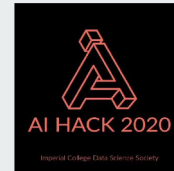




# Data-driven analysis to improve quality of life

George Hajivassiliou, Andreas Maos, Marios Kassapis, Kyriacos Theocharides

March 2020

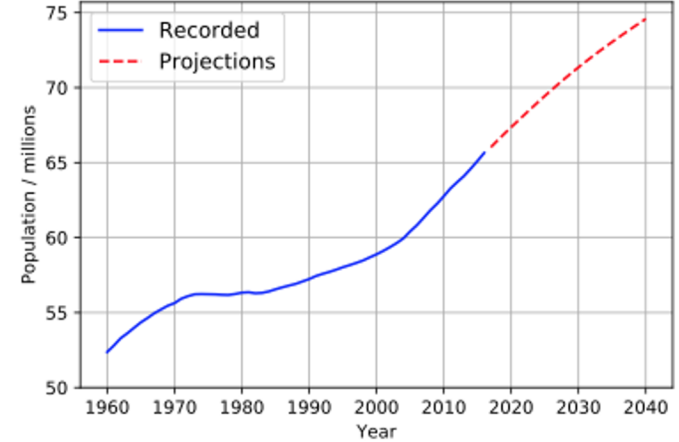


# Motivation

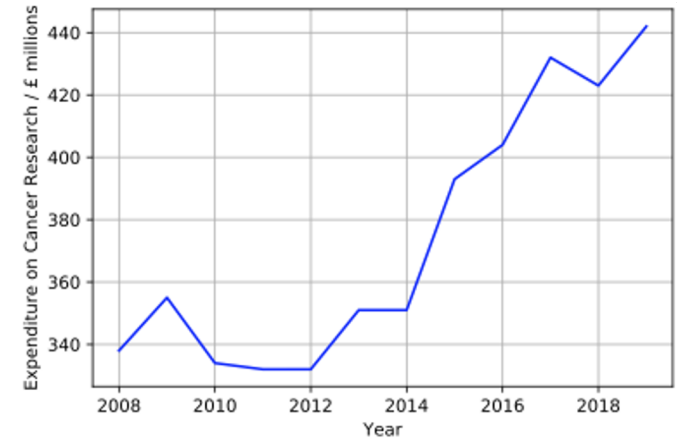
- Increasing burden from ageing population
- Elderly much more prone to disease & virus

❑ Simply increasing lateral expenditure is not the optimal solution

→ **How does the NHS prepare for this?**



[1]



[2]

[1] Source: Office for National Statistics

[2] Source: Annual Report and Accounts, Cancer Research UK



**Can we use existing data to predict  
future disease incidence?**



# Do we have adequate datasets?

## We had some:

- ONS - Consumer trends time series
- Cancer incidence and mortality rates ([www.cancerdata.nhs.uk](http://www.cancerdata.nhs.uk))

## And made some:

- Web-scraped drug prescriptions for cancer and dementia (<https://openprescribing.net/api/>)



# Probed Dementia using sales of Rivastigmine

Basic Info:

- Commonly administered by NHS to treat Alzheimer's disease; the most common form of dementia [3]
- May be prescribed by a GP on advice of a specialist [4]
- Also known as Exelon (brand name)
- Donepezil and Galantamine are also used and have very similar effects[3,4]
- Rivastigmine was chosen as it had the most complete, available dataset

[3] Source: NHS.UK, Dementia Treatment

[4] Source: NHS.UK, Alzheimer's Treatment




# Methods

- Data wrangling using R to generate ML-ready data frames
- Used linear regression and long short-term memory (LSTM) to make predictions
- Features used:
  - Tobacco, alcohol, narcotics, health, condition-specific drug prescriptions
- Predicted dementia prescriptions until 2020



# Linear Regression for Time Series



Tobacco	Alcohol	Drugs	Health	Dementia Probe
0.123	0.124	0.567	0.543	0.12
0.146	0.204	0.755	0.621	0.33
0.153	0.243	0.801	0.721	0.24
0.165	0.255	0.769	0.711	0.51



# Linear Regression for Time Series

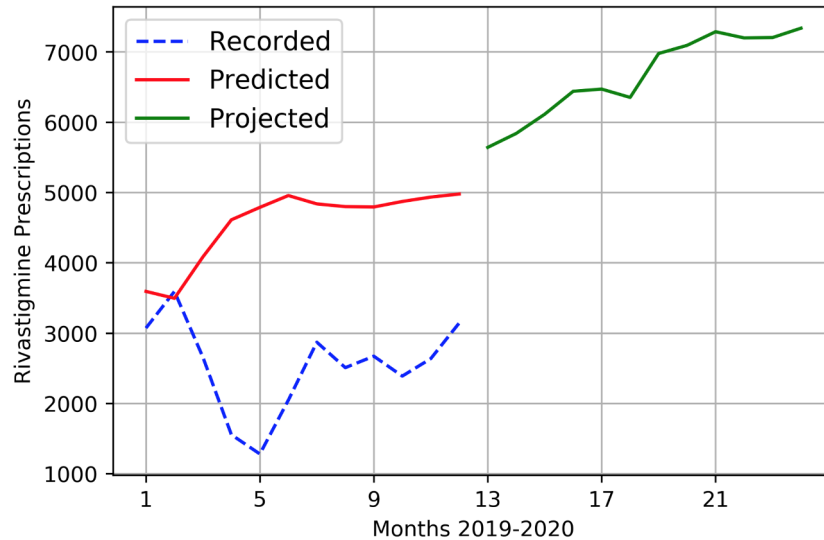
time ↓	Tobacco	Alcohol	Drugs	Health	Dementia Probe
	0.123	0.124	0.567	0.543	0.12
	0.146	0.204	0.755	0.621	0.33
	0.153	0.243	0.801	0.721	0.24
	0.165	0.255	0.769	0.711	0.51



# Linear Regression for Time Series

Linear Regression for Time Series								Dementia Probe	
time ↓					Tobacco	Alcohol	Drugs	Health	0.12
	Tobacco	Alcohol	Drugs	Health	0.123	0.124	0.567	0.543	0.33
	0.123	0.124	0.567	0.543	0.146	0.204	0.755	0.621	0.24
	0.146	0.204	0.755	0.621	0.153	0.243	0.801	0.721	0.51
	0.153	0.243	0.801	0.721	0.165	0.255	0.769	0.711	
	0.165	0.255	0.769	0.711					
t - 2					t - 1				t

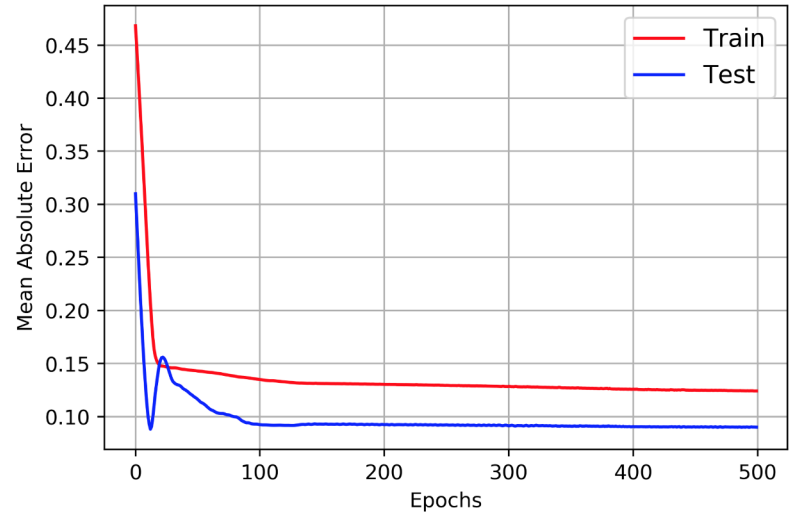
# Results: Linear Regression



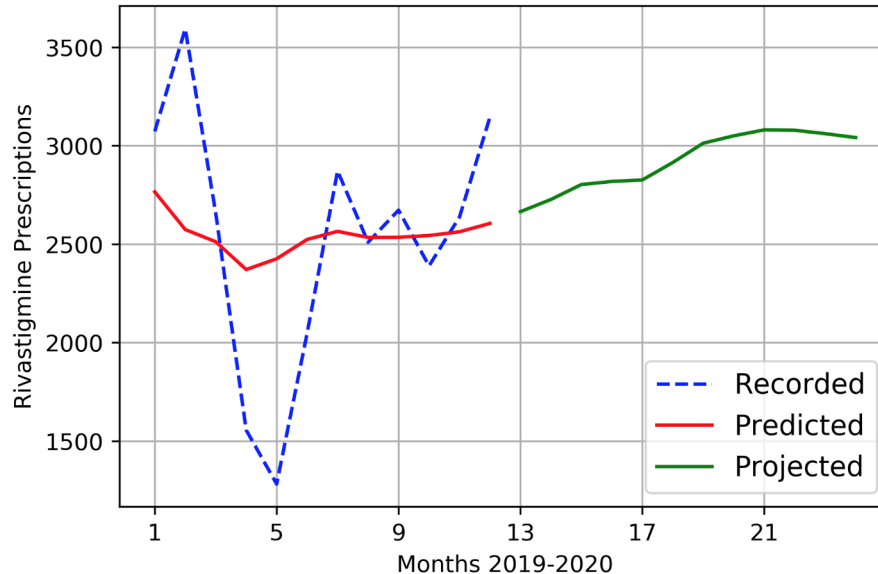
- Mean Absolute Error on (Normalised): 0.353
- Model is unable to capture the complexity of the data

# Long-Short Term Memory Networks (LSTMs)

- Neural Networks suited for time series analysis
- Used data from multiple times steps to predict a single time point in the future (Same arrangement as Linear Regression)



## Results: LSTM



- Mean Absolute Error on (Normalised): 0.0909
- Much Lower Error
- Reasonable Prediction
- We predict a slight increase in Rivastigmine predictions over 2020.



# Conclusion

- Combined existing datasets with web scraping to build a machine learning ready dataset
- Manipulated that dataset to predict using multiple timesteps
- Applied Linear Regression and LSTM
- Predicted dementia trend up to the end of 2020