→ Recently Established Centre of Excellence for Analytics

→ 21 academic staff, RFs and PhD Students

→ In the UK, partnerships with Co-op, Boots, NHS, Olio, BBC, ONS

→ Core Expertise is in use of "Big Data" + Machine Learning" for Social Good

- Diagnosing disease with shopping data
  - https://www.nlab.org.uk/project/shopping-data-disease/

- Donating personal transactional data for research
  - https://www.turing.ac.uk/research/research-projects/donating-personal-transactional-data-research

CENTRE FOR DOCTORAL TRAINING

Forecasting local COVID-19/Respiratory Disease mortality via national longitudinal shopping data: the case for integrating digital footprint data into early warning systems

J. Goulding (University of Nottingham), E. Dolan (University of Nottingham), G Long (University of Nottingham), J. Flanagan (Imperial College London), J. Harvey (University of Nottingham), L. Tata (University of Nottingham), A. Skatova (University of Bristol)

Elizabeth Dolan, PhD Candidate, *Horizon Digital Economy Research* elizabeth.dolan@nottingham.ac.uk

Engineering and Physical Sciences Research Council

What is respiratory disease? Respiratory Disease ICD 10 coding: **J00–J99**

Infections and diseases of the lungs and respiratory system https://icd.who.int/browse10/2016/en#/J00-J06
These can be due to **COVID-19**.

Disease	ICD-10 Codes	Respiratory Disease ICD-10: J00-J99	Notes
Asthma	J45-46	✓	
Chronic Obstructive Pulmonary Disease (COPD)	J40-47	✓	Includes Bronchitis, Emphysema
Lung Cancer	C33-34	✗	
Pneumonia and Influenza	J10-18	✓	
Respiratory Tuberculosis	A16-19	✗	J65 Pneumoconiosis is associated with tuberculosis
Cystic Fibrosis	E84.8-84.9	✗	

- COVID-19 led to unparalleled pressure on healthcare services, with improved healthcare planning respiratory diseases becoming a key concern.

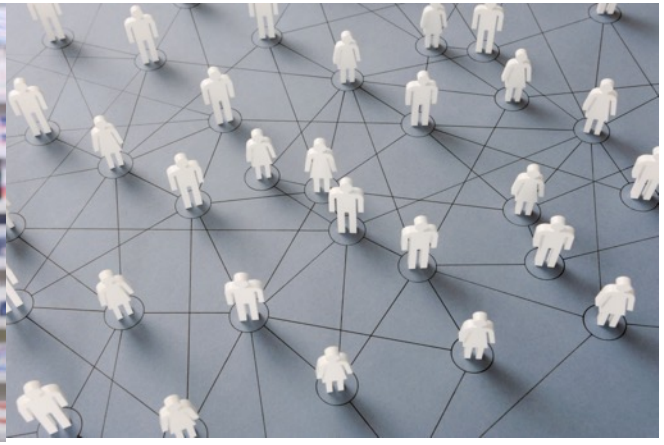
- We present the results of **two related studies** investigating the potential of using digital footprint data, in the form of non-prescription medication sales, to improve forecasting of weekly registered deaths of such diseases at local levels.

Study: A new experimental design to compare the use of sales data against other datasets used in the prediction of respiratory deaths

Applying Hofman et al.'s [1] recommendations on computational social science:

- Baseline model
- Out-of-sample testing
- Combine prediction and explanation - Here we use **a new AI Explainability Tool Model Class Reliance (MCR) for Random Forest Regressor**





[1] Hofman, J. M. et al. Integrating explanation and prediction in computational social science. Nature 595, 181-188 (2021).

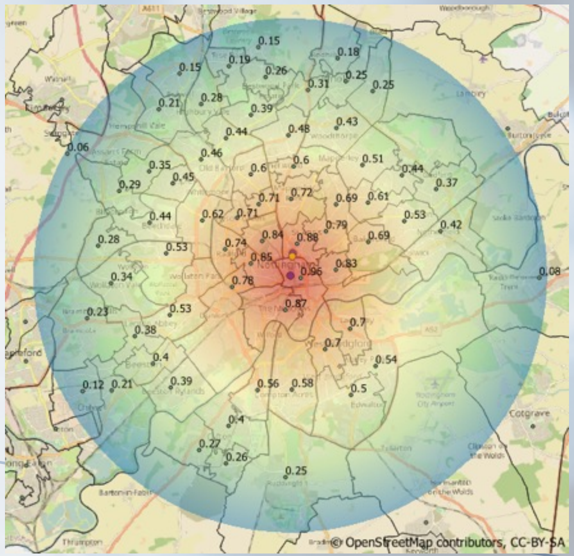
Explaining the Commercial Sales features

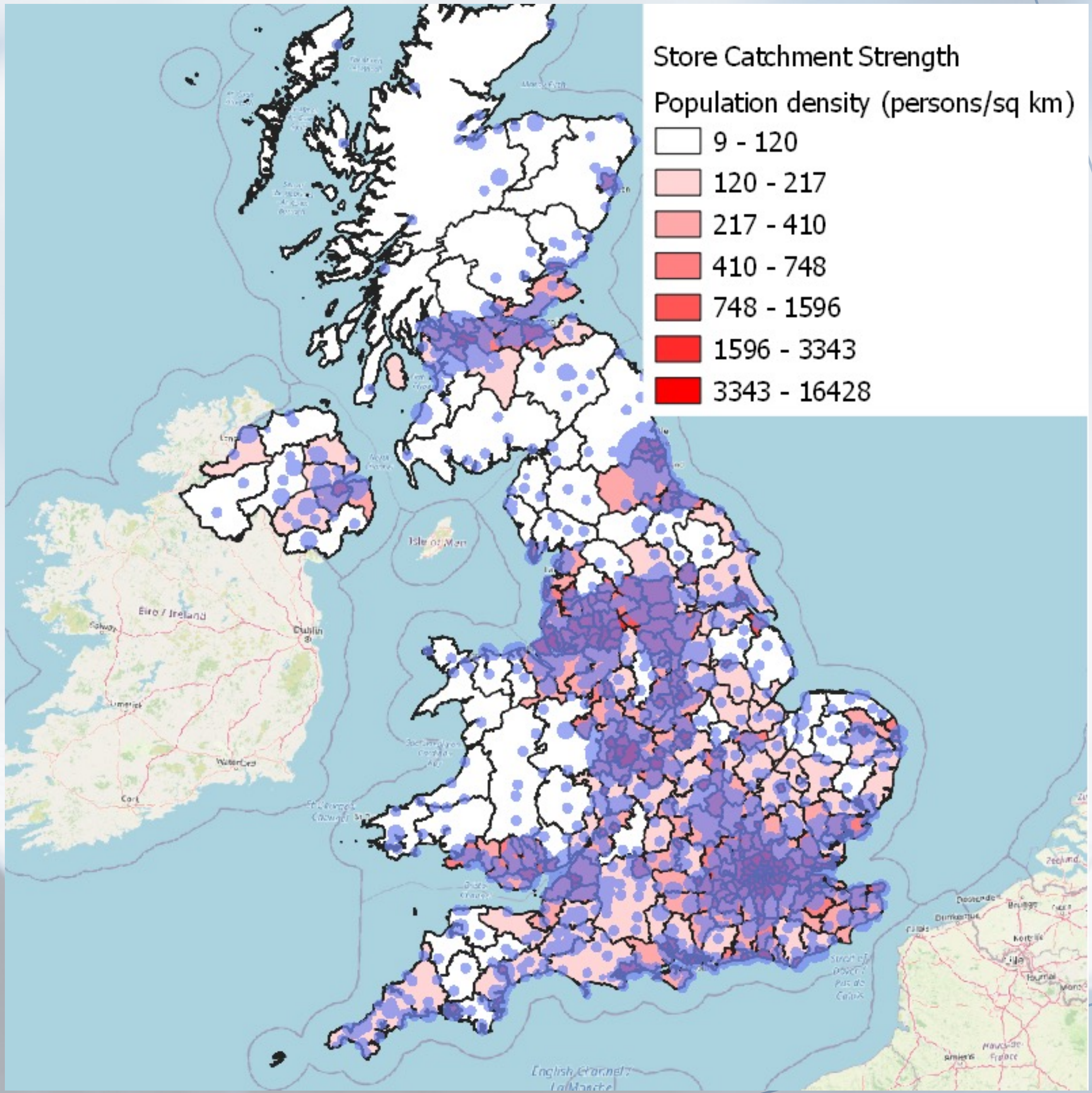
BASIC STATS:

Data using in models (2016 to 2020) is over 2 billion in store sale units in England (not including online)

For example over 14 million cough medicine sales

Sales are allocated from stores to 314 LTLAs in England





N/LAB

horizon
CENTRE FOR DOCTORAL TRAINING

Creating the PADRUS* model

***Predicting amount of deaths from respiratory disease using sales data**

Model type: Random Forest Regressor. Optimized using a time series cross-validation grid-search on training data to prevent over-fitting.

Target: Predict weekly respiratory deaths 17 days in advance for each of the 314 LTLA (Lower Tier Local Authority) areas in England from 18th March 2016 to 27th March 2020.

Input features: 56 features (static and dynamic including traditional variables associated with respiratory disease)

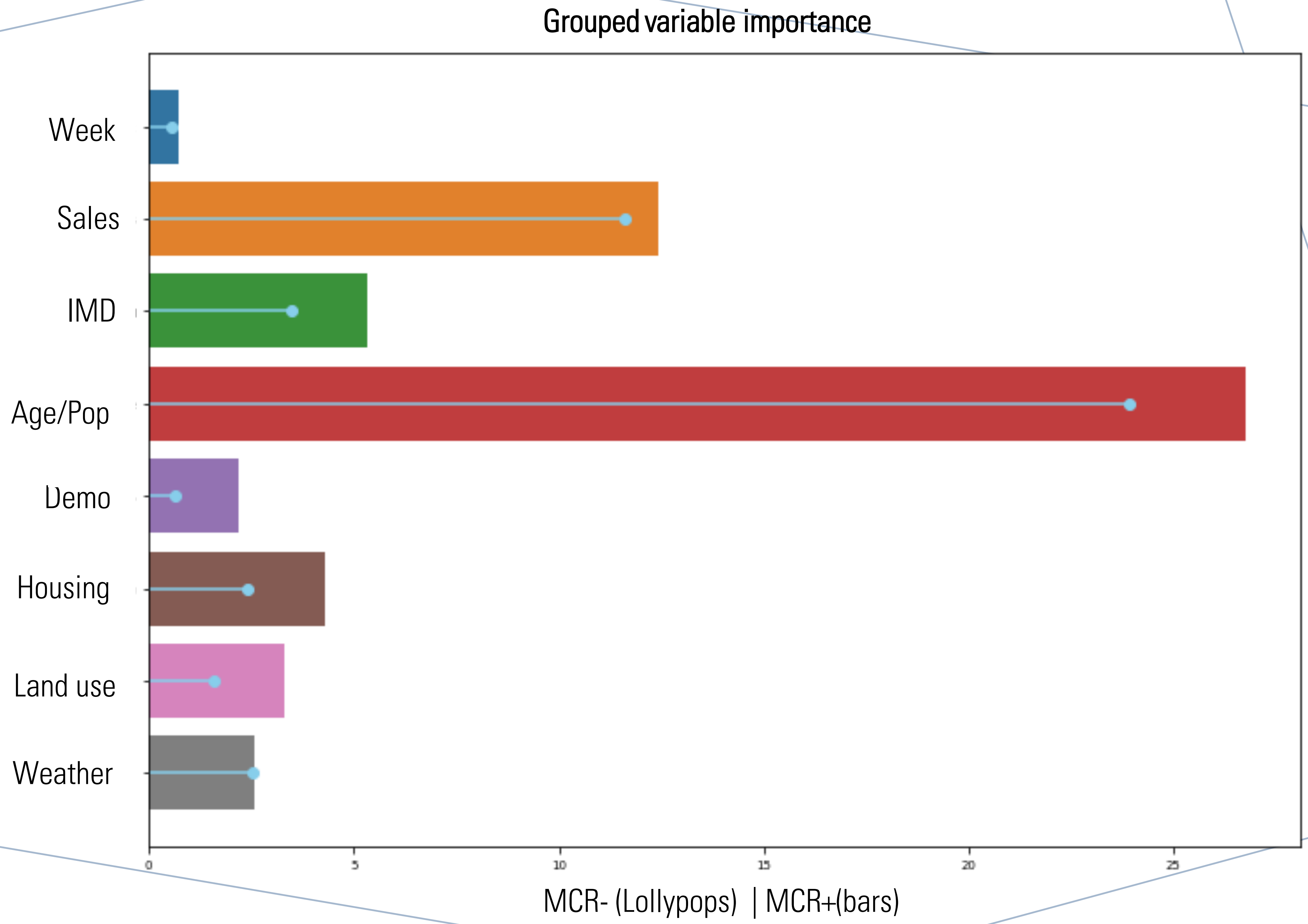
Training Datapoints: 45844

Out-of-sample Testing
Testing Datapoints: 20410

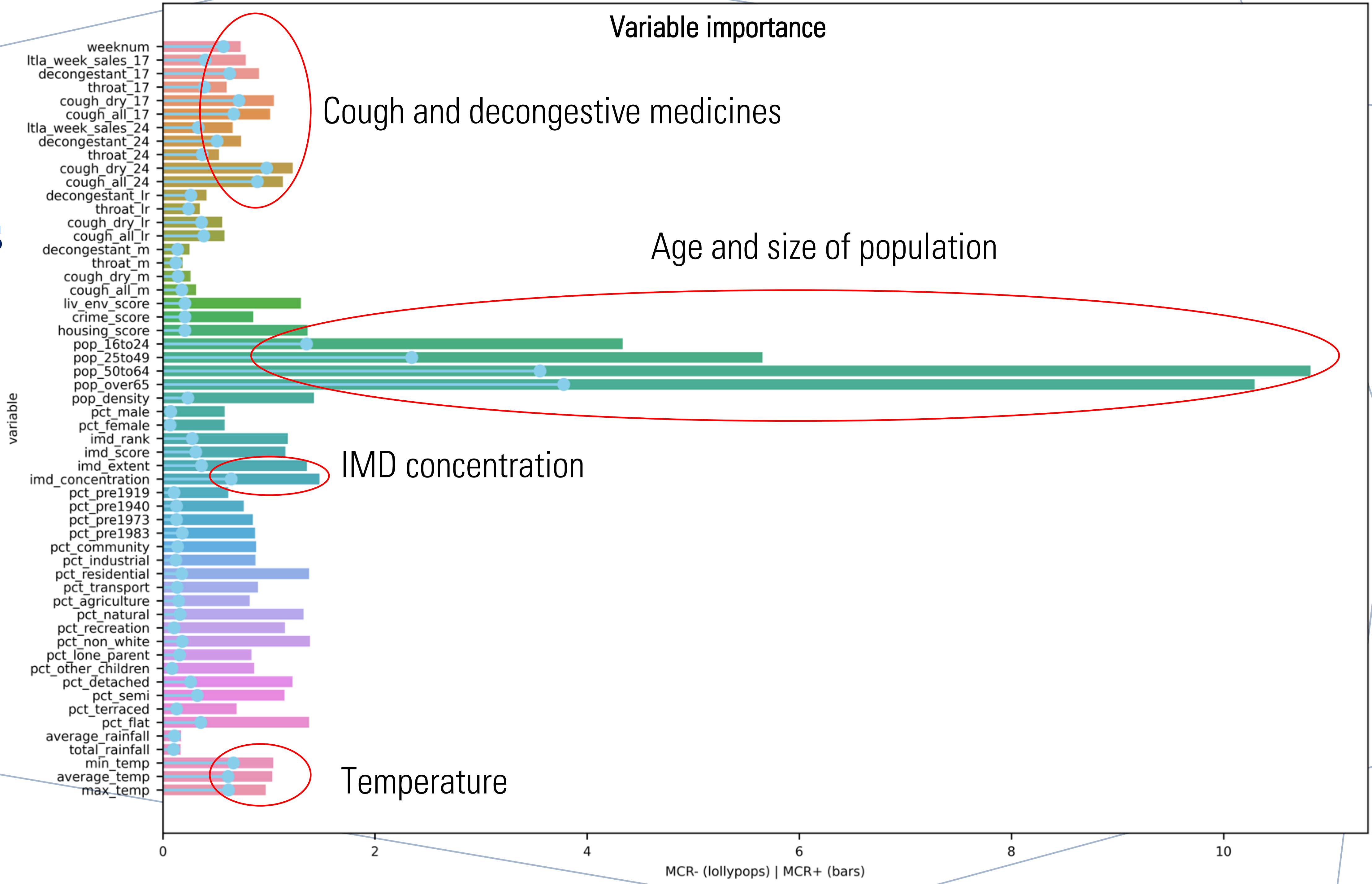
Results from predicting on test data:		
Mean Absolute Error	Root Squared Mean Error	R^2
2.39 (Baseline model 2.78)	3.42 (Baseline model 4)	0.78 (Baseline model 0.71)

With highest accuracy gains over models without digital footprint data (increases in R^2 between 0.09 to 0.11) occurring in periods of maximum risk to the general public.

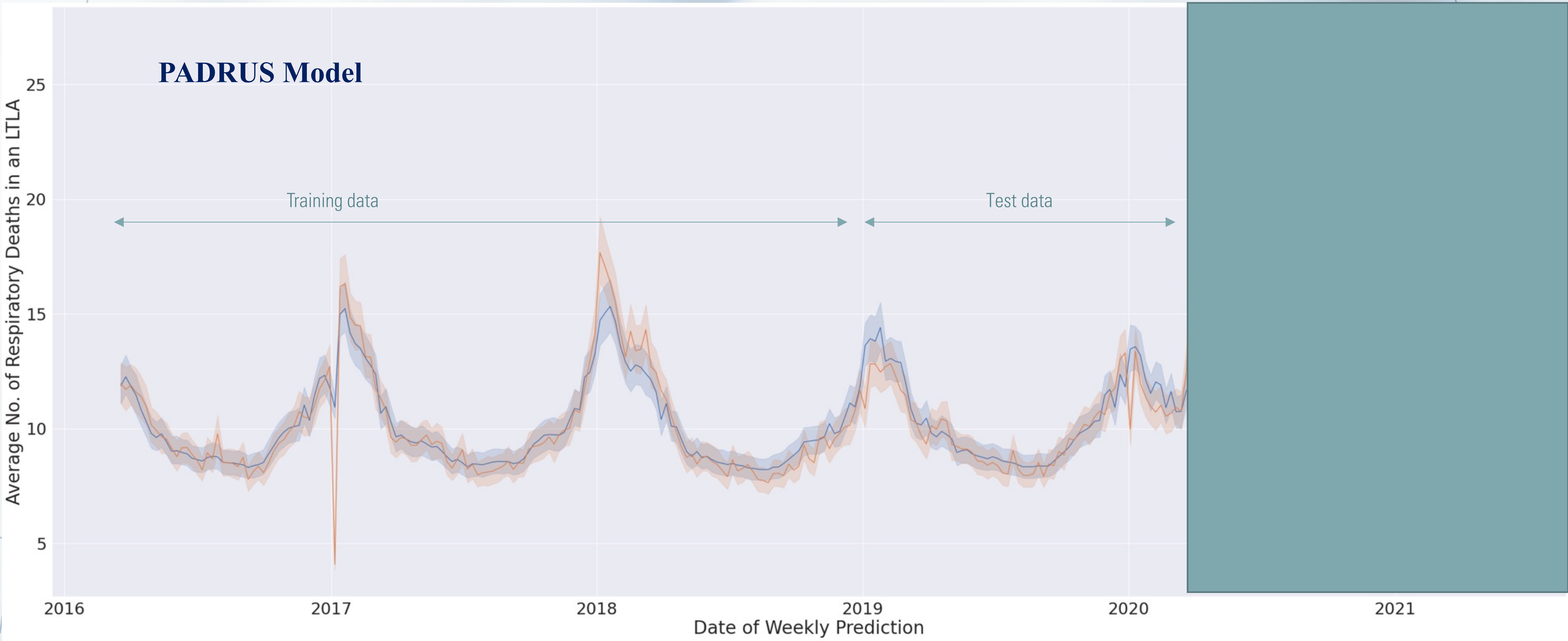
Group MCR PADRUS model



MCR PADRUS model (run on training data)…



Running the model during COVID-19

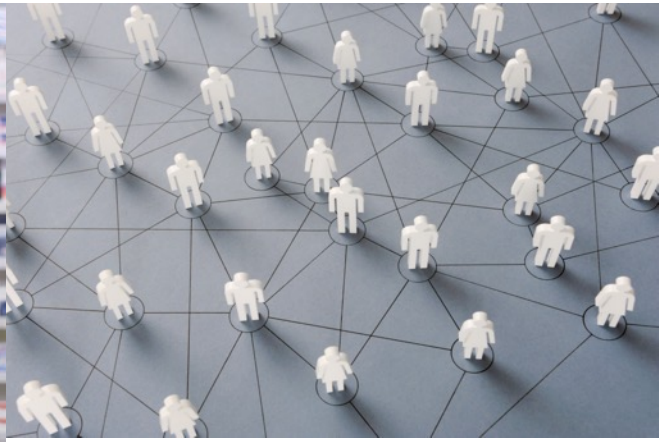


Study: Creating the CIVIC model: Predicting Covid-19 Impact on Vulnerable Individuals and Communities via health, deprivation, and transactional sales data

- Addressed the feature drift of sales variables due to government lockdowns

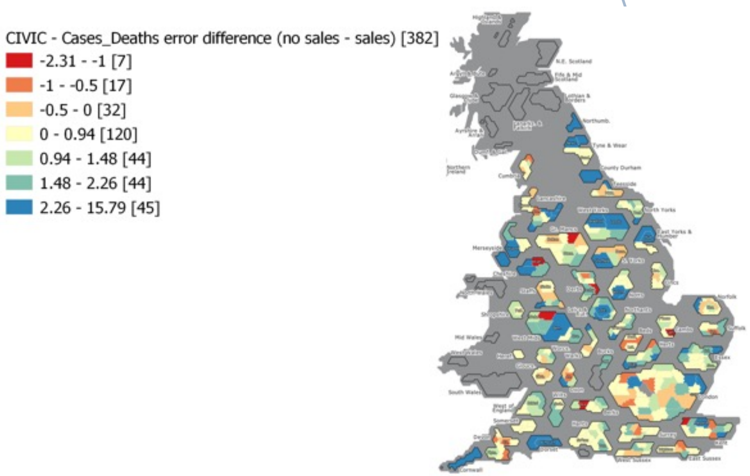
- Used additional data resources available in the pandemic,
  i.e. easily accessible and regularly updated COVID-19 test, case and mortality data,
  newly available mobility data





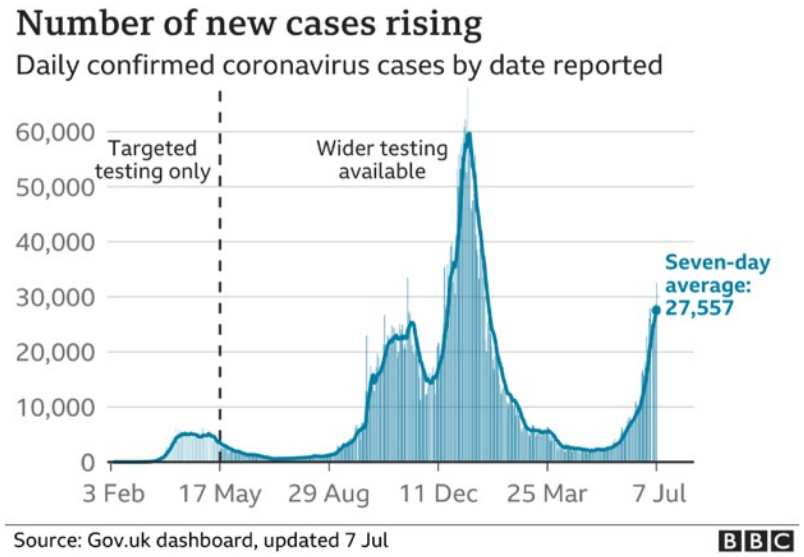
N/LAB
horizon
CENTRE FOR DOCTORAL TRAINING

- MODEL TYPE : A range of ML regression models (Linear, SVR, Random Forest & XGBoost) were developed and tested across the initial waves of the COVID-19 pandemic



- TARGET : Forecast of COVID-19 cases/deaths in 7/14/21 days at the 314 Lower Tier Local Authorities in England

- INPUT FEATURES: Grouped feature sets(e.g. cases, sales of over-the-counter medications, demographics, mobility, etc.) combined across multiple runs to test predictive performance and identify optimal set of predictors. (Retail sales attributed to LTLAs using simulated catchment model)



- Datapoints : Weekly time series covering period from April 2020 to June 2021. Split into two waves (April-July 2020, July 2020-March 2021)

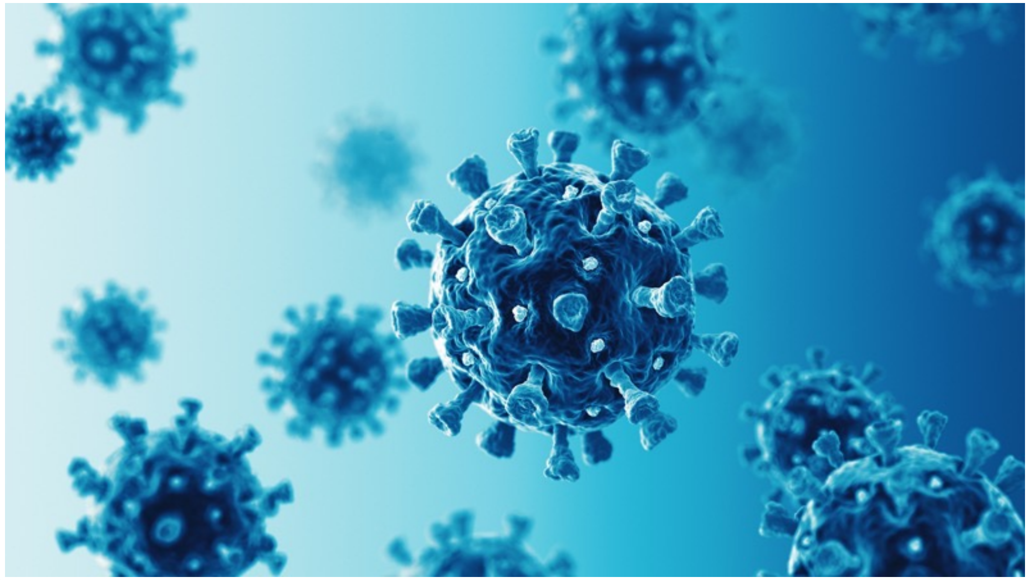
Results

- XGBoost models containing sales data optimally predicted the number of COVID deaths 21 days in advance (R2=0.68**), significantly outperforming models based on official COVID case data alone at local-area levels (R2=0.44**)
- Demographics and mobility inputs less useful than retail sales.

(Without demographics – 0.61**)

Conclusions:

- Over-the-counter medication purchases related to management of respiratory illness are correlated with registered deaths at a 17-21 day window.
- Results demonstrate the potential for sales data to support early warning population health mechanisms at local area levels.



Any questions?

Forecasting local COVID-19/Respiratory Disease mortality via national longitudinal shopping data: the case for integrating digital footprint data into early warning systems

@3lizabeth_Dolan

Useful Links:

https://www.nlab.org.uk/project/shopping-data-disease/

https://www.researchsquare.com/article/rs-2226531/v1

https://github.com/nhsx/commercial-data-healthcare-predictions