

Abstract

Full title: Machine Learning on biochemical data for the prediction of mutation presence in suspected Familial Hypercholesterolaemia

Background

Familial hypercholesterolemia (FH) is a common monogenic disorder and, if not diagnosed and treated early, results in premature atherosclerotic cardiovascular disease. Most individuals with FH are undiagnosed due to limitations in current screening and diagnostic approaches, but the advent of machine learning (ML) offers a new prospect to identify these individuals. Our objective was to create a ML model from basic lipid profile data with better screening performance than low-density lipoprotein cholesterol (LDL-C) cut-off levels and diagnostic performance comparable to the Dutch Lipid Clinic Network (DLCN) criteria.

Methods

The ML model was developed using a combination of logistic regression, deep learning and random forest classification and was trained on a 70% split of an internal dataset consisting of 555 individuals clinically suspected of having FH. The performance of the model, as well as that of the LDL-C cut-off and DLCN criteria, were assessed on both the internal 30% testing dataset and a high prevalence external dataset by comparing the area under the receiver operator characteristic (AUROC) curves. All three methodologies were measured against the gold standard of FH diagnosis by mutation identification. Furthermore, the ML model was also tested on two lower prevalence datasets derived from the same external dataset.

Results

The ML model achieved an AUROC curve of 0.711 on the high prevalence external dataset (n=1376; FH prevalence=64%), which was superior to that of the LDL-C cut off alone (AUROC=0.642) and comparable to that of the DLCN criteria (AUROC=0.705). The model performed even better when tested on the medium prevalence (n=2655; FH prevalence=20%) and low prevalence (n=1616; FH prevalence=1%) datasets, with AUROC curve values of 0.801 and 0.856 respectively.

Conclusions

Despite the absence of clinical information, the ML model was better at correctly identifying genetically confirmed FH in a cohort of individuals suspected of having FH than the LDL-C cut-off tool and comparable to the DLCN criteria. The same ML model performed even better when tested on two cohorts with lower FH prevalence. The application of ML is therefore a promising tool in both the screening for, and diagnosis of, individuals with FH.