Deep Learning for Predicting Medical Conditions

Australian Government
Department of Health

Australian National University

Motivation

- Currently it is challenging to identify medical conditions using MADIP because this data source does not contain diagnosis information for the vast majority of individuals.
- There is survey data that has been linked to MADIP (NHS & SDAC), but they only contain information for a very small portion of the population.
- However MBS and PBS are more up-to-date datasets compared to hospital data.
- We wanted to know if we could use a machine learning approach to predict if someone has a medical condition.
- This approach would mean that targeted patient cohorts are easier to identify.

Background

- Recent advances in deep learning have drastically improved ability to solve various natural language processing tasks. The sequence of item numbers that different individuals utilise can be seen as a form of language that can be processed using natural language algorithms.
- The results presented today are preliminary and as such are not for further distribution. We have other projects that are examining this issue using other methodologies.
- If you are interested in learning more about this project please contact either myself (richard.hurley@health.gov.au) or Dr Allison Clarke (allison.clarke@health.gov.au).

Aims

- Determine the extent to which medical conditions can accurately be predicted solely from PBS and MBS claim history.
- Evaluate how much deep learning models can improve performance.

Data

- **Inputs**
  - PBS & MBS data from MADIP (aggregated to monthly level from 2011 – 2016)

- **Outputs**
  - Main condition identified on Survey of Disability, Aging and Carers (SDAC) in 2015

Data Modelling

- We treat the person’s PBS and MBS history as a sequence of discrete tokens.

- Each PBS/MBS item is associated with a unique id.

Models

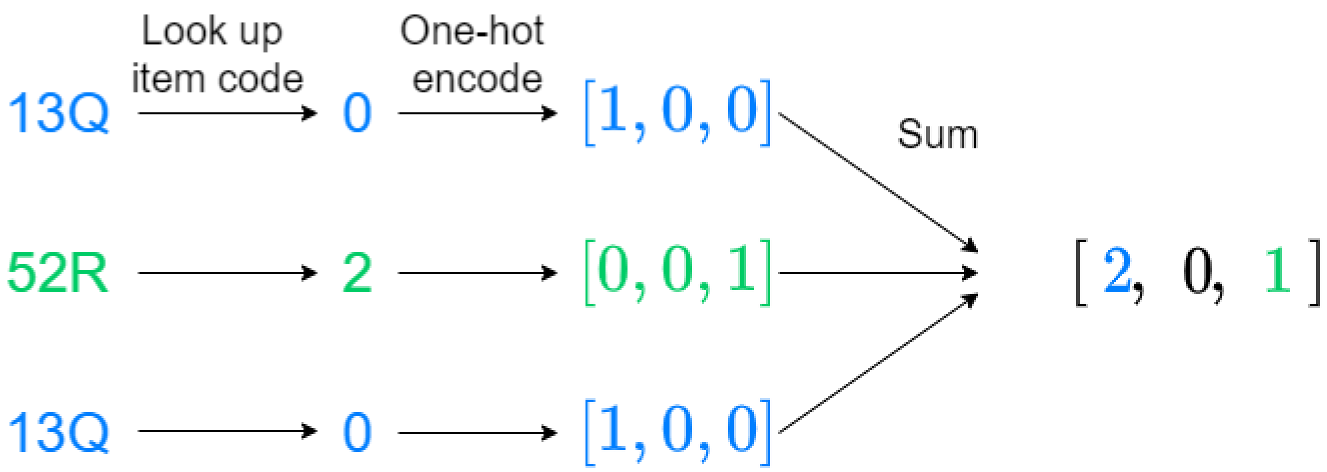
- We used four models:
  - Linear model (baseline model)
  - Recurrent Neural Network (RNN)
  - Convolutional Neural Network (CNN)
  - Transformer (TN)

Australian Government
Department of Health

Australian National University

Baseline Model

- To serve as a simple baseline, we train a linear classifier.
  - Each item is represented with one-hot encoding.
  - A person’s history is represented as the sum of the encodings of each item in their history.
  - Linear classifier is fit to people’s history encodings.



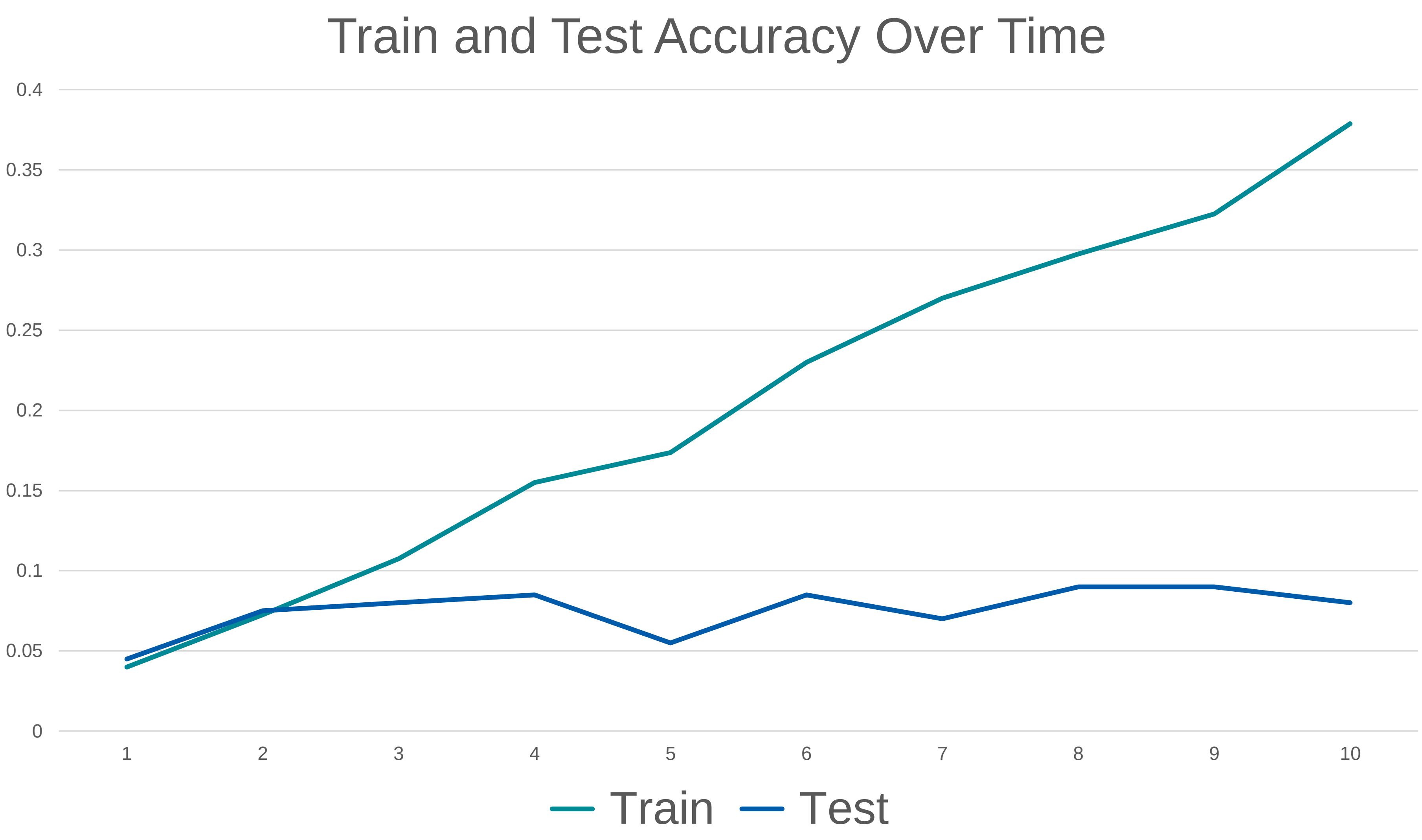
Deep Learning Models

- The baseline model does not take into account the *relationships between MBS and PBS* items or the *sequential nature of the data*. We therefore also train a variety of Deep Neural Network models to see if capturing this information helps.

Results - Initial

- We use a random sample of 1000 people from the SDAC, split 90%/10% for training/testing.
- Accuracy reported is highest achieved within 10 epochs of training.

Model	Accuracy
Linear	**20%**
RNN	**14%**
CNN	**18%**
TN	**9%**



Australian Government
Department of Health

Australian National University

Discussion of Results

- These results indicate that the deep learning models are overfitting heavily. This could be because
  - not enough regularization.
  - not enough training data.

Results – Full SDAC

- We use the entire linked SDAC population of around16,000, split 90%/10% for training/testing.
- Accuracy reported is highest achieved within 10 epochs of training.

Model	Accuracy
Linear	**29.2%**
Linear+Self-Attention	**29.4%**
CNN	**28.4%**

Results – Easiest Conditions to Predict

Condition	F1 Score
ADHD	**0.593**
Autism	**0.591**
Diabetes	**0.582**
Asthma	**0.559**
Epilepsy	**0.516**
Hypertension	**0.457**

Results – Single Condition Classification

- For each of the most common conditions we train a linear model to predict whether or not a person has that condition.

Condition	Accuracy
Anxiety	**66.0**
Depression	**78.5**
Hypertension	**73.1**
Asthma	**73.5**
Diabetes	**80.5**

Future Work

- Investigate the impact that the size of the training dataset has
  - Train models on larger subsets of PBS/MBS
- Use National Health Survey (NHS) data instead of SDAC
  - Unlike SDAC, NHS is representative of general population
- Compare results using this methodology to the other projects that use different methods.

- Self-supervised pre-training.

Self-supervised Training

- We have a small amount of labelled data (only those included in NHS or SDAC survey). But we have an enormous amount of unlabelled data
- We can use self-supervised learning methods (such as training to predict which item code comes next in a sequence) to improve performance.
- Increased computation will be crucial, so we will move the project to ABS cloud datalab.

Thank you

Australian Government
Department of Health

Australian National University