Scottish Government Health Directorates Chief Scientist Office

FOCUS ON RESEARCH

APPLICATION OF SUPPORT VECTOR MACHINE LEARNING TO PREDICT THE RISK OF DEATH FROM CHRONIC OBSTRUCTIVE PULMONARY DISEASE USING ELECTRONIC PRIMARY CARE MEDICAL RECORDS

Researchers

Dr Daniel Morales
Dr Jianguo Zhang
Dr Rob Flynn
Mr Kris Zutis

Aim

To determine whether routinely collected data in electronic primary care medical records can help predict death in people with Chronic Obstructive Pulmonary Disease (COPD) whilst comparing the perfomance of a support vector machine learning method of analysis with standard logistic regression.

Project Outline/Methodology

Data from primary care electronic medical records contained with the Clinical Practice Research Datalink were used to validate three existing clinical prediction rules (CPRs) that predict the risk of death in people with COPD. These CPRs were created using small numbers of people in different clinical settings. The CPRs were ADO, DOSE and COTE that use information on a persons lung fuction, the degree of breathlessness, age, smoking status, the number of exacerbations of COPD in the last year and the comorbidities people may have. To be eligible for analysis, patients were required to have been coded with a diagnosis for COPD in their medical record and to have been treated with COPD medicines.

The main outcome was death from any cause, assessed over a 1, 2 and 3 year period. Clinical information required to assess each of the three CPRs was identified and modelled using logistic regression, a “traditional” statistical approach for modelling such data. We then compared the performance of a machine learning method of analysis called Support Vector Machine Learning that required the data to be split into testing, training and validation data sets. Each models predictive performance was assessed by calculating the c-statistic, which is a standard method of assessing the predictive accuracy of such models. The c-statistic is a number that lies between 0.5 and 1, with values greater than 0.7-0.8 indicating good predictive performance and values greater than or equal to 0.8 indicating excellent predictive performance.

Key Results

A total of 204,473 people with COPD where identified in CPRD between 01/01/2000 and 01/08/2014. Of this cohort, the optimal period of clinical data recording to take a cross-section of the cohort for analysis was 01/04/2011 and included approximately 54,000 people for analysis. Of the three CPRs, ADO performed the best, following by DOSE, then the COTE comorbidities using multivariable logistic regression. Results were similar for a 1 year, 2 year and 3 year period. Predictive performance improved significantly when the COTE comorbidities were modelled with either ADO or DOSE clinical features. Comparable c-statistics were obtained using a Support Vector Machine method of analysis with the pre-defined feature set.

Conclusions

The risk of death in people with COPD can be predicted reasonably well using data contained in electronic medical records. The best performing CPR in UK primary care data was ADO when combined with COTE comorbidities. A support vector machine learning method of analysis is therefore a valid alternative to risk prediction modelling using electronic medical record data and could be applied more widely to larger more complicated data sets.

What does this study add to the field?

This study demonstrates that the performance of existing CPRs varies significantly when applied to patients in primary care (where the majority of people with COPD are managed). This study has also helped to validate an alternative analytical approach for predicting the risk of death in people with COPD that could be used to create better models in future.

Implications for Practice or Policy

This study may help clinical practice by helping clinicians to identify people with COPD at high risk of death in the following year that may help planning of palliative care services

Where to next?

We plan to develop a new clinical prediction rule by using machine learning approaches to identify novel features that predict the risk of death using primary care data.

Further details from:

d.r.z.morales@dundee.ac.uk/
danielmorales@nhs.net