Can machine-learning improve cardiovascular risk prediction using routine clinical data?

Stephen F. Weng1,2*, Jenna Reps3,4*, Joe Kai1,2‡, Jonathan M. Garibaldi3,4‡, Nadeem Qureshi1,2‡

1 NIHR School for Primary Care Research, University of Nottingham, Nottingham, United Kingdom, 2 Division of Primary Care, School of Medicine, University of Nottingham, Nottingham, United Kingdom, 3 Advanced Data Analysis Centre, University of Nottingham, Nottingham, United Kingdom, 4 School of Computer Science, University of Nottingham, Nottingham, United Kingdom

* These authors contributed equally to this work.
‡ These authors also contributed equally to this work.

*stephen.weng@nottingham.ac.uk

Abstract

Background

Current approaches to predict cardiovascular risk fail to identify many people who would benefit from preventive treatment, while others receive unnecessary intervention. Machine-learning offers opportunity to improve accuracy by exploiting complex interactions between risk factors. We assessed whether machine-learning can improve cardiovascular risk prediction.

Methods

Prospective cohort study using routine clinical data of 378,256 patients from UK family practices, free from cardiovascular disease at outset. Four machine-learning algorithms (random forest, logistic regression, gradient boosting machines, neural networks) were compared to an established algorithm (American College of Cardiology guidelines) to predict first cardiovascular event over 10-years. Predictive accuracy was assessed by area under the ‘receiver operating curve’ (AUC); and sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) to predict 7.5% cardiovascular risk (threshold for initiating statins).

Findings

24,970 incident cardiovascular events (6.6%) occurred. Compared to the established risk prediction algorithm (AUC 0.728, 95% CI 0.723–0.735), machine-learning algorithms improved prediction: random forest +1.7% (AUC 0.745, 95% CI 0.739–0.750), logistic regression +3.2% (AUC 0.760, 95% CI 0.755–0.766), gradient boosting +3.3% (AUC 0.761, 95% CI 0.755–0.766), neural networks +3.6% (AUC 0.764, 95% CI 0.759–0.769). The highest achieving (neural networks) algorithm predicted 4,998/7,404 cases (sensitivity 67.5%, PPV 18.4%) and 53,458/75,585 non-cases (specificity 70.7%, NPV 95.7%), correctly predicting 355 (+7.6%) more patients who developed cardiovascular disease compared to the established algorithm.
Conclusions

Machine-learning significantly improves accuracy of cardiovascular risk prediction, increasing the number of patients identified who could benefit from preventive treatment, while avoiding unnecessary treatment of others.

Introduction

Globally, cardiovascular disease (CVD) is the leading cause of morbidity and mortality. In 2012, there were 17.5 million deaths from CVD with 7.4 million deaths due to coronary heart disease (CHD) and 6.7 million deaths due to stroke [1]. Established approaches to CVD risk assessment, such as that recommended by the American Heart Association/American College of Cardiology (ACC/AHA), predict future risk of CVD based on well-established risk factors such as hypertension, cholesterol, age, smoking, and diabetes. These risk factors have recognised aetiological associations with CVD and feature within most CVD risk prediction tools (e.g. ACC/AHA [2], QRISK2 [3], Framingham [4], Reynolds [5]. There remain a large number of individuals at risk of CVD who fail to be identified by these tools, while some individuals not at risk are given preventive treatment unnecessarily. For instance, approximately half of myocardial infarctions (MIs) and strokes will occur in people who are not predicted to be at risk of cardiovascular disease [6].

All standard CVD risk assessment models make an implicit assumption that each risk factor is related in a linear fashion to CVD outcomes [7]. Such models may thus oversimplify complex relationships which include large numbers of risk factors with non-linear interactions. Approaches that better incorporate multiple risk factors, and determine more nuanced relationships between risk factors and outcomes need to be explored.

Machine-learning (ML) offers an alternative approach to standard prediction modelling that may address current limitations. It has potential to transform medicine by better exploiting 'big data' for algorithm development [7]. ML developed from the study of pattern recognition and computational learning (so-called 'artificial intelligence'). This relies on a computer to learn all complex and non-linear interactions between variables by minimising the error between predicted and observed outcomes [8]. In addition to potentially improving prediction, ML may identify latent variables, which are unlikely to be observed but might be inferred from other variables [9].

To date, there has been no large-scale investigation applying machine-learning for prognostic assessment in the general population, using routine clinical data. The aim of this study was to evaluate whether machine-learning can improve accuracy of cardiovascular risk prediction within a large general primary care population. We also sought to determine which class of machine-learning algorithm has highest predictive accuracy.

Methods

Data source

The cohort of patients was derived from the Clinical Practice Research Datalink (CPRD), anonymized electronic medical records from nearly 700 UK family practices documenting demographic details, history of medical conditions, prescription drugs, acute medical outcomes, referrals to specialists, admissions to hospitals, and biological results. The database is representative of the UK general population and linked to hospital (secondary care) records [10]. Ethical and research approvals were granted by the Independent Scientific Advisory Committee (ISAC) at CPRD (number 14_205).
Study population
The cohort of patients were registered with a family practice between the ages of 30 to 84 years at baseline, who had complete data for the eight core baseline variables (gender, age, smoking status, systolic blood pressure, blood pressure treatment, total cholesterol, HDL cholesterol, and diabetes) used in the established ACC/AHA 10-year risk prediction model [2]. The baseline date was set as the 1st of January 2005, thus allowing all patients within the cohort to be followed-up for 10 years. The end of the study period was specified as the 1st of January 2015, the latest date for which CPRD had provided an updated dataset. Individuals with a previous history of CVD, lipid disorders which are inherited, prescribed lipid lowering drugs, or outside the specified age range prior to or on the baseline date were excluded from the analysis.

Risk factor variables
The eight core risk variables (above) were used to derive a baseline risk prediction model using the published equations in the 2013 ACC/AHA guidelines for assessment of CVD risk [2]. To compare the machine-learning algorithms, an additional 22 variables with potential to be associated with CVD were included in the analysis. These variables were selected based on their inclusion in published CVD risk algorithms [2–5], within literature on other potential CVD risk factors [11–21], and further reviewed by practising clinicians (NQ, JK).

In nine of the additional continuous variables, there were some levels of missing data. Median imputation, a common approach to dealing with missing values in machine-learning algorithms [22] was used. It was also hypothesized that missing values in certain clinical variables (e.g. BMI and laboratory results) may indicate a perception of reduced relevance in certain patients, given the under recording of normal BMI values in primary care medical records [23]. Dummy variables were created to indicate whether these continuous variable values were missing. For demographic categorical variables, Townsend deprivation index (28) and ethnicity, missing values were given a separate category of ‘unknown’ in the analyses. In total, there were 30 variables (excluding dummy variables for missing values) analysed in the machine-learning models prior to baseline (Table 1).

Outcome
The primary outcome was the first recorded diagnosis of a fatal or non-fatal cardiovascular event documented in the patient’s primary or secondary care computerised record. In primary care, CVD is labelled and electronically recorded by UK National Health Service (NHS) Read codes. Further, confirmation of outcomes in secondary care (Hospital Episodes Statistics) utilised ICD-10 codes, specifically I20 to I25 for coronary (ischaemic) heart conditions and I60 to I69 for cerebrovascular conditions.

Machine-learning algorithms
To compare machine-learning risk algorithms, the study population was split in the data set into a ‘training’ cohort in which the CVD risk algorithms were derived and a ‘validation’ cohort in which the algorithms were applied and tested. The ‘training’ cohort was derived from random sampling of 75% of the extracted CPRD cohort, and the ‘validation’ cohort comprised the remaining 25%. Four commonly used classes of machine-learning algorithms were utilised: logistic regression [25], random forest [26], gradient boosting machines [27], and neural networks [28]. These algorithms were selected based on the ease of implementation into current UK primary care electronic health records. Development of the risk algorithms in the training cohort and application of the risk algorithms to the validation cohort was completed.
using RStudio with library packages caret (http://CRAN.R-project.org/package=caret) for neural networks and h2o (http://www.h2o.ai) for the remaining algorithms. Each model’s hyper parameters were determined by using a grid search and two fold cross-validation on the training cohort to determine the values which led to the best performance. Further details on machine-learning models are described in the S1 Text.

### Statistical analysis

Descriptive characteristic of the study population were provided, including number (%) and mean (SD) for categorical and continuous variables, respectively. The performance of the machine-learning prediction algorithms, developed from the training cohort, was assessed using the validation cohort by calculating Harrell’s c-statistic [29], a measure of the total area
under the receiver operating characteristic curve (AUC). Standard errors and 95% confidence intervals were estimated for the c-statistic using a jack-knife procedure [30]. Additionally, using thresholds corresponding to the 10-year CVD risk of > 7.5% as recommended by the ACC/AHA guidelines [2] for initiating lipid lowering therapy, binary classification analysis was used to compare observed and expected prediction of cases and non-cases in the validation cohort. This process provided sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). The statistical analyses assessing algorithm performance were performed using STATA 13 MP4.

Results

Data extraction

There were a total of 383,592 patients from 12 million patients in the CPRD database at baseline (1 Jan 2005) who met eligibility criteria. After excluding 5,336 patients with coding errors (i.e. non-numerical entries for blood pressure/cholesterol) and extreme outlying observations (> 5 SDs from the mean), the analysis cohort consisted of 378,256 patients. This cohort was then randomly split into a 75% sample of 295,267 patients to train the machine-learning algorithms and the remaining sample of 82,989 patients for validation (Fig 1).

Study population characteristics

From a total cohort of 378,256 patients who were free from CVD at baseline, there were 24,970 incident cases (6.6%) of CVD during the 10-year follow-up period. There were significantly fewer women than men (42% F, 52% M) in CVD cases while there was only slightly more women than men in non-CVD cases (52% F, 48% M). The mean baseline age of CVD patients was 65.3 years compared to 57.3 years in non-CVD patients (p < 0.001). Further characteristics of CVD and non-CVD patients are presented in Table 2.

Machine-learning variable rankings

All variables listed in Table 2 were inputs for the machine-learning models and trained using a cohort of 295,267 patients with 19,487 incident CVD cases (6.6%) of developing over the 10-year follow-up period. Variable importance was determined by the coefficient effect size for the ACC/AHA baseline model and machine-learning logistic regression. Random forest and gradient boosting machine models, based on decision-trees, rank variable importance on the selection frequency of the variable as a decision node while neural networks use overall weighting of the variable within the model. The top 10 risk factors for the CVD prediction algorithms are presented in Table 3.

The standard risk factors in the ACC/AHA algorithm stratified by gender were age, total cholesterol, HDL cholesterol, smoking, blood pressure, and diabetes. Several of these risk factors in the ACC/AHA model (age, gender, smoking) were present as top ranked risk factors for all four machine-learning algorithms. However, diabetes, which is prominent in many CVD algorithms, was not present in the top ranked risk factors for any of the machine-learning models (though HbA1c was included as a proxy in random forest models). Other new risk factors not found in any previous risk prediction tools but determined by machine-learning included medical conditions such as COPD and severe mental illness, prescribing of oral corticosteroids, as well as biomarkers such as triglyceride levels. Random forest and gradient boosting machines were most similar in risk factor selection and rankings, with some discrepancies in ranking order and substitution of BMI for systolic blood pressure. Logistic regression and neural networks prioritised medical conditions such as atrial fibrillation, chronic kidney
disease, and rheumatoid arthritis over biometric risk factors. Neural networks also put less weighting on age as a risk factor, and included ‘BMI missing’ as a protective risk factor of CVD. Full variable selection rankings can be found in S1 Table.

Prediction accuracy
The prediction accuracy according to the discrimination (AUC c-statistic) is shown in Table 4 for all models.

Fig 1. Patient cohort data extraction procedures.
https://doi.org/10.1371/journal.pone.0174944.g001
Table 2. Characteristics of patients aged 30 to 84 in the CPRD study cohort who were free from CVD at baseline. Patients are stratified by first CVD event during the 10-year follow-up period.

<table>
<thead>
<tr>
<th>Risk Factor Variables</th>
<th>Units</th>
<th>CVD (n = 24,970)</th>
<th>No CVD (n = 353,286)</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age*</td>
<td>years (SD)</td>
<td>65.3 (11.1)</td>
<td>57.6 (12.8)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>BMI*</td>
<td>kg/m² (SD)</td>
<td>27.9 (4.94)</td>
<td>27.9 (5.21)</td>
<td>0.323</td>
</tr>
<tr>
<td>Systolic blood pressure*</td>
<td>mm HG (SD)</td>
<td>141 (17.6)</td>
<td>137 (17.2)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Total cholesterol*</td>
<td>mmol/L (SD)</td>
<td>5.60 (1.11)</td>
<td>5.56 (1.06)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>HDL cholesterol*</td>
<td>mmol/L (SD)</td>
<td>1.39 (0.41)</td>
<td>1.46 (0.43)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>LDL cholesterol*</td>
<td>mmol/L (SD)</td>
<td>3.45 (0.91)</td>
<td>3.40 (0.88)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Triglycerides*</td>
<td>mmol/L (SD)</td>
<td>1.69 (0.85)</td>
<td>1.57 (0.83)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>CRP*</td>
<td>mg/L (SD)</td>
<td>10.0 (13.7)</td>
<td>8.37 (11.5)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Serum fibrinogen*</td>
<td>g/L (SD)</td>
<td>3.86 (1.22)</td>
<td>3.73 (1.33)</td>
<td>0.129</td>
</tr>
<tr>
<td>gamma GT*</td>
<td>IU/L (SD)</td>
<td>41.3 (33.7)</td>
<td>39.3 (33.6)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Serum creatinine*</td>
<td>umol/L (SD)</td>
<td>91.9 (17.3)</td>
<td>87.6 (16.0)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>HbA1c*</td>
<td>% (SD)</td>
<td>7.26 (1.61)</td>
<td>7.14 (1.64)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>FEV1*</td>
<td>% (SD)</td>
<td>66.2 (16.3)</td>
<td>67.8 (16.9)</td>
<td>0.007</td>
</tr>
<tr>
<td>AST/ALT ratio*</td>
<td>— (SD)</td>
<td>1.04 (0.36)</td>
<td>1.01 (0.35)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Female*</td>
<td>%</td>
<td>41.8</td>
<td>52.8</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Smoking*</td>
<td>%</td>
<td>23.4</td>
<td>20.5</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Family history CHD &lt; 60 years</td>
<td>%</td>
<td>5.00</td>
<td>5.51</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Ethnicity*: South Asian</td>
<td>%</td>
<td>2.27</td>
<td>1.90</td>
<td>0.004</td>
</tr>
<tr>
<td>Ethnicity*: Black/Afro-Caribbean</td>
<td>%</td>
<td>0.66</td>
<td>1.20</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Ethnicity*: Chinese/East Asian</td>
<td>%</td>
<td>0.54</td>
<td>0.58</td>
<td>0.465</td>
</tr>
<tr>
<td>Ethnicity*: Other/Mixed</td>
<td>%</td>
<td>0.85</td>
<td>1.32</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Ethnicity*: Unknown</td>
<td>%</td>
<td>43.5</td>
<td>57.1</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>SES*: 2nd Townsend quintile</td>
<td>%</td>
<td>15.8</td>
<td>16.0</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>SES*: 3rd Townsend quintile</td>
<td>%</td>
<td>13.7</td>
<td>13.6</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>SES*: 4th Townsend quintile</td>
<td>%</td>
<td>12.6</td>
<td>11.8</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>SES*: 5th Townsend quintile (most deprived)</td>
<td>%</td>
<td>7.95</td>
<td>6.91</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>SES*: Unknown</td>
<td>%</td>
<td>34.6</td>
<td>34.5</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Hypertension</td>
<td>%</td>
<td>31.8</td>
<td>25.2</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Diabetes</td>
<td>%</td>
<td>15.0</td>
<td>10.1</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Blood pressure treatment*</td>
<td>%</td>
<td>28.3</td>
<td>21.9</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>%</td>
<td>1.55</td>
<td>0.91</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Chronic kidney disease</td>
<td>%</td>
<td>0.99</td>
<td>0.48</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Atrial fibrillation</td>
<td>%</td>
<td>4.64</td>
<td>2.20</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>COPD</td>
<td>%</td>
<td>3.97</td>
<td>2.02</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Severe mental illness</td>
<td>%</td>
<td>0.34</td>
<td>0.32</td>
<td>0.563</td>
</tr>
<tr>
<td>Anti-psychotic drug prescribed</td>
<td>%</td>
<td>15.2</td>
<td>12.7</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Oral corticosteroid prescribed</td>
<td>%</td>
<td>13.2</td>
<td>9.55</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Immunosuppressant prescribed</td>
<td>%</td>
<td>13.3</td>
<td>9.70</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>BMI missing</td>
<td>%</td>
<td>3.48</td>
<td>5.87</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>LDL cholesterol missing</td>
<td>%</td>
<td>25.1</td>
<td>24.6</td>
<td>0.041</td>
</tr>
<tr>
<td>Triglycerides missing</td>
<td>%</td>
<td>11.7</td>
<td>12.3</td>
<td>0.004</td>
</tr>
<tr>
<td>CRP missing</td>
<td>%</td>
<td>88.5</td>
<td>89.9</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Serum fibrinogen missing</td>
<td>%</td>
<td>99.0</td>
<td>99.0</td>
<td>0.207</td>
</tr>
<tr>
<td>gamma GT missing</td>
<td>%</td>
<td>64.8</td>
<td>69.1</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Serum creatinine missing</td>
<td>%</td>
<td>16.1</td>
<td>21.5</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>HbA1c missing</td>
<td>%</td>
<td>79.6</td>
<td>85.9</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

(Continued)
The ACC/AHA risk model served as a baseline for comparison (AUC 0.728, 95% CI 0.723–0.735). All machine-learning algorithms tested achieved statistically significant improvements in discrimination compared to the baseline models (from 1.7% for random forest algorithms to 3.6% for neural networks).

Classification analysis

The ACC/AHA baseline model predicted 4,643 cases correctly from 7,404 total cases, resulting in a sensitivity of 62.7% and PPV of 17.1%. The random forest algorithm resulted in a net increase of 191 CVD cases from the baseline model, increasing the sensitivity to 65.3% and PPV to 17.8% while logistic regression resulted in a net increase of 324 CVD cases (sensitivity 67.1%; PPV 18.3%). Gradient boosting machines and neural networks performed best, resulting in a net increase of 354 (sensitivity 67.5%; PPV 18.4%) and 355 CVD (sensitivity 67.5%; PPV 18.4%) cases correctly predicted, respectively.

The ACC/AHA baseline model correctly predicted 53,106 non-cases from 75,585 total non-cases, resulting in a specificity of 70.3% and NPV of 95.1%. The net increase in non-cases

Table 2. (Continued)

<table>
<thead>
<tr>
<th>Risk Factor Variables</th>
<th>Units</th>
<th>CVD (n = 24,970)</th>
<th>No CVD (n = 353,286)</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEV1 missing</td>
<td>%</td>
<td>96.3</td>
<td>97.7</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>AST/ALT ratio missing</td>
<td>%</td>
<td>85.2</td>
<td>88.2</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

*core risk factor for ACC/AHA 10-year CVD risk equations
*missing values present
*reference category is White Caucasian
*reference category is 1st Townsend quintile (most affluent)

The ACC/AHA risk model served as a baseline for comparison (AUC 0.728, 95% CI 0.723–0.735). All machine-learning algorithms tested achieved statistically significant improvements in discrimination compared to the baseline models (from 1.7% for random forest algorithms to 3.6% for neural networks).

Table 3. Top 10 risk factor variables for CVD algorithms listed in descending order of coefficient effect size (ACC/AHA; logistic regression), weighting (neural networks), or selection frequency (random forest, gradient boosting machines). Algorithms were derived from training cohort of 295,267 patients.

<table>
<thead>
<tr>
<th>ACC/AHA Algorithm</th>
<th>Men</th>
<th>Women</th>
<th>ML: Logistic Regression</th>
<th>ML: Random Forest</th>
<th>ML: Gradient Boosting Machines</th>
<th>ML: Neural Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age</td>
<td>Ethnicity</td>
<td>Age</td>
<td>Age</td>
<td>Atrial Fibrillation</td>
<td></td>
</tr>
<tr>
<td>Total Cholesterol</td>
<td>HDL Cholesterol</td>
<td>Age</td>
<td>Gender</td>
<td>Gender</td>
<td>Ethnicity</td>
<td></td>
</tr>
<tr>
<td>HDL Cholesterol</td>
<td>Total Cholesterol</td>
<td>SES: Townsend Deprivation Index</td>
<td>Ethnicity</td>
<td>Ethnicity</td>
<td>Oral Corticosteroid Prescribed</td>
<td></td>
</tr>
<tr>
<td>Smoking</td>
<td>Smoking</td>
<td>Gender</td>
<td>Smoking</td>
<td>Smoking</td>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>Age x Total Cholesterol</td>
<td>Age x HDL Cholesterol</td>
<td>Smoking</td>
<td>HDL cholesterol</td>
<td>HDL cholesterol</td>
<td>Severe Mental Illness</td>
<td></td>
</tr>
<tr>
<td>Treated Systolic Blood Pressure</td>
<td>Age x Total Cholesterol</td>
<td>Atrial Fibrillation</td>
<td>HbA1c</td>
<td>Triglycerides</td>
<td>SES: Townsend Deprivation Index</td>
<td></td>
</tr>
<tr>
<td>Age x Smoking</td>
<td>Treated Systolic Blood Pressure</td>
<td>Chronic Kidney Disease</td>
<td>Triglycerides</td>
<td>Total Cholesterol</td>
<td>Chronic Kidney Disease</td>
<td></td>
</tr>
<tr>
<td>Age x HDL Cholesterol</td>
<td>Untreated Systolic Blood Pressure</td>
<td>Rheumatoid Arthritis</td>
<td>SES: Townsend Deprivation Index</td>
<td>HbA1c</td>
<td>BMI missing</td>
<td></td>
</tr>
<tr>
<td>Untreated Systolic Blood Pressure</td>
<td>Age x Smoking</td>
<td>Family history of premature CHD</td>
<td>BMI</td>
<td>Systolic Blood Pressure</td>
<td>Smoking</td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>Diabetes</td>
<td>COPD</td>
<td>Total Cholesterol</td>
<td>SES: Townsend Deprivation Index</td>
<td>Gender</td>
<td></td>
</tr>
</tbody>
</table>

Italics: Protective Factors

https://doi.org/10.1371/journal.pone.0174944.t003
correctly predicted compared to the baseline ACC/AHA model ranged from 191 non-cases for the random forest algorithm to 355 non-cases for the neural networks. Full details on classification analysis can be found in S2 Table.

**Discussion**

Compared to an established AHA/ACC risk prediction algorithm, we found all machine-learning algorithms tested were better at identifying individuals who will develop CVD and those that will not. Unlike established approaches to risk prediction, the machine-learning methods used were not limited to a small set of risk factors, and incorporated more pre-existing medical conditions. Neural networks performed the best, with predictive accuracy improving by 3.6%. This is an encouraging step forward. For example, the addition of emerging biochemical risk factors, such as high sensitivity C-reactive protein, has recently achieved less than 1% improvement in CVD risk prediction [31].

**Strengths**

To our knowledge, this is the first investigation applying machine-learning to routine data in patients’ electronic records, demonstrating improved prediction of CVD risk in a large general population. The study also illustrates use of a range of machine learning methods, as well as evaluation techniques, that are lacking in existing applications of machine-learning to clinical data [32]. Our results are consistent with much smaller studies [33,34] in more selected populations. For example, a cohort study of 5,159 men in Northern Germany [34] found a similar 3.2% improvement in accuracy of prediction of coronary risk using a probabilistic neural network model.

The current study’s use of an array of machine-learning algorithms has suggested intriguing variations in the importance of different risk factors depending on the modelling technique. Models based on decision trees resembled closely to each other, with gradient boosting machines out-performing random forests. Neural networks and logistic regression placed far more importance on categorical variables and CVD-associated medical conditions, clustering patients with similar characteristics in each groups. This may help inform further exploration of diverse predictive risk factors, and future development of new risk prediction approaches and algorithms.

Finally, the importance of missing values or non-response are not often assessed in development of conventional CVD risk prediction tools [2–5]. This study suggests that missing values, in particular, for routine biometric variables such as BMI, are independent predictors of CVD.
This is consistent with subjective assessment by clinicians who may not record normal BMI values if patients appear at lower CVD risk [23].

**Limitations**

It is acknowledged that the “black-box” nature of machine-learning algorithms, in particular neural networks, can be difficult to interpret. This refers to the inherent complexity in how the risk factor variables are interacting and their independent effects on the outcome. However, improvements in data visualization methods have improved understanding of these models, illustrating the importance of network connections between risk factors [35] (See example visualising our neural network model in Fig 2).

It is also recognised that as the number of potential risk factors increases, the complexity of the models can cause over-fitting, yielding implausible results. We addressed this by active and appropriate choice of pre-training, hyper-parameter selection, and regularisation [36].

Although we have cross-validated the performance of the machine-learning algorithms using an independent dataset, an approach commonly used for the development of established cardiovascular risk algorithms applied to clinical practice [2–5,24,37], it must be acknowledged that the jack-knife procedure may yield more accurate results as demonstrated in genomic or proteomic datasets [38,39]. Moreover, these established risk prediction algorithms for use in clinical practice have been developed from a binary classification framework which can often result in an unbalanced dataset. Ensemble learning have been demonstrated as a solution to construct balanced datasets to enhance prediction performance [40]. These methods are not yet commonplace for developing risk prediction models in clinical datasets but their utility should be explored in future studies.

Finally, we note the study was performed in a large cohort of primary care patients in the UK. However, its demonstration of machine-learning methods, and use of routine clinical data available within electronic records in several countries [41], underline applicability to other populations and health systems.
Future implications

CVD risk prediction has become increasingly important in clinical decision-making since the introduction of the recent ACC/AHA and similar guidelines internationally [2,42]. Machine-learning approaches offer the exciting prospect of achieving improved and more individualised CVD risk assessment. This may assist the drive towards personalised medicine, by better tailoring risk management to individual patients [43,44].

The improvement in predictive accuracy found in the current study should be further explored using machine learning with other large clinical datasets, in other populations, and in predicting other disease outcomes. Future investigation of the feasibility and acceptability of machine-learning applications in clinical practice will be needed. As the computational capacity in health care systems is improving, the opportunities to exploit machine-learning to enhance prediction of disease risk in clinical practice will become a realistic option [7]. This might increasingly include predicting protein structure and function from genetic sequences from patients’ clinical profiles [7]. This will inevitably require exploration in future studies on utility and clinical applicability other computationally demanding machine-learning algorithms such as support vector machines and deep learning for integration into primary care electronic health records. In several countries, electronic health records across health care organisations are held on central servers. This may allow new algorithm development to be performed off-site using cloud computing software, and then returned to the clinical setting as applications programme interfaces (APIs) for PCs, mobile devices and tablets.

Conclusion

Compared to an established risk prediction approach, this study has shown machine-learning algorithms are better at predicting the absolute number of cardiovascular disease cases correctly, whilst successfully excluding non-cases. This has been demonstrated in a large and heterogeneous primary care patient population using routinely collected electronic health data.

Supporting information

S1 Table. Full ranking of important variables for four machine-learning 10-year CVD risk prediction algorithms. Variable importance determined based on coefficient effect sizes (logistic regression), frequency (random forest, gradient boosting machines), or weighting (neural networks) developed from the training CPRD training cohort of 295,267 patients.

S2 Table. Classification analysis showing sensitivity, specificity, positive predictive value (PPV), and negative predictive (NPV) value of 10-year CVD machine-learning prediction algorithms. Thresholds are determined corresponding to the ACC/AHA guideline recommendation determining ‘high risk’ > 7.5% for initiating of lipid modification.

S1 Text. Machine-learning algorithms.

Acknowledgments

This paper presents independent research funded by the National Institute for Health Research School for Primary Care Research (NIHR SPCR): post-doctoral training fellowship award (URL: https://www.spcr.nihr.ac.uk/trainees). The views expressed are those of the authors and not necessarily those of the NIHR, the NHS, or the Department of Health.
Can machine-learning improve cardiovascular risk prediction using routine clinical data?

Author Contributions

Conceptualization: SW JR JK NQ JG.

Data curation: SW JR.

Formal analysis: SW JR.

Funding acquisition: SW NQ JK.

Investigation: SW JR.

Methodology: SW JR JK NQ JG.

Project administration: SW.

Resources: SW JR.

Software: SW JR.

Supervision: SW NQ.

Validation: SW JR JK NQ JG.

Visualization: SW JR JK NQ JG.

Writing – original draft: SW JR.

Writing – review & editing: SW JR JK NQ JG.

References


Can machine-learning improve cardiovascular risk prediction using routine clinical data?


