Predicting mortality in both diabetes and open-source clinical datasets from free text entries using machine learning (natural language processing)

CAR Sainsbury [1], A Conkie [2]
1 glucose.ai Research Group, Department of Diabetes, Gartnavel General Hospital, Glasgow, Scotland, UK
2 Red Star Consulting, UK

http://glucose.ai
@csainsbury
c.sainsbury@nhs.net

Background

To date, the focus of predictive modelling in both diabetes and clinical medicine as a whole has been on the use of numerical features. With the wide adoption of electronic health record (EHR) systems, text - both structured and unstructured - has become available for analysis. Historically unstructured (free) text has often been excluded from analysis due to the technical difficulty of extracting useful insight.

Within the Scottish healthcare system, a single clinical database covering 99.7% of all individuals with diabetes (SCI Diabetes) is used nationally. This dataset contains demographic information, numerical features over time (including weight, BMI, blood pressure, HbA1c and all relevant laboratory results), and free text entered by clinicians (including podiatrists, dieticians, specialist nurses and medical staff) predominantly as the record of a clinical contact.

Aims

Our group (and others within Scotland), have previously developed a number of models predicting clinical outcomes of interest using numerical data using traditional statistical - and a variety of machine learning techniques - using the SCI Diabetes dataset. The fundamental aims of the current work were to:

(i) explore approaches to extract insight / information from unstructured free text records of clinical consultations, and

(ii) to explore the extent to which this information is additive to numerical data when constructing predictive models. The results presented here relate to aim (i).

We aimed to test the utility of a number of technical approaches to the task of semantic analysis of clinical free-text entries. The outcome chosen was all-cause mortality due to its unequivocal nature, and completeness of recording of the outcome within the datasets of interest. Approaches were developed on data from our national diabetes dataset, and on an open source clinical dataset (MIMIC-III) [1]. In the presented results, analysis was constrained to text alone, to explore the level of extractable information contained solely within text data.

Methods

Data was prepared for analysis using R [2], and analysis code was written in both R and Python [3].

Diabetes dataset. An analysis period of 3 years was defined for each individual, during which all recorded unstructured clinical free-text data were extracted. The majority of this data represented recording of clinical consultations. Structured content was removed. Mortality status at 1 year (following the end of the text analysis period) was identified from linked data. Data were pre-processed - numbers and special characters were removed, and all text entries were concatenated into a single string. Data were divided randomly into training/validation and test sets with a 0.8:0.2 ratio. The training/validation set was further randomly divided using an 0.8:0.2 ratio. Dimensionality reduction was performed using embedding, and a combined convolutional and recurrent (LSTM) neural network was trained on the training subset for 20 epochs. Class imbalance was managed by applying class weights. A prediction of outcome was made on the withheld test set using the trained model, and area under receiver operator characteristic curve (AUROC) was calculated. k-fold (k=4) cross validation was employed to explore optimum hyperparameter values.

MIMIC-III dataset. A similar methodology was applied, with dataset-specific pre processing. Text was extracted from MIMIC-III discharge documentation, with death at 1 year remaining the target for prediction. A similar CNN/LSTM model was applied, as well as an implementation of FastText [4].

Results

Diabetes dataset: 53954 individuals with data were identified. 2292 deaths were recorded at 1-year post analysis. AUROC of model predictions when applied to withheld test set was 0.62.

MIMIC-III: 11518 individuals identified, with 2045 deaths at 1 year. AUROC applied to withheld test set 0.86. AUROC on an expanded dataset (n=46070; 8070 deaths) was 0.87 (Figure 1). Risk of death at 1 year for each decile of prediction is shown in Figure 2.

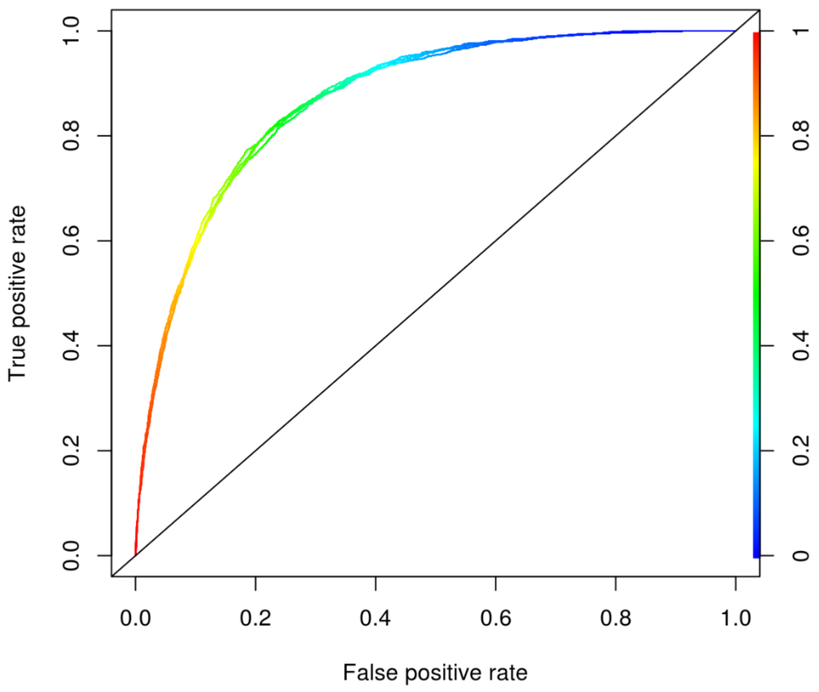

Figure 1. Receiver Operator Characteristic Curve, prediction of death at 1 year. n = 46070

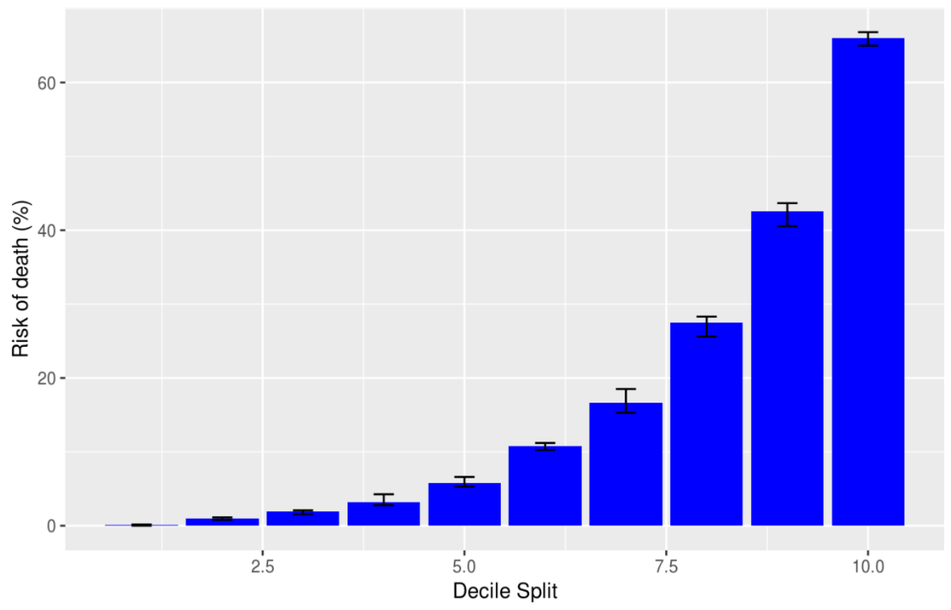

Figure 2. Risk of death at 1 year by decile of prediction, illustrating discriminating ability.

Summary and Conclusions. By learning from clinician's summaries, natural language processing has the potential to leverage the clinical understanding of multiple clinicians across multiple contact episodes. We have demonstrated that these models perform reasonably in 2 disparate datasets to predict mortality in the short term. Subsequent work has improved performance on the diabetes dataset to an AUROC of 0.76. Using a similar technique, models of this may additionally be trained on outcomes such as response to particular therapeutic agents, or development of complications, and may either be used stand-alone, or to provide input features to existing predictive models built on numerical information.

References

1. DOI: 10.1038/sdata.2016.35.
2. https://www.r-project.org/
3. https://www.python.org/
4. https://fasttext.cc/