170

Predicting Progression of Chronic Kidney Disease Using Machine Learning

Dr Thomas Oates[1], Dr Rohini Mathur[2], Dr Neil Ashman[1], Dr Sally Hull[2]
[1]Barts Health Nhs Trust, London, United Kingdom, [2] Institute of Population Health Sciences, Queen Mary University of London, London, United Kingdom

Introduction
Analysis of routinely collected data stored in primary care electronic health records (EHR) may allow development of tools to predict adverse outcomes in patients with chronic kidney disease (CKD). Concurrent increases in the amount of data stored in EHRs, and the computational power and sophistication available to analyse these data, has raised interest in the use of machine learning (ML) techniques to predict patient outcomes. We examined if ML methods might be better than a standard statistical technique at predicting progression of CKD in a large cohort derived from a primary care EHR.

Methods
We used a previously published dataset (Mathur et al BMJ Open 2018; 8: e020145) of 25 to 85 year old individuals with baseline diagnoses of Type 2 Diabetes Mellitus and CKD (eGFR < 60ml/min). The outcome was an eGFR fall of 5ml/min or greater in a single year. Logistic regression was used as the standard statistical technique, and compared to the ML modeling techniques of random forests (RF) and radial support vector machines (SVM). Highly correlated variables and variables with near zero variance were removed in pre-processing leaving 21 predictor variables for modeling. K-nearest neighbour imputation was used for missing values and predictor variables were centred and scaled prior to analysis. All models were trained to optimise the Receiver-Operator Characteristic (ROC) on 75% of the dataset. The remaining 25% was used for model testing. All analysis was performed using the R statistical computing language.

Results
The dataset comprises 33,171 patient years of CKD follow up from 6,631 individuals. 7,119 (21.5%) of observations showed the outcome of CKD progression of >5ml/min/year.

The logistic regression model predicted CKD progression with an accuracy of 77.5%, which was similar to if no information were known above the proportions of progression/non-progression in the data (the no information rate (NIR)), (P-value for accuracy above NIR 0.92). Of the two ML models, RF outperformed SVM in prediction of CKD progression (accuracy of 80.1% and 79.0% respectively. P-values for accuracy above NIR 3.9x10-6 and 0.034 respectively. See Table).

Positive predictive values for ML methods were 0.66 for RF and 0.61 for SVM, comparing favourably to traditional prediction methods such as the Kidney Failure Risk Equation (KFRE) whose PPV varies between approximately 0.15-0.30. However, negative predictive values and ROCs were modest for all models (see Table and Figure), and inferior to that seen with the KFRE.

Conclusion
This preliminary analysis illustrates that ML methods may be useful to predict progression of CKD from routinely collected EHR data. However, the performance metrics of all techniques illustrate a need for further research into the optimal use of ML in this setting.

Differences between predictive models and the KFRE may be explained by the highly diverse nature of this dataset (40% of individuals identified as BME), the scarcity of proteinuria measurements in primary care records, and the fact the models in our study remain unvalidated in an independent dataset.