguish between factors associated with the local area or the individual.
  - Future work would need direct measure of patient deprivation (e.g. education level).

- The Diabetes Centre dataset is an incomplete record of a patient's interaction with healthcare services for the diabetes treatment.
  - Interactions with GPs are not be recorded.
  - No clear pattern of patient interaction with the Diabetes Centre.
    - Why are there large gaps between appointments?
      - Could be due to lack of patient engagement.
      - Alternatively, patient may not need the service due to positive progress.

-

Future Work

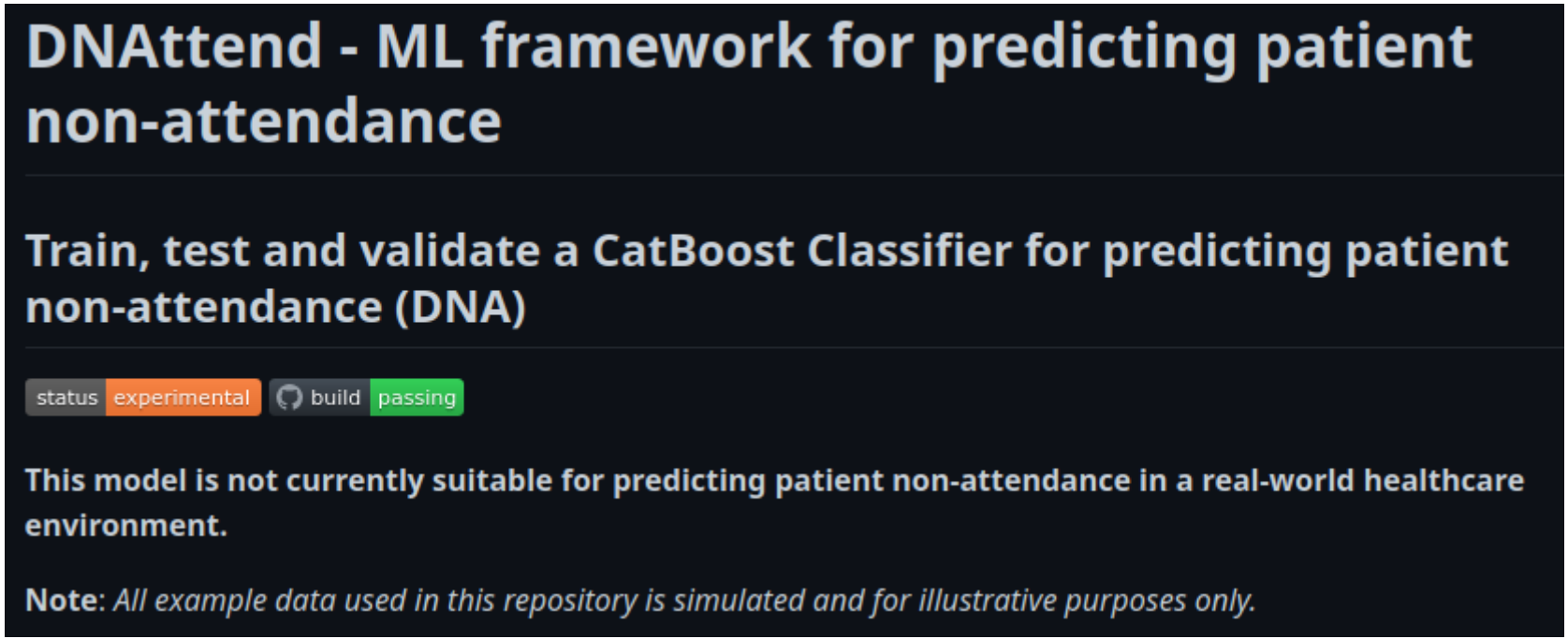
- Future work needs to identify the underlying reasons why patients from deprived LSOAs have worse metrics on referral?

- Patient's may be developing diabetes earlier, or more rapidly, so it is not immediately picked up in primary care.
- Patient's may be under-utilising primary care.
  - Can we compare patient pathways leading up to referral?
  - Are patients being referred from different routes (e.g. primary vs secondary).
- Deprivation may be associated with area, rather than individual.
  - e.g. perhaps GP practises in deprived LSOAs are less resourced.

- A more complete record of the patient interactions for diabetes treatment is required.
  - Individual patient deprivation measures (e.g. education level)
  - Metadata describing *why* patients were referred and the referral source.

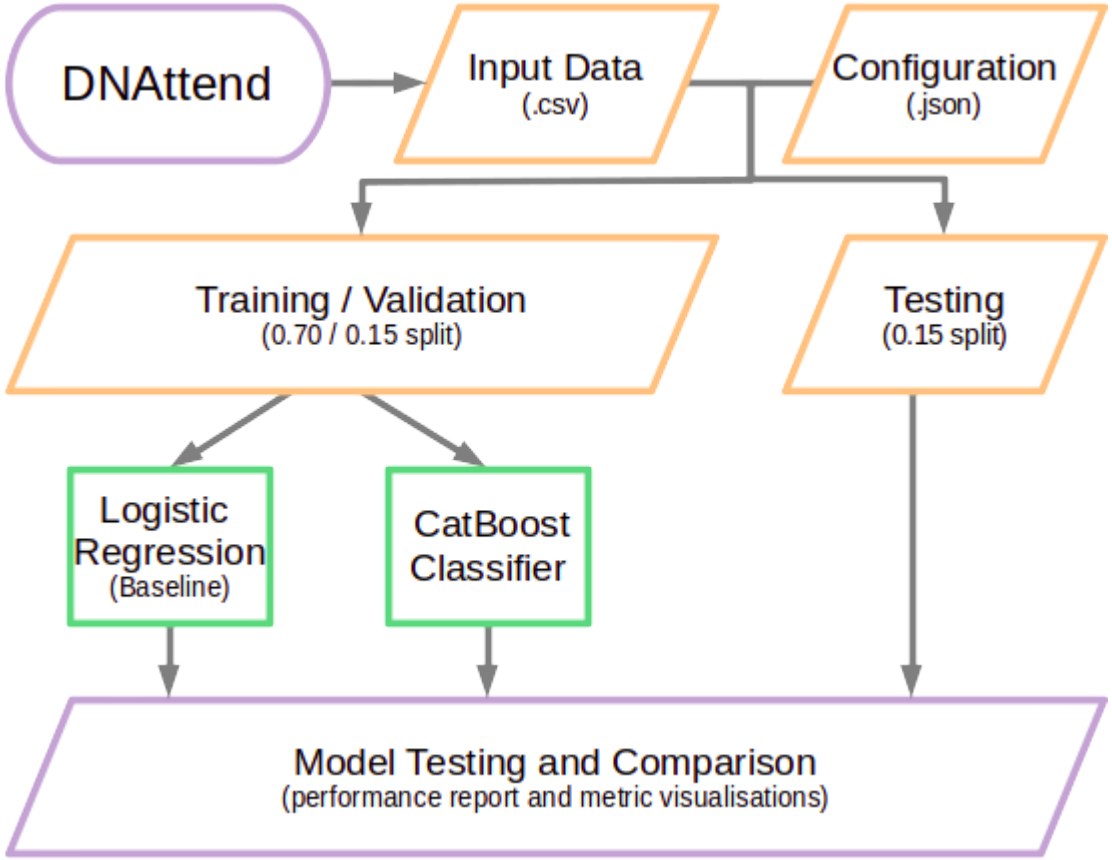
2) DNAttend

AutoML framework for predicting patient non-attendance.



Full Documentation Available At:
https://github.com/nhsx/dna-risk-predict

DNAttend: Simplified Workflow



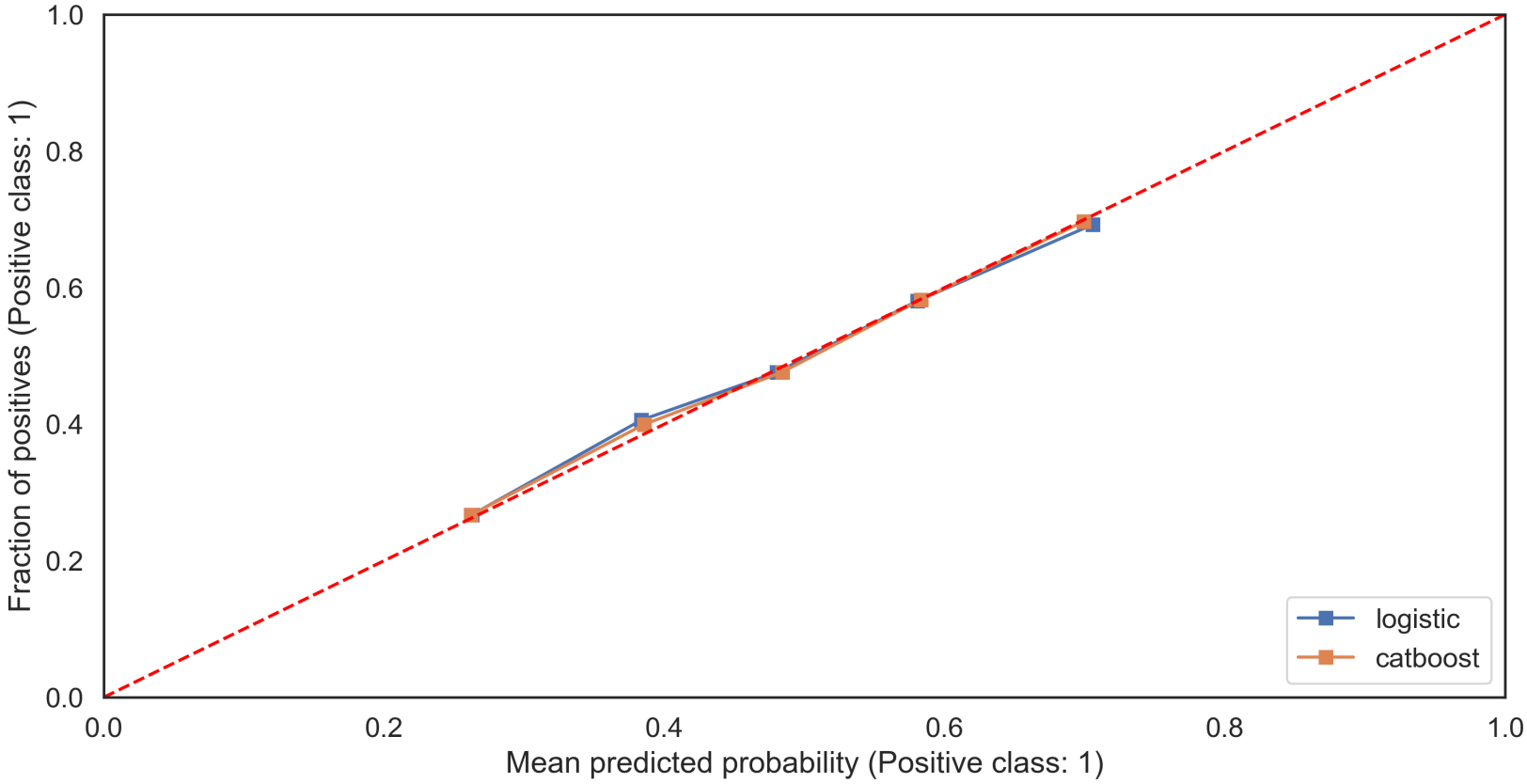
DNAttend: Key Features

- Patient non-attendance is a significant burden to the NHS with approximately 5% of all primary healthcare appointments listed as Did Not Attend (DNA).
- In addition to putting patients at risk, DNAs reduce clinical capacity and are estimated to cost the NHS as much as £216 million annually.
- DNAttend is a supervised machine learning framework for identifying hidden patterns that determine an individual's DNA risk.
- Command-line utility for training and evaluating binary classification models for predicting patient non-attendance.
  - Enables non-expert users to quickly build and evaluate robust models.
- DNAttend builds two models:
  - Logistic regression (simple baseline model)
  - CatBoost (gradient-boosted decision tree)
    - Hyper-parameters are automatically tuned via random search.
- A full example workflow, with simulated data, is available at the GitHub repository.

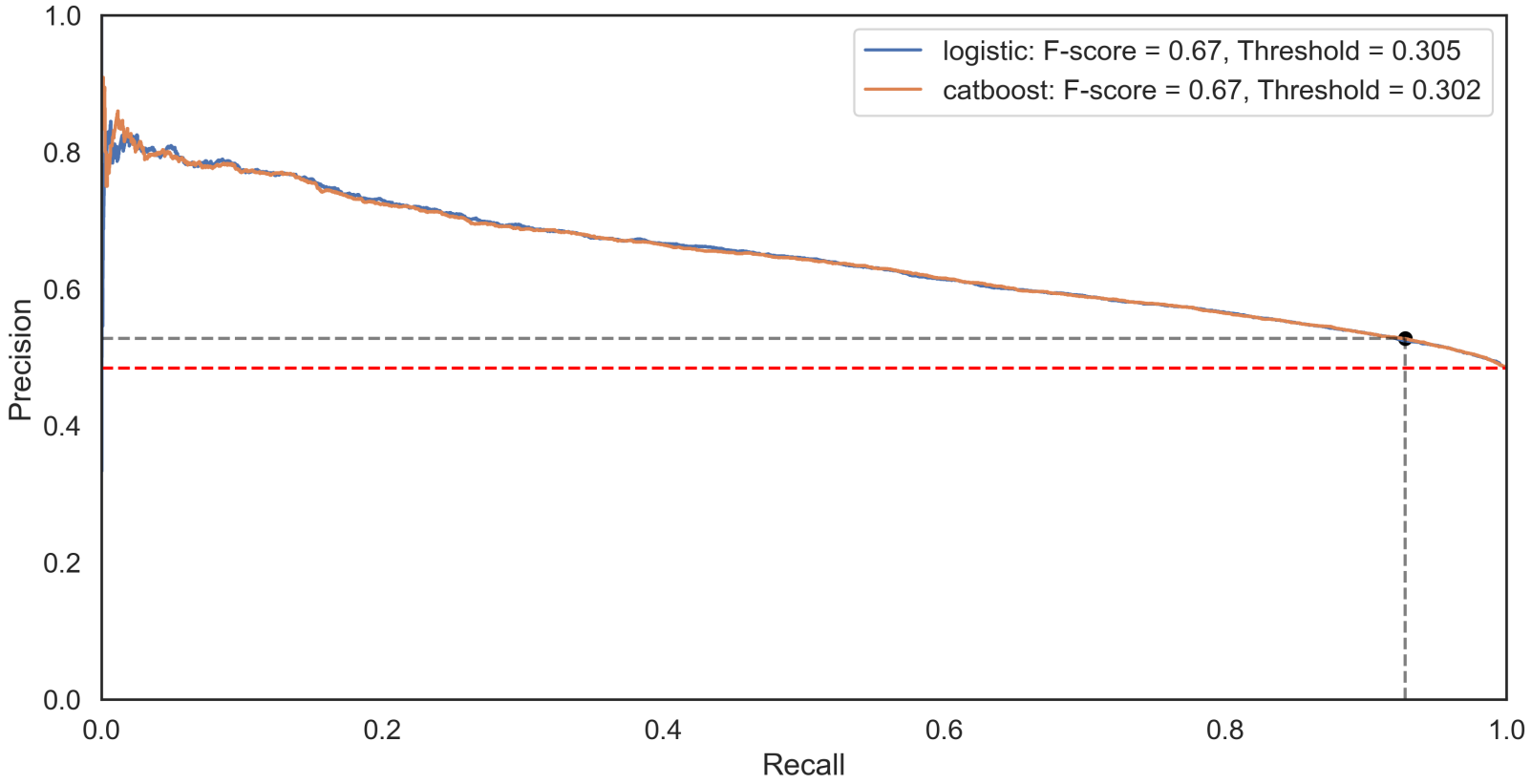
DNAttend: Command Line Interface

```
(base) jovyan@ceb90354e88b:~/work$ dnattend --help
usage: dnattend [-h] [--version] [--verbose] Commands ...

DNAttend

options:
  -h, --help  show this help message and exit
  --version   show program's version number and exit
  --verbose   verbose logging for debugging

required commands:

  Commands    Description:
    train     Train model.
    test      Test model.
    retrain   Retrain model.
    predict   Run predictions model.
    simulate  Simulate test data.

Stephen Richer, NHS England (stephen.richer@nhs.net)
```

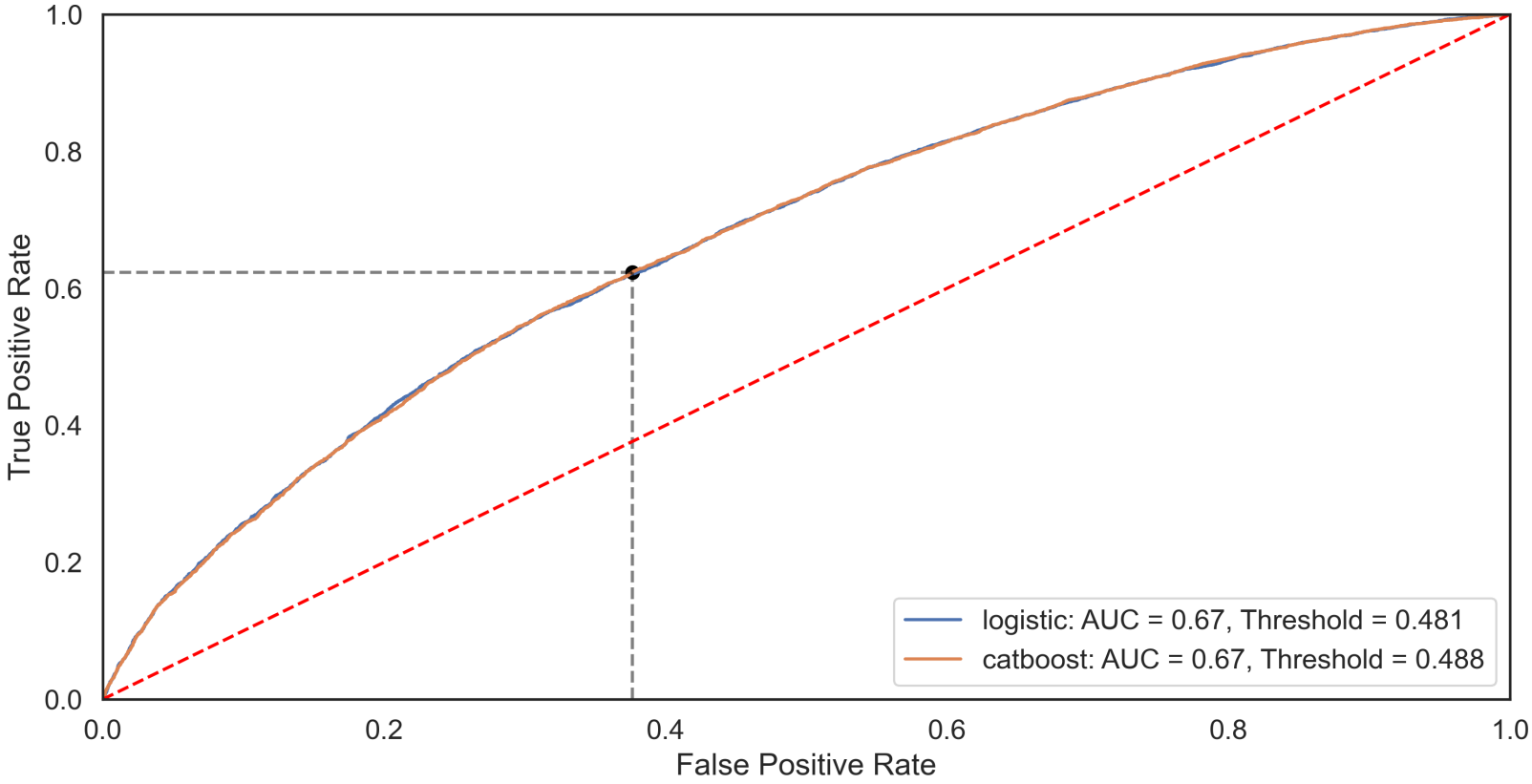
Example Output: Calibrated Probabilities



Example Output: Precision-Recall Curve

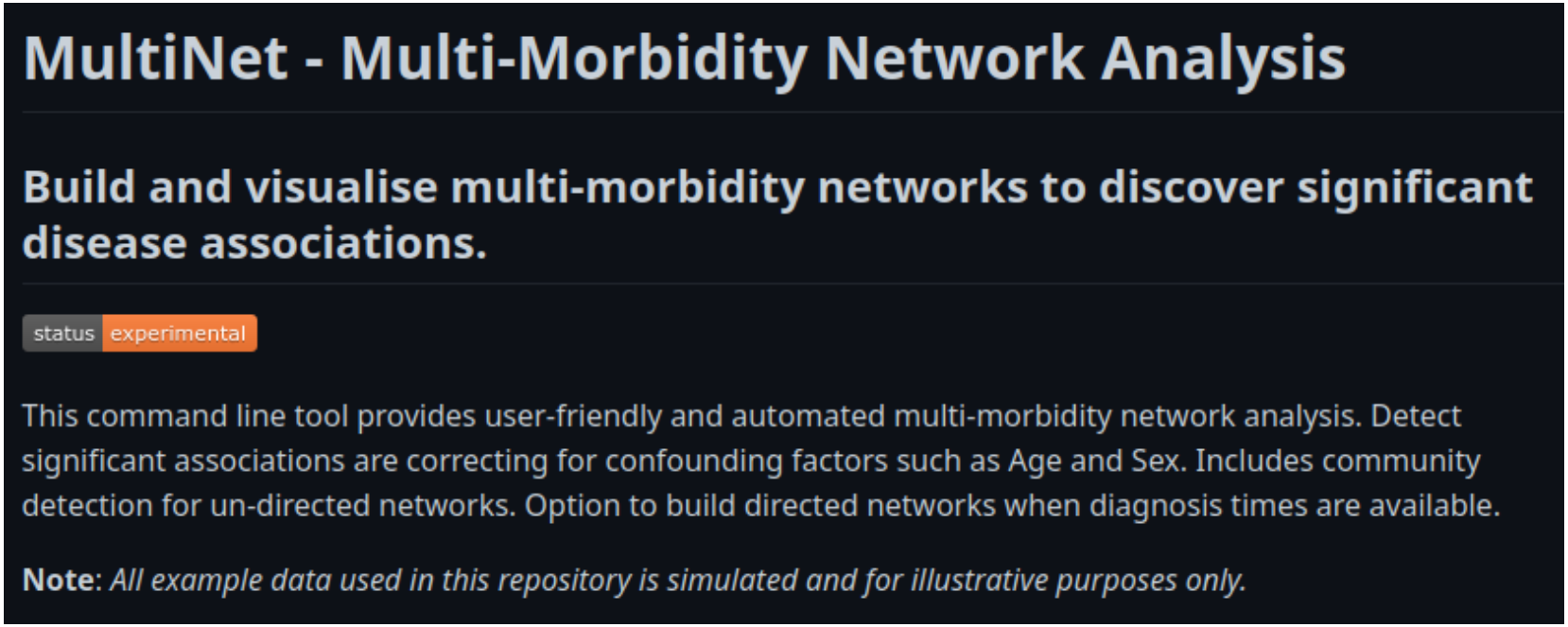


Example Output: ROC Curve



3) MultiNet

Build and visualise multi-morbidity networks to discover significant disease associations.



Full Documentation Available At:
https://github.com/nhsx/morbidity_network_analysis

MultiNet: Key Features

- Multi-morbidity represents the co-occurrence of two of more chronic medical conditions.
- A substantial proportion of patients have multi-morbidities and the associated costs of treating such patients are often significantly higher.
- Multi-morbidities can be studied using network analysis to identify significant disease associations and pathways.
  - The methodology for simple network analysis is established [Aguado et al., 2020].
  - However, the approaches remain inaccessible to non-exert users – simple interfaces for performing these analyses are not widely available.
- MultiNet is a user-friendly command-line utility for generating multi-morbidity networks and discovering significant associations between pairs of diseases.
- A full example workflow, with simulated data, is available at the GitHub repository.

MultiNet: Command Line Interface

```
(base) jovyan@ceb90354e88b:~/work$ multinet --help
usage: multinet [-h] [--version] [--verbose] Commands ...

MultiNet - Multimordibity Network Analysis

options:
  -h, --help  show this help message and exit
  --version   show program's version number and exit
  --verbose   verbose logging for debugging

required commands:

  Commands    Description:
    process   Pre-process data and compute edge weights.
    network   Build and visualise network.
    simulate  Simulate test data.
    enriched  Estimate morbidity enrichment by strata.

Stephen Richer, NHS England (stephen.richer@nhs.net)
```

MultiNet: Example Configuration

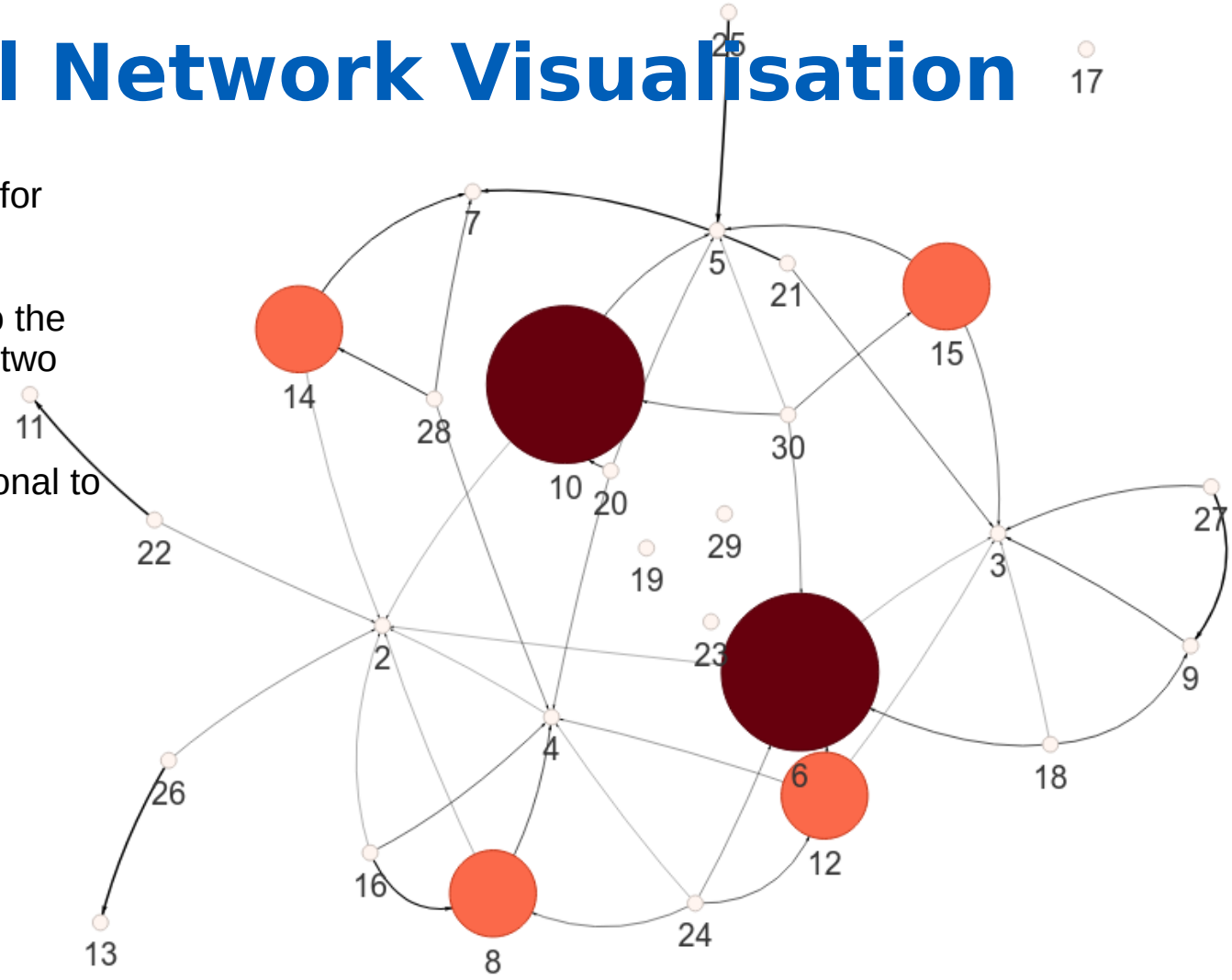
```yaml
input: MultiNet-data.csv
edgeData: MultiNet-processed.csv.gz
networkPlot: MultiNet.html
codes:
    code1: time1
    code2: time2
    code3: time3
    code4: time4
strata:
    - Age
refNode: 30
maxNode: 10
wordcloud: MultiNet-wordcloud.svg
fromRef: true
excludeNode: 1
enrichmentPlot: MultiNet-enrichment.svg
enrichmentNode: 1
demographics: Age
seed: 42
```

- MultiNet is configured via a single YAML file that describes input data and paramers.

- The command `multinet simulate` generates a full example dataset with configuration file.
  - The dataset describes directional relationships between numeric factors (e.g. 12 -> (2, 3, 4, 6)).
  - The simulated data allows users to explore functionality and serve as a the template for real-world clinical data.

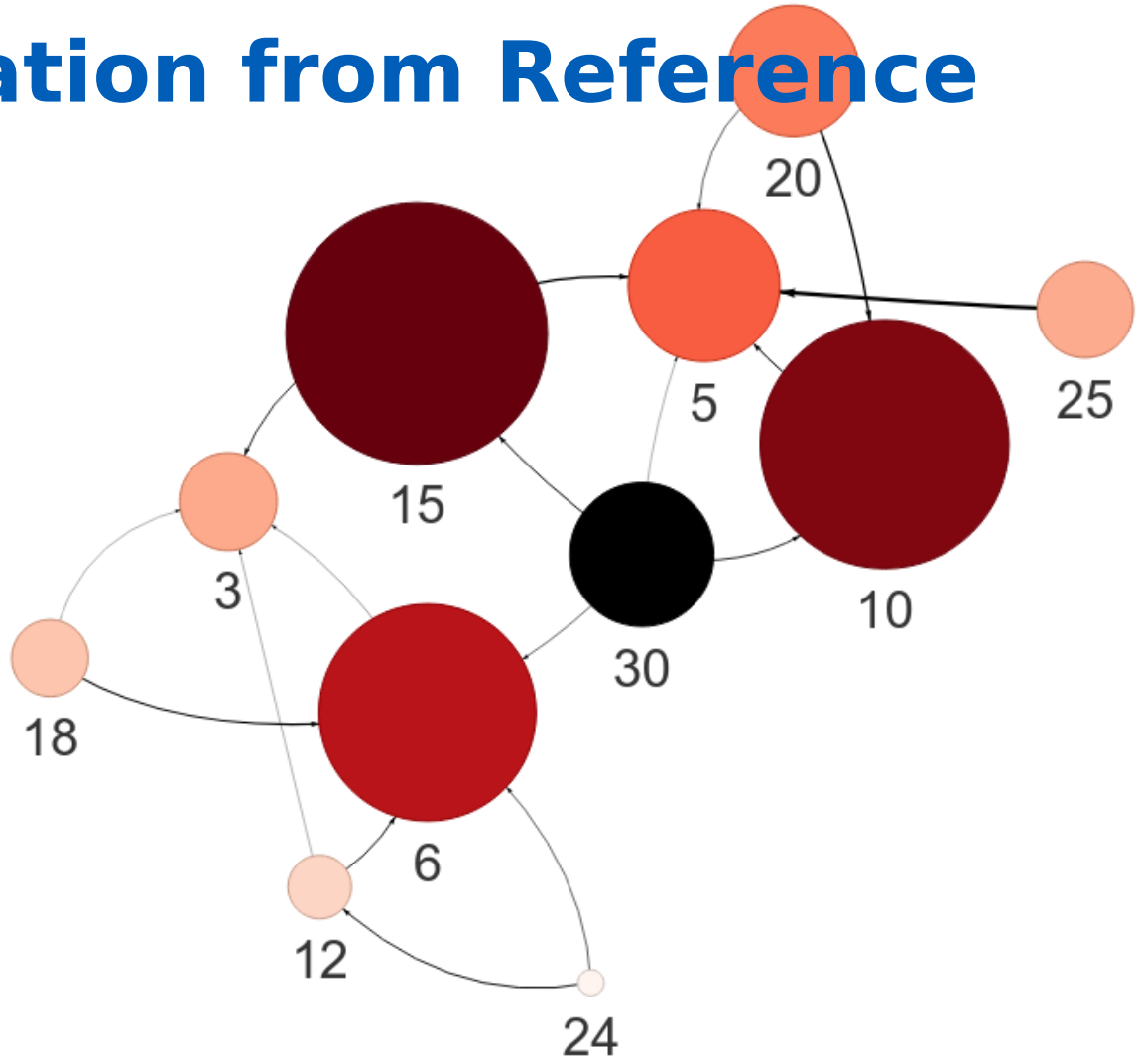
Example: Full Network Visualisation

- Small networks may be suitable for visualisation in their entirety.
- Edge thickness is proportional to the strength of association between two nodes.
- Node size and colour is proportional to betweenness centrality.



Example: Visualisation from Reference

- A reference node(s) is usually for focussing or graphs of larger networks.
- Edge thickness is proportional to the strength association between two nodes.
- Node size and colour is scaled to the distance the reference node(s).



Example: WordCloud from/to Reference

- A WordCloud is alternate approach to visualising node association from reference.
- For directional graphs, the WordCloud is configured either *from* the reference node or *to* the reference node.
  - from:
    - e.g. 30 -> (15, 10, 6, 5, 3, 2)
  - to:
    - e.g. (120, 90, 60) -> 30
- Word size is scaled to the distance from the reference node(s).