BCS Health Informatics in Primary Care

Dr Ralph Kwame Akyea

Doctoral Student and Research Associate

Centre for Academic Primary Care,
Population Health and Lifespan Sciences,
School of Medicine, University of Nottingham

Performance and clinical utility of supervised machine-learning approaches in detecting familial hypercholesterolaemia in primary care

Familial hypercholesterolaemia (FH) is a high cholesterol condition that runs in families. It is an inherited condition, rather than a result of lifestyle choices. FH is thought to affect about 1 in 250 people, or around 120,000 people in the UK, although most of these people are undiagnosed and untreated. Without treatment, people with FH are much more likely to develop heart disease earlier than most people. When untreated, more than 50% of men with FH develop coronary heart disease by the age of 50, and about 30% of women by the age of 60.

Why was this study needed?

The National Institute for Heath and Care Excellence (NICE) clinical guideline on FH provides recommendations for identifying and managing the condition. However, people with raised cholesterol levels are often not recognised as potentially having FH. Several approaches have been developed to improve identification of individuals with high risk of FH in electronic health records, but we are still unclear how using machine learning (ML) can improve this. In this study we explored how well different ML approaches performed to improve FH detection in line with NICE guidance and the impact of each approach on GP workload to review patients.

What did this study do?

A retrospective cohort study was done using routine primary care clinical records of 4,027,775 individuals from the United Kingdom Clinical Practice Research Datalink (CPRD GOLD) with total cholesterol measured from 1 January 1999 to 25 June 2019. Patients aged 16 years and younger were excluded as the cholesterol level thresholds for the diagnosis and treatment of FH vary when compared to adults. Patients with a FH diagnosis prior to study entry date (1 January 1999) or with a prior diagnosis of other inherited lipid disorders were excluded.

Machine-learning algorithms/models

The total study cohort was randomly split into a 'training' cohort (75% of the study cohort) in which the FH algorithms were derived and a 'validation' cohort (remaining 25% of the cohort) in which the algorithms were applied and tested. The data split was computer-generated using a uniform distribution to generate random numbers in Stata. The five commonly used algorithms were used – logistic regression, random forest, gradient boosting machines, deep-learning neural networks, and ensemble learning. Ensemble learning model was a combination of the four (4) other ML algorithms. Using the library package h2o (http://www.h2o.ai) in R Studio, the risk algorithms were developed in the training cohort and applied to the validation cohort. A grid search was used to determine the hyper parameters for each model and ten-fold cross-validations was done to determine the values for the best performance using the training cohort.

Predictive accuracy (how well each of the five common ML approaches [logistic regression, random forest, gradient boosting machines, neural network, and ensemble learning] detect FH) was assessed by area under the receiver operating curves (AUC) and expected versus observed calibration slope; with clinical utility assessed by expected case-review workload and likelihood ratios.

What did it find?

There were 7,928 new diagnoses of FH. In addition to known clinical features of FH (raised total cholesterol or LDL-C and family history of premature coronary heart disease), machine-learning (ML) algorithms identified features such as raised triglycerides which reduced the likelihood of FH. Apart from logistic regression (AUC, 0.81), all four other ML approaches had similarly high predictive accuracy (AUC > 0.89). Calibration slope ranged from 0.997 for gradient boosting machines to 1.857 for logistic regression. Among those screened, high probability cases requiring clinical review varied from 0.73% using ensemble learning to 10.16% using deep learning, but with positive predictive values of 15.5% and 2.8% respectively. Ensemble learning exhibited a dominant positive likelihood ratio (45.5) compared to all other ML models (7.0–14.4).

What are the implications?

Machine-learning models show similar high accuracy in detecting FH, offering opportunities to increase diagnosis. With increasing processing power, primary care IT systems have the capacity to incorporate ML algorithms. This describes a practical application of this approach. However, the ML approach adopted will impact on the GP workload to review patients likely to have FH.

Citation

Akyea, R.K., Qureshi, N., Kai, J. Weng, S.F. Performance and clinical utility of supervised machine-learning approaches in detecting familial hypercholesterolaemia in primary care. npj Digit. Med. 3, 142 (2020). https://doi.org/10.1038/s41746-020-00349-5

Funding

This project was funded by the National Institute for Health Research School for Primary Care Research programme (NIHR-SPCR FR 17).