Journal of the American Heart Association

ORIGINAL RESEARCH

Improving the Detection of Potential Cases of Familial Hypercholesterolemia: Could Machine Learning Be Part of the Solution?

Christophe A. T. Stevens, MSc; Antonio J. Vallejo-Vaz, PhD; Joana R. Chora, PhD; Fotis Barkas, PhD; Julia Brandts, MD; Alireza Mahani, PhD; Leila Abar, PhD; Mansour T. A. Sharabiani, PhD*; Kausik K. Ray, FMedSci*

BACKGROUND: Familial hypercholesterolemia (FH), while highly prevalent, is a significantly underdiagnosed monogenic disorder. Improved detection could reduce the large number of cardiovascular events attributable to poor case finding. We aimed to assess whether machine learning algorithms outperform clinical diagnostic criteria (signs, history, and biomarkers) and the recommended screening criteria in the United Kingdom in identifying individuals with FH-causing variants, presenting a scalable screening criteria for general populations.

METHODS AND RESULTS: Analysis included UK Biobank participants with whole exome sequencing, classifying them as having FH when (likely) pathogenic variants were detected in their *LDLR*, *APOB*, or *PCSK9* genes. Data were stratified into 3 data sets for (1) feature importance analysis; (2) deriving state-of-the-art statistical and machine learning models; (3) evaluating models' predictive performance against clinical diagnostic and screening criteria: Dutch Lipid Clinic Network, Simon Broome, Make Early Diagnosis to Prevent Early Death, and Familial Case Ascertainment Tool. One thousand and three of 454 710 participants were classified as having FH. A Stacking Ensemble model yielded the best predictive performance (sensitivity, 74.93%; precision, 0.61%; accuracy, 72.80%, area under the receiver operating characteristic curve, 79.12%) and outperformed clinical diagnostic criteria and the recommended screening criteria in identifying FH variant carriers within the validation data set (figures for Familial Case Ascertainment Tool, the best baseline model, were 69.55%, 0.44%, 65.43%, and 71.12%, respectively). Our model decreased the number needed to screen compared with the Familial Case Ascertainment Tool (164 versus 227).

CONCLUSIONS: Our machine learning–derived model provides a higher pretest probability of identifying individuals with a molecular diagnosis of FH compared with current approaches. This provides a promising, cost-effective scalable tool for implementation into electronic health records to prioritize potential FH cases for genetic confirmation.

Key Words: cardiovascular disease prevention ■ familial hypercholesterolemia ■ genetic ■ machine learning ■ screening

Familial hypercholesterolaemia (FH) is a highly prevalent autosomal codominant genetic condition occurring with a prevalence of ≈1:311 in the worldwide general population.[1,2] FH results in lifelong (from birth) exposure to high low-density lipoprotein cholesterol (LDL-C) levels, increasing the risk of premature atherosclerotic cardiovascular disease (ASCVD), more notably coronary artery disease (CAD).[1,3] However, if individuals with FH are identified and prescribed effective cholesterol-lowering treatments early, then much of their future adverse health outcomes can be avoided.[4] Yet FH remains largely underdiagnosed, with

Correspondence to: Christophe A. T. Stevens, MSc, Department of Primary Care and Public Health, Imperial Centre for Cardiovascular Disease Prevention, Imperial College London, School of Public Health Building, 90 Wood Lane, London W12 7TA, United Kingdom. Email: christophe.stevens@imperial.ac.uk

*M. T. A. Sharabiani and K. K. Ray are co-senior authors.

This manuscript was sent to Manju Bengaluru Jayanna, MD, MS, Assistant Editor, for review by expert referees, editorial decision, and final disposition.

Supplemental Material is available at https://www.ahajournals.org/doi/suppl/10.1161/JAHA.123.034434

For Sources of Funding and Disclosures, see page 15.

JAHA is available at: www.ahajournals.org/journal/jaha

CLINICAL PERSPECTIVE

What Is New?

- The study derived and evaluated the utility of novel models derived from machine learning (ML) for the automated identification of genetically confirmed cases of familial hypercholesterolemia (FH) in the general population using variables available within electronic health records .
- Our novel ML-derived model exhibits superior sensitivity and a lower number needed to screen for FH when compared with clinical diagnostic criteria and the automated screening criteria currently recommended in the United Kingdom for large-scale screening of EHRs.
- The data set, derived from the UK Biobank, includes 454 710 adults with whole exome sequencing, of which 1003 have genetically validated (likely) pathogenic FH variants, providing a robust foundation for model development.

What Are the Clinical Implications?

- Integration of ML-derived models into electronic health recaords for prioritizing genetic confirmation of FH offers a more effective and efficient approach in finding FH in adult populations in general populations and diverse clinical settings.
- Superior sensitivity and lower number needed to screen of the new ML model offer promise for the detection of a higher percentage of true individuals with FH with fewer tests, enabling targeted preventive measures.
- Implementing ML screening *criteria* may enhance early identification and management, potentially reducing acute myocardial infarctions, revascularizations, and cardiovascular death resulting from undetected cases.

Nonstandard Abbreviations and Acronyms

ASCVD	atherosclerotic cardiovascular disease
AUC	area under the receiver operating characteristic curve
DLCN	Dutch Lipid Clinic Network
FAMCAT	Familial Hypercholesterolaemia Case Ascertainment Identification Tool
FH	familial hypercholesterolemia
LLM	lipid-lowering medication
LR	logistic regression
ML	machine learning
NNS	number needed to screen
UKB	UK Biobank
XGBoost	eXtreme Gradient Boosting

an estimated 22.5 million adults with FH remaining to be identified globally (<10% identified).[1,2]

The definitive method for diagnosing FH is through molecular (genetic) testing, recognized as the gold standard. There are currently no population-wide genetic testing programs. Typically, genetic testing for FH is initiated by secondary and tertiary care specialists following a referral from primary care. These referrals are based on the physician's clinical suspicion of FH, for example, prompted by a premature ASCVD event such as an acute coronary syndrome. Specialists then use clinical diagnostic criteria, such as the Dutch Lipid Clinic Network (DLCN), Make Early Diagnosis to Prevent Early Death, or Simon Broome,[5–7] which were derived and validated on genetically confirmed cohorts of patients, to assess the likelihood of a positive, confirmatory genetic diagnosis in a patient with suspected FH. Yet these clinical diagnostic criteria demand thorough phenotypical assessments and detailed family histories, necessitating direct patient consultations. Such comprehensive data may not be consistently documented in electronic health records (EHRs), rendering these criteria less practical for automated, large-scale screening approaches within the general population through primary care EHR systems, where many undiagnosed cases of FH likely exist. Screening criteria, on the other hand, are specifically developed for identifying potential FH cases in the general population through automated screening of EHRs.

Screening criteria may circumvent the limitations of clinical diagnostic criteria, as they do not rely on the availability of detailed information on family history and physical signs and thus offer the ability to identify true cases of FH at scale. However, examples like the Familial Case Ascertainment Tool (FAMCAT) currently recommended in the United Kingdom, have not been exclusively derived and validated in molecularly confirmed cohorts of patients with FH, potentially leading to the identification of individuals with hypercholesterolemia due to secondary or polygenic causes rather than (monogenic) FH itself.[8,9] This distinction is crucial because individuals with FH have a higher average risk of ASCVD than patients with secondary or polygenic hypercholesterolemia, and their identification not only offers an opportunity to prevent ASCVD events but also facilitates cascade screening among relatives. Moreover, applying screening criteria that were not exclusively derived from molecularly defined individuals, like FAMCAT, may result in the referral of numerous unnecessary cases for genetic testing. This approach may lead to increased health care costs due to a higher number of individuals needing to be screened to find true FH diagnoses, exemplifying a low-yield strategy.

In our study, we aimed to assess whether statistical and machine learning (ML) algorithms encompassing diverse variables captured in EHRs could outperform

clinical diagnostic criteria and screening criteria in providing an automated process for identifying individuals subsequently confirmed to have an FH-causing variant and thus potential clinical utility. We derived and evaluated several models for identifying a molecular diagnosis of FH in the UK Biobank (UKB). We designed our models to operate within the constraints of real-world EHR data, where detailed clinical signs used in diagnostic criteria used by specialists in secondary care are often absent. Afterwards, we compared their effectiveness in identifying individuals with a confirmed genetic diagnosis of FH against both widespread clinical diagnostic criteria and the FAMCAT screening criteria. Our evaluation included assessing sensitivity and determining the number needed to screen (NNS) for those with positive genetic testing for FH. To conclude our analysis, we calculated and compared the clinical utility of screening and diagnostic criteria against our de novo ML-derived model within a hypothetical large-scale automated screening program targeting the UK population aged >40 years. This assessment focused on the number of identified patients with FH and the tests required by each criterion and our chosen model.

METHODS

Availability of Data and Source Code

The data used in this study are available from the UK Biobank, but restrictions apply to the availability of these data and are therefore not publicly available. Researchers will need to apply to access the UK Biobank database at www.ukbiobank.ac.uk. The source code is withheld due to intellectual property rights.

Data Source

We used data from the UKB study (application 67789), a prospective study of >500 000 participants aged 40 to 69 years when recruited from 2006 to 2010, all of whom granted written informed consent for participation. The UKB encompasses comprehensive data pertaining to participants' lifestyle, environmental influences, genotype, and phenotype information, with the ongoing monitoring of their health status facilitated through the integration of EHRs.[10,11] Only participants with whole exome sequencing data were retained for the analysis. The UKB study has been approved from the Northwest Multi-centre Research Ethics Committee as a Research Tissue Bank.

Definition of FH

FH was defined as the presence of a pathogenic variant in 1 of 3 genes, namely, low-density lipoprotein receptor (*LDLR*), apolipoprotein B (*APOB*), and proprotein convertase subtilisin/kexin type 9 (*PCSK9*). The *LDLR* variants were classified according to a simplified version of the FH *LDLR* variant interpretation guidelines, an implementation of the American College of Medical Genetics and Genomics guideline,[12] proposed by the ClinGen FH Variant Curation Expert Panel,[13] and to the level of pathogenicity reported in the ClinVar database (Data S1). Given the absence of specific *APOB* and *PCSK9* gene variant classification guidelines, we assessed variant pathogenicity using the ClinVar database. Individuals with at least 1 "pathogenic" or "likely pathogenic" variant, hereafter simply referred to as "pathogenic variants," were considered FH carriers (Data S2).

Data Preparation

Variables used to predict the FH-carrying status were used "as-is" or derived from the occurrence of *International Classification of Diseases, Tenth Revision* (*ICD-10*) and Office of Population Censuses and Surveys version 4 codes (comorbidities and procedures), and medication labels within patients' records (Data S3). Variables with >20% missing values were excluded. Categorical variables were converted into binary vectors. Numerical variables were centered and scaled. The multivariate imputation by the chained equations method was used to impute missing data in the remaining variables.

The data were separated into 3 data sets, all stratified by age, sex, and FH-carrying status: (1) feature importance data set (20% of participants) using 3-fold cross-validation, (2) derivation data set (60%) using 5-fold cross-validation, and (3) validation data set (20%) (Figure 1). The relatively low numbers of folds used in the feature importance and derivation data sets were arbitrarily chosen to ensure that each fold contained a sufficient number of participants across various age groups, by sex, and FH-carrying status. Missing data imputation and data transformations were applied separately within each data set and cross-validation fold to avoid data leakage that may lead to bias. Unscaled versions of the imputed derivation and validation data sets were kept to respectively recalibrate the clinical diagnostic criteria and FAMCAT and to establish the performances of the comparators in the validation data set (Data S4).

Statistical and ML Models

The UKB variables included in our analysis were selected on the basis of the established definitions of how FH can be clinically identified, including variables used as predictors in diagnostic criteria and other variables readily available within EHRs. Each feature's importance in predicting the outcome was established by deriving an eXtreme Gradient Boosting (XGBoost)

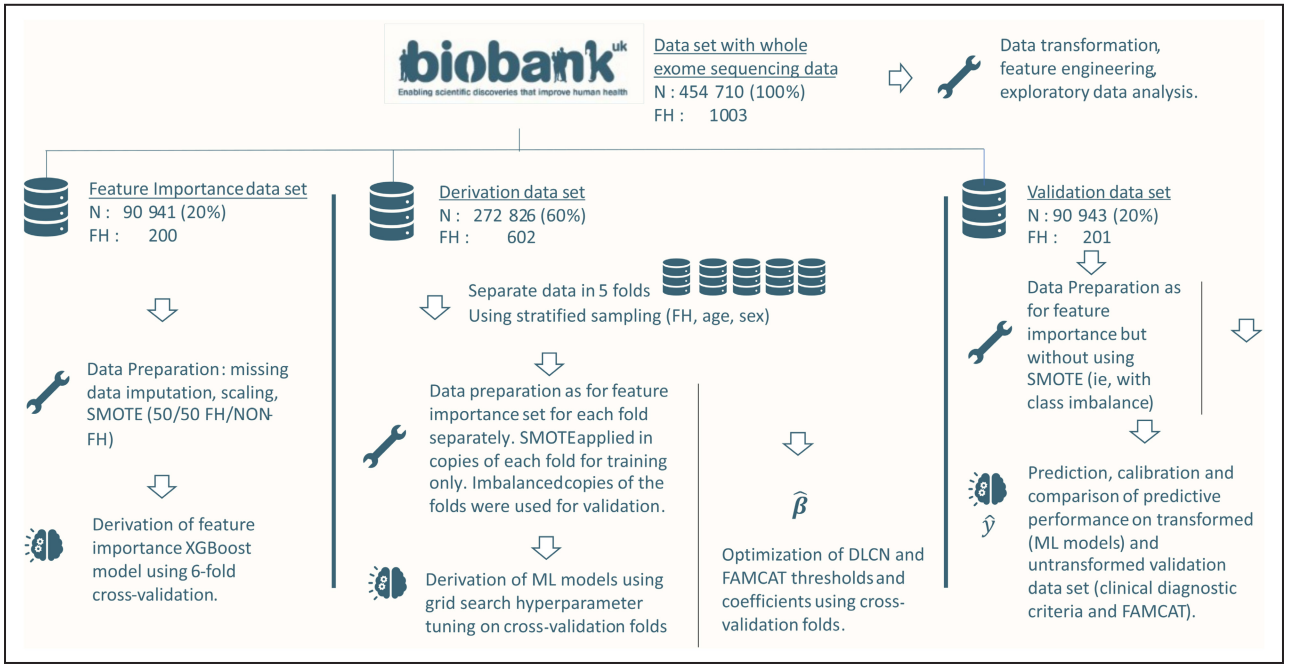

Figure 1. Data partitioning and ML methodology.
FH indicates familial hypercholesterolemia; DLCN, Dutch Lipid Clinic Network; FAMCAT, Familial Hypercholesterolemia Case Ascertainment Tool; ML, machine learning; SMOTE, synthetic minority oversampling technique; and XGBoost, eXtreme Gradient Boosting.

model on the feature importance data set. New models were derived using 9 algorithms (logistic regression [LR], least absolute shrinkage and selection perator (L1), Ridge (L2), Elastic Net(L1+L2), random forest, Support Vector Machines, Artificial Neural Networks, XGBoost, Stacked Ensemble) and different hyperparameter configurations within our cross-validation scheme (Data S5). The synthetic minority oversampling technique was used to decrease the effect of class imbalance, but it was not performed when assessing the performance of our models in the validation folds of our cross-validation scheme and within the validation data set.

Comparators: Clinical Diagnostic Criteria and FAMCAT Algorithm

The DLCN, Simon Broome, and Make Early Diagnosis to Prevent Early Death clinical diagnostic criteria were adapted, keeping only variables present in the UKB, to compare their predictive performances against our models. This is a common practice when applying clinical diagnostic criteria to EHRs, as some variables are infrequently recorded in clinics and unavailable in EHRs.[14] The variable indicative of a positive genetic test was removed from the Simon Broome and DLCN criteria, as this is the primary outcome of our study. We pooled the different diagnosis categories into "unlikely" and "possible" FH for each clinical diagnostic criteria. For the DLCN criteria, we scaled the total score achieved by a patient between 0 and 1 and applied a receiver operating characteristic curve analysis to establish the threshold with the highest combined sensitivity and specificity within the derivation data set. We also implemented a modified version of FAMCAT, a recent EHRs-derived linear model (Data S6).[15]

Statistical Analysis

The sensitivity and positive predictive value (PPV) metrics were used to select the best model in the derivation data set's validation folds, whereas the sensitivity, PPV, specificity, area under the receiver operating characteristic curve (AUC), logloss, and negative predictive value were used to report models' performances in the validation data set. The calibration of our final model and comparators was calculated using the Hosmer–Lemeshow goodness-of-fit statistic, Spiegelhalter's z-test, Brier score statistics, and calibration plots. The reciprocal of the PPV, which is conceptually equivalent to the NNS for an FH variant carrier, was also used.

Clinical characteristics were compared using χ^2 tests for binary variables, hypergeometric tests for categorical (nonbinary) variables, and Wilcoxon rank-sum tests (or Mann–Whitney *U* test) for continuous variables. Statistical significance was defined as $P<0.05$ with Bonferroni correction. Statistical analyses were conducted using R software version 4.0.2 (R Foundation for Statistical Software, Vienna, Austria).

This article complies with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis statement (Data S7).[16]

RESULTS

In the first step of our analysis, we annotated the genetic variants of the 454 710 UKB participants included in our analysis and found 1003 carriers of known pathogenic variants; the prevalence of genetically defined FH was 1:453 (95% CI, 1:426–1:482). The list of genetic variants identified and the number of genetic variants linked to FH, their pathogenicity, the numbers of carriers by gene, their treatment status, and associated LDL-C levels are available in Tables S1 and S2.

Individuals with a pathogenic variant in the *LDLR* gene, the most common locus for pathogenic variants in our data set (n=999/1003 [99.6%]), had higher LDL-C levels than carriers of benign variants (+0.51 mmol/L median difference overall, q-value <0.0001) regardless of their treatment status. The 2 carriers of a pathogenic variant in the *APOB* or *PCSK9* genes (n=2/1003; 0.2% in each gene) had higher median differences in LDL-C levels (*APOB*: +1.01 mmol/L; *PCSK9*: +1.14 mmol/L) than carriers of benign variants but lacked sufficient power to meet the significance threshold (*q*-value >0.99 and 0.43, respectively). No homozygous or double-heterozygous carriers of pathogenic variants were identified, but 5 individuals were compound heterozygotes in the *LDLR* gene (Table S3). In Figure 2, a violin plot shows LDL-C levels as a function of treatment and variant pathogenicity. LDL-C levels in this plot largely overlap but with noticeable right skew in the LDL-C levels of pathogenic variants carriers, which illustrates the difficulty in identifying individuals with FH on the basis of LDL-C levels alone.

Table 1 and Table S4 summarize the clinical characteristics of the UKB participants stratified by FH-carrying status (ie, participants with FH: individuals carrying [likely] pathogenic variants; participants without FH: individuals not carrying any variant, carrying [likely] benign variants, or variants of uncertain significance), and the strength and directions of association between each clinical feature and FH status. Statistically significant *P* values, marked by an asterisk, were defined as those <0.0003 (ie, Bonferroni adjusted α of 0.05 divided by 166 tests). Women were 0.052% more likely to carry a pathogenic FH variant than men (*P**=0.0002) and median age was similar in individuals with and without FH (58 years). More than 95% of participants were White individuals.

Differences in comorbidities, procedures of the heart and arteries, family history, laboratory measurement, and prescriptions between individuals with and without FH are available in Table 1 and summarized here for convenience. No significant differences were found in traditional cardiovascular disease risk factors, such as hypertension, diabetes, body mass index, and smoking between individuals with and without FH. Absolute differences between these 2 groups show that prevalence of ASCVD, CAD, and peripheral artery disease were higher in individuals with FH (+8.02% ASCVD, +4.96% premature ASCVD; +9.38% CAD; +2.4% peripheral artery disease; *P* *≤0.0001), whereas the prevalence of cerebral vascular disease overall or stroke specifically did not significantly differ by FH status. Percutaneous coronary procedures were more frequent among individuals with FH occurring in 8.97% versus 5.64%, which represents an absolute difference of +3.34% (*P**<0.0001). Patients with FH had higher proportions of self-reported family history of heart disease (57.93% in FH versus 44.08% in non-FH) or documented (*ICD-10*) family history of CAD (10.37% in FH versus 4.96% in non-FH), which represent absolute differences of +13.85% and +5.40%, respectively (*P**<0.0001).

In terms of lipid measurements, the levels of total cholesterol, non–high-density lipoprotein cholesterol, and

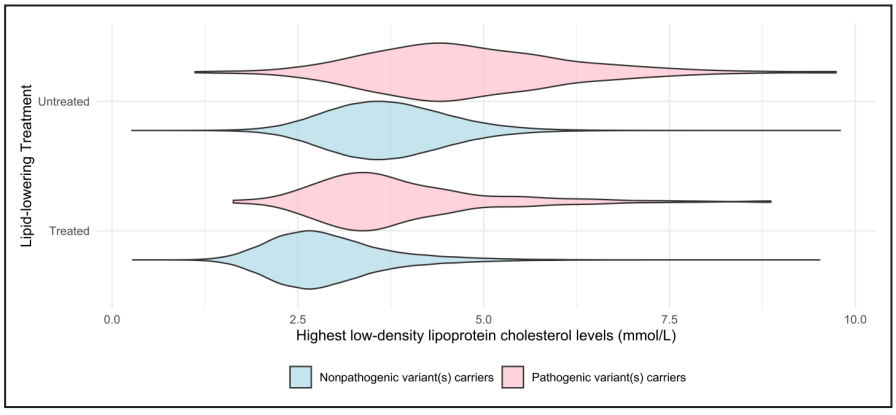

Figure 2. Violin plot of LDL-C levels by pathogenicity of carried variants and treatment status in the UK Biobank participants.
LDL-C indicates low density lipoprotein cholesterol.

Table 1. Characteristics of Participants Included in the Study Stratified by FH Status

	Missing data*, %	Not FH (N=453 707)†, n (%)	FH (N=1003)†, n (%)	Difference‡, %	95% CI‡, %	*P* value§,‖
Sex, male	0.00	207 720/453 707 (45.78)	400/1003 (39.88)	−5.90	−8.99 to −2.82%	0.0002**
Sex, female	0.00	245 987/453 707 (54.22)	603/1003 (60.12%)	5.90	2.82 to 8.99%	0.0002**
Age at recruitment, y	0.00	58.00 (50.00 to 63.00)	58.00 (50.00 to 63.00)	0	−1.00 to 0.00	0.42
Race or ethnicity	1.39					
White		427 508/447 484 (95.54)	935/991 (94.35)	−1.19	−2.68 to 0.3	0.09
Asian		10 257/447 484 (2.29)	27/991 (2.72)	0.43	−0.63 to 1.5	0.31
Black		7088/447 484 (1.58)	23/991 (2.32)	0.74	−0.25 to 1.73	0.06
Mixed		2631/447 484 (0.59)	6/991 (0.61)	0.02	−0.48 to 0.52	0.73
Waist, cm	0.22	90.00 (80.00 to 99.00)	89.00 (79.00 to 98.00)	−1.00	−2.00 to 0.00	0.006
Body mass index, kg/m^2	0.39	26.82 (24.21 to 29.99)	26.94 (23.94 to 30.01)	0.12	−0.17 to 0.41	0.97
Body massindex category	0.39					
Overweight		193 143/451 956 (42.73)	417/1001 (41.66)	−1.08	−4.18 to 2.03	0.51
Normal		143 848/451 956 (31.83)	330/1001 (32.97)	1.14	−1.83 to 4.1	0.42
Obese		112 817/451 956 (24.96)	252/1001 (25.17)	0.21	−2.53 to 2.95	0.84
Smoking	0.11					
Never		247 095/453 224 (54.52)	548/1003 (54.64)	0.12	−3.02 to 3.25	0.92
Previous		156 783/453 224 (34.59)	358/1003 (35.69)	1.10	−1.92 to 4.12	0.44
Current		47 501/453 224 (10.48)	89/1003 (8.87)	−1.61	−3.42 to 0.2	0.10
Prefer not to answer		1845/453 224 (0.41)	8/1003 (0.80)	0.39	−0.21 to 0.99	0.05
Diabetes, type 2	0.00	35 378/453 707 (7.80)	79/1003 (7.88)	0.08	−1.64 to 1.8	0.97
Diabetes, any type	0.00	37 462/453 707 (8.26)	86/1003 (8.57)	0.32	−1.47 to 2.1	0.76
Hypertension	0.00	132 295/453 707 (29.16)	304/1003 (30.31)	1.15	−1.75 to 4.05	0.44371
Atherosclerotic cardiovascular disease	0.00					
No disease		385 138/453 707 (84.89)	771/1003 (76.87)	−8.02	−10.68 to −5.36	<0.0001**
Nonpremature disease		50 359/453 707 (11.10)	142/1003 (14.16)	3.06	0.85 to 5.27	0.003
Premature disease		18 210/453 707 (4.01)	90/1003 (8.97)	4.96	3.14,6.78	<0.0001**
Coronary artery disease	0.00					
No disease		403 536/453 707 (88.94)	798/1003 (79.56)	−9.38	−11.93 to −6.83	<0.0001**
Nonpremature disease		36 870/453 707 (8.13)	125/1003 (12.46)	4.34	2.24,6.43	<0.0001**
Premature disease		13 301/453 707 (2.93)	80/1003 (7.98)	5.04	3.32 to 6.77	<0.0001**
Myocardial Infarction	0.00					
No disease		432 260/453 707 (95.27)	919/1003 (91.63)	−3.65	−5.41,-1.88	<0.0001**
Nonpremature disease		16 956/453 707 (3.74)	60/1003 (5.98)	2.24	0.73 to 3.76	<0.001
Premature disease		4491/453 707 (0.99)	24/1003 (2.39)	1.40	0.41 to 2.4	0.0001**
Angina	0.00					
No disease		425 360/453 707 (93.75)	879/1003 (87.64)	−6.12	−8.2 to −4.03	<0.0001**
Nonpremature disease		21 785/453 707 (4.80)	85/1003 (8.47)	3.67	1.9 to 5.45	<0.0001**
Premature disease		6562/453 707 (1.45)	39/1003 (3.89)	2.44	1.2 to 3.69	<0.0001**
Peripheral artery disease	0.00					
No disease		439 723/453 707 (96.92)	948/1003 (94.52)	−2.40	−3.86 to −0.94	0.0001**
Nonpremature disease		11 620/453 707 (2.56)	44/1003 (4.39)	1.83	0.51 to 3.14	<0.001
Premature disease		2364/453 707 (0.52)	11/1003 (1.10)	0.58	−0.12 to 1.27	0.02
Cerebrovascular disease	0.00					
No disease		438 102/453 707 (96.56)	976/1003 (97.31)	0.75	−0.31,1.8	0.16

(*Continued*)

Table 1. Continued

	Missing data*, %	Not FH (N=453 707)†, n (%)	FH (N=1003)†, n (%)	Difference‡, %	95% CI‡, %	*P* value§,‖
Nonpremature disease		12 106/453 707 (2.67)	20/1003 (1.99)	−0.67	−1.59 to 0.24	0.21
Premature disease		3499/453 707 (0.77)	7/1003 (0.70)	−0.07	−0.64 to 0.49	0.98
Stroke	0.00					
Nonpremature disease		441 905/453 707 (97.40)	985/1003 (98.21)	0.81	−0.07 to 1.68	0.08
No disease		8697/453 707 (1.92)	12/1003 (1.20)	−0.72	−1.44 to 0	0.11
Premature disease		3105/453 707 (0.68)	6/1003 (0.60)	−0.09	−0.61 to 0.44	0.94
Heart percutaneous intervention	0.00	25 570/453 707 (5.64)	90/1003 (8.97)	3.34	1.52 to 5.16	<0.0001**
Arteries percutaneous intervention	0.00	9386/453 707 (2.07)	36/1003 (3.59)	1.52	0.32 to 2.72	0.001
Family history of heart disease (self-assessed)	0.00	199 975/453 707 (44.08)	581/1003 (57.93)	13.85	10.74 to 16.96	<0.0001**
Family history of coronary artery disease (documented)	0.00	22 524/453 707 (4.96)	104/1003 (10.37)	5.40	3.47 to 7.34	<0.0001**
Total cholesterol, mmol/L	4.44	5.66 (4.92 to 6.44)	6.24 (5.34 to 7.48)	0.58	0.46 to 0.72	<0.0001**
Non-HDL cholesterol, mmol/L	12.30	4.19 (3.49 to 4.94)	4.73 (3.90 to 5.87)	0.54	0.46 to 0.74	<0.0001**
HDL cholesterol, mmol/L	12.29	1.40 (1.17 to 1.68)	1.39 (1.18 to 1.66)	−0.01	−0.05 to 0.03	0.64
LDL cholesterol (measured), mmol/L	4.61	3.53 (2.95 to 4.13)	4.07 (3.33 to 5.06)	0.55	0.43 to 0.62	<0.0001**
Triglycerides, mmol/L	4.51	1.49 (1.05 to 2.16)	1.34 (0.97 to 1.91)	−0.15	−0.21 to −0.10	<0.0001**
Apolipoprotein A, g/L	12.77	1.51 (1.35 to 1.70)	1.47 (1.31 to 1.64)	−0.04	−0.06 to −0.03	<0.0001**
Apolipoprotein B, g/L	4.91	1.02 (0.87 to 1.18)	1.18 (1.01 to 1.40)	0.16	0.14 to 0.18	<0.0001**
C-reactive protein, mg/L	4.64	1.34 (0.66 to 2.79)	1.15 (0.52 to 2.44)	−0.19	−0.26 to −0.13	<0.0001**
Glycated hemoglobin, HbA_{1c}, mmol/mol	4.66	35.30 (32.80 to 38.00)	35.70 (33.20 to 38.50)	0.4	0.10 to 0.80	<0.001**
Statin	0.00	80 846/453 707 (17.82)	491/1003 (48.95)	31.13	27.99 to 34.28	<0.0001**
Ezetimibe	0.00	3546/453 707 (0.78)	126/1003 (12.56)	11.78	9.68 to 13.88	<0.0001**
Fibrate	0.00	1549/453 707 (0.34)	14/1003 (1.40)	1.05	0.28 to 1.83	<0.0001**
Bile acid sequestrant	0.00	301/453 707 (0.07)	7/1003 (0.70)	0.63	0.07 to 1.2	<0.0001**
β Blocker	0.00	33 796/453 707 (7.45)	130/1003 (12.96)	5.51	3.38 to 7.64	<0.0001**
Antiplatelet	0.00	67 280/453 707 (14.83)	246/1003 (24.53)	9.70	6.98 to 12.41	<0.0001**
Medication: ATC A	0.00	174 494/453 707 (38.46)	456/1003 (45.46)	7	3.87 to 10.14	<0.0001**
Medication: ATC B	0.00	81 857/453 707 (18.04)	280/1003 (27.92)	9.87	7.05 to 12.7	<0.0001**
Medication: ATC C	0.00	196 087/453 707 (43.22)	656/1003 (65.40)	22.18	19.19 to 25.18	<0.0001**

ATC indicates Anatomical Therapeutic Chemical classification system (A, alimentary tract and metabolism; B, blood and blood-forming organs; C, cardiovascular system); FH, familial hypercholesterolemia; HbA_{1c}, glycated hemoglobin; HDL, high-density lipoprotein; and LDL-C, low-density lipoprotein.

*Percentage of missing values overall. Morbidities, family history, operation, and medication values were assumed to be negative when no codes were found.

†Median (25% to 75%) for numerical; count/$N_{case/controls}$ (%) for categorical variable (binary or >2 levels).

‡Difference in percentage for categorical variables and in median unit for continuous variables.

§Wilcoxon test rank sum test for continuous variables; 2-sample test for equality of proportions for binary categorical variable; hypergeometric test for categorical variables (>2 levels).

‖*P* values marked with a superscript double star "**" are statistically significant at the 5% significance level after Bonferroni correction for 166 tests.

LDL-C were 0.54 to 0.58 mmoL/L higher (0.2 mmoL/L higher for apolipoprotein B) among individuals with FH compared with individuals without FH (*P**<0.0001). By contrast, the triglyceride levels were 0.2 mmoL/L lower among individuals with FH versus those without FH (*P**<0.0001). The proportions of documented prescriptions of lipid-lowering medications (LLMs; statins, ezetimibe, fibrates, or bile acid sequestrants) were higher in individuals with FH than in individuals without FH (+31%, +12%, +1%, and +0.63% higher, respectively, *P**<0.0001). Similarly, the proportions of documented prescriptions of non-LLMs (antiplatelets and β blockers) were 9.70% and 5.51% higher for individuals with FH than for individuals without FH (*P**<0.0001).

Figure 3 shows the 10 most important features identified by our XGBoost Feature Importance algorithm (Data S8) that yielded a 3-fold cross-validated AUC of 86% in the feature importance data set. The 4 most important features were LDL-C levels, statin use, triglycerides, and apolipoprotein B levels. Other important predictors included waist circumference, a recorded family history of CAD, age, diastolic blood pressure, apolipoprotein A levels, and the use of any medication linked to

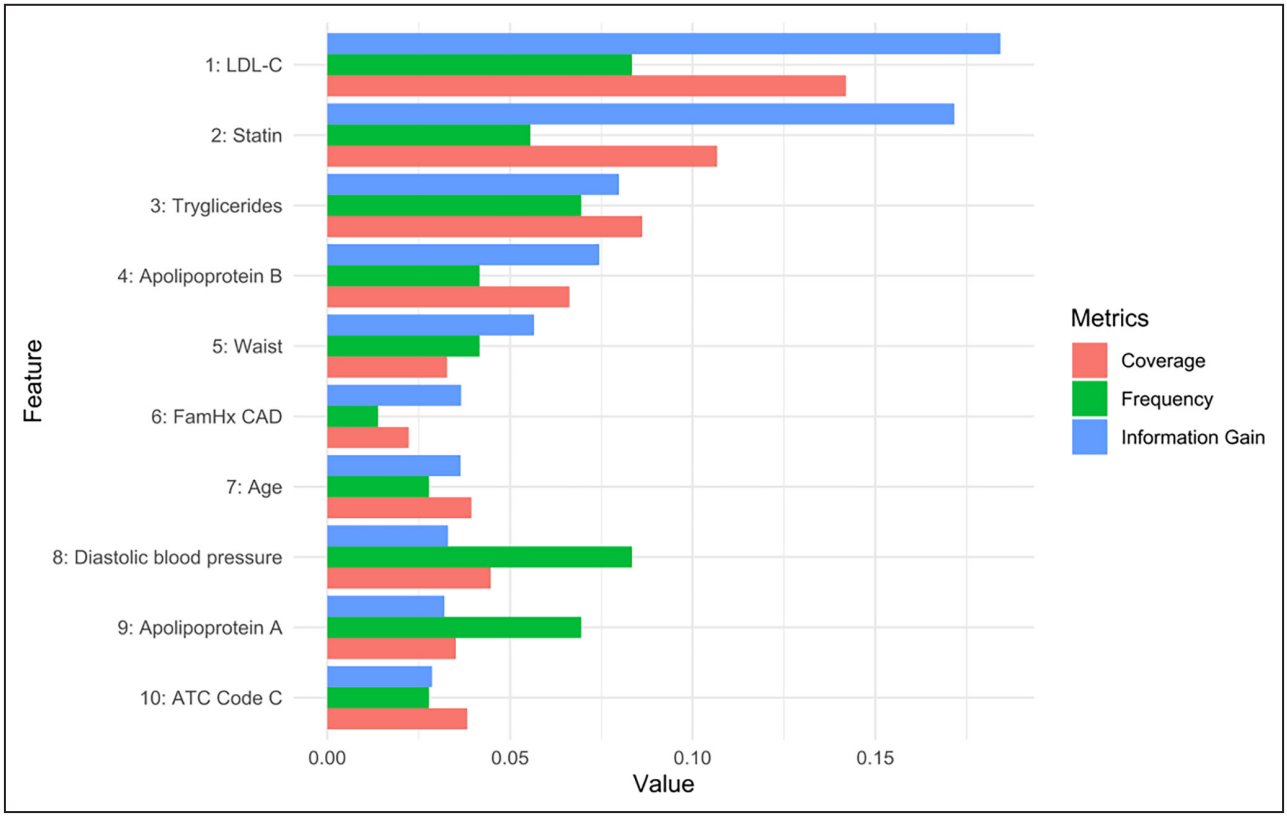

Figure 3. Ten most important features using XGBoost feature importance and the information gain metrics in the UK Biobank.
Information gain quantifies the reduction in uncertainty or randomness achieved by including a specific feature in a model. Coverage refers to the proportion of data points or instances explained by a specific feature. Frequency refers to the relative number of instructions in which a feature was used to predict the outcome by an algorithm. ATC code C indicates Anatomical Therapeutic Chemical code for medications that impact the cardiovascular system; FamHx CAD, family history of coronary artery disease; and LDL-C, low density lipoprotein cholesterol.

the cardiovascular disease system (ie, Anatomical Therapeutic Chemical code C).

We derived a series of models that included 27 predictor variables and the FH status as the binary outcome variable. We report our models' performances in the training and validation data set in Table 2. The final list of predictors was chosen on the basis of the results of our feature importance analysis, statistical significance, clinical knowledge, and expected availability within EHRs (Table S5). The 7 ML algorithms initially included in our analysis (ie, LR, least absolute shrinkage and selection operator, Elastic Net, Ridge, random forest, XGBoost, Artificial Neural Networks, and Support Vector Machines) provided similar performances on the validation folds of the derivation data set and in the validation data set. However, the random forest algorithm provided marginally higher PPV than the other algorithms.

We generated 3 stacking ensembles using (1) all our initial ML-derived models (ie, excluding LR) as base learners; (2) the top 5 models in terms of PPV as base learners; and (3) the best model, in terms of PPV, for each of our 7 initial algorithms as base learners. The stacking ensembles yielded lower logloss than any of our initial models, meaning that their probabilistic estimates for each observation were closer to the observed FH status. The stacking ensemble model combining the top 5 models by PPV was chosen as our best model due to its ability to retain a high sensitivity, specificity, and AUC in the cross-validation folds of the derivation data set (Table 2).

In Figure 4 and Table 3, we compared the performances of our chosen model in the validation data set, that is, the stacking ensemble, to those of modified clinical diagnostic criteria and FAMCAT. In terms of AUC, our model outperformed other criteria and our LR model, but especially the clinical diagnostic criteria methods (AUC ML-derived model, 79.12%; LR, 77.05%; FAMCAT, 71.12%; DLCN, 67.13%; Simon Broome, 61.30%). Our model had the highest sensitivity (ML-derived model, 74.93%; LR, 69.10%; FAMCAT, 69.55%; DLCN, 64.23%; Make Early Diagnosis to Prevent Early Death, 7.16%; Simon Broome, 35.12%) and was also more specific and precise than screening

and diagnostic criteria with sensitivity >50% but less specific and precise than the logistic regression model (specificity: ML-derived model, 72.80%; LR, 78.10%; FAMCAT, 65.43%; DLCN, 62.72%).The NNS of our model and LR were also lower than those of screening and diagnostic criteria overall (ML-derived model, 164 [95% CI, 154–175]; LR, 144 [95% CI, 130–164]; FAMCAT, 227 [95% CI, 208–250]; DLCN, 263 [95% CI, 239–293]) and across tertiles of age (Table 3), more notably in older individuals. While the difference in sensitivity and NNS among the different approaches may seem small, when applying these approaches at the population level, the numbers become clinically relevant. Table 4 shows the difference in the yield of patients with FH identified and number of tests required by our model compared with DLCN, FAMCAT, and our logistic regression model. Our model had higher calibration performance than FAMCAT (Hosmer–Lemeshow test: P=0.015 versus <2.2e-16; Spiegelhalter's z-test: P=0.001 versus <2.2e-16; Figure 5 and Table S6).

DISCUSSION

The present study demonstrates that ML algorithms could improve the detection of individuals likely to have genetically confirmed FH. Our top-performing model exhibits superior predictive accuracy compared with FAMCAT, the algorithm currently recommended in primary care settings in England. Moreover, application of our ML algorithm would reduce the NNS for genetic confirmation by about one-quarter, which would reduce unnecessary referrals to lipid clinics and potentially unnecessary genetic testing, a relevant consideration at a time of increasing health care workload and ever restricted health care budgets. Our results suggest that the advantage of using ML over simpler models, like our logistic regression model, is mainly reflected in the sensitivity metric and at the expense of specificity and PPV. Nonetheless, even a minor improvement in sensitivity could lead to a substantial increase in identifying FH through population screening. Additionally, the explainability (ie, the capability to comprehend how the model arrived at its prediction) offered by simpler models may be less essential in screening compared with diagnostic applications. To date, in England, the path to diagnosing FH typically begins with a clinician's suspicion on the basis of factors like premature CAD, severely elevated cholesterol levels, or a suggestive family history, followed by application of the Simon Broome criteria or other clinical diagnostic criteria to assess the likelihood of an FH diagnosis, which should be subsequently confirmed by a specialist, usually through genetic testing. FAMCAT offers an alternative approach to screening primary care EHRs for potential FH cases, leading to specialist referrals. This automated approach enhances detection rates and makes genetic testing more cost-effective by prioritizing high-probability candidates. Yet the validity of screening algorithms relies on the accuracy and completeness of the underlying data set. Real-world EHRs vary in data quality and frequently show missing values. Anamnesis-related variables, especially family history, are less likely to be recorded compared with procedure codes, laboratory results, and medication.

Consequently, when developing a screening algorithm, it is imperative to emphasize commonly recorded variables to ensure its applicability in routine clinical practice. For example, although the DLCN criteria is the most widely used clinical diagnostic criteria in specialist secondary or tertiary care clinics[17,18] and the best performing among all diagnostic criteria in terms of predictive accuracy in the validation data set, our findings suggest that for improving pretest probability in EHRs, it is less useful than our model or FAMCAT. This may in part reflect the fact that clinical signs used in specialist clinics are not routinely captured in nonspecialist settings, and hence modified clinical diagnostic criteria did not perform well compared with newer models for EHR screening. Furthermore, FAMCAT necessitates data related to family history of elevated cholesterol and FH, which may not be consistently available, or incorrectly reported/captured. In contrast, our model encompasses a single family history variable, along with other variables, such as 10 prescribed medications and procedures that are more likely to be consistently recorded in EHRs. A projection of the clinical utility of our model and the currently recommended screening criteria (ie, FAMCAT) show that our model could identify 74.93% of the 106 109 patients with FH aged ≥40 years in the United Kingdom versus 69.55% with FAMCAT (ie, 79 507 versus 73 798). This represents an additional 5709 patients identified using our model. Furthermore, with an NNS of 167, our model would require 13 039 148 tests to identify 79 507 patients, but with an NNS of 227, contrary FAMCAT would require 16 752 146 tests to identify 73 798 patients. This means that, at the population level, our model would require around 4 million fewer tests to identify around 6000 more patients with FH than FAMCAT. Therefore, integrating our model into EHRs as a pretest tool to prioritize those most likely to yield a positive molecular diagnosis would translate into considerable financial savings for health care systems.

In contrast with clinical diagnostic criteria and FAMCAT, which respectively rely on point-based scoring system and statistical linear models, our model is developed through ML techniques that excel in capturing high-degree interactions and nonlinear associations among predictor variables and outcomes.[19] Several studies have examined the use of ML algorithms for identifying patients with FH in different countries and EHR data sets, including the United States,

Table 2. Machine Learning Model Performances on the Validation Folds of the Cross-Validation Derivation Data Set and the Validation Data Set

No.	Algorithm	Validation cross-validation folds derivation data set, mean (SD on CV folds)							Validation data set, mean (SD on 10 bootstrapped samples)						
		Sensitivity*	Specificity*	Accuracy*	PPV*	NPV*	Logloss†	AUC*	Sensitivity*	Specificity*	Accuracy*	PPV*	NPV*	Logloss†	AUC*
0	Logistic regression	71.93 (1.50)	78.78 (2.24)	78.76 (2.23)	0.75 (0.07)	99.92 (0.00)	0.40 (0.02)	82.07 (1.12)	69.10 (3.96)	78.10 (0.17)	78.08 (0.17)	0.69 (0.04)	99.91 (0.01)	0.48 (0.00)	77.05 (2.74)
1	LASSO	55.81 (3.96)	64.33 (2.85)	64.31 (2.84)	0.35 (0.02)	99.85 (0.01)	0.62 (0.01)	65.64 (2.37)	57.66 (3.34)	64.97 (0.16)	64.96 (0.16)	0.36 (0.02)	99.86 (0.01)	0.66 (0.00)	66.99 (2.10)
2	LASSO‡,§	71.75 (2.38)	79.83 (0.93)	79.81 (0.93)	0.78 (0.03)	99.92 (0.01)	0.43 (0.02)	81.89 (1.64)	64.28 (3.00)	81.38 (0.17)	81.34 (0.17)	0.76 (0.04)	99.90 (0.01)	0.48 (0.00)	77.87 (1.63)
3	LASSO	72.08 (2.71)	79.00 (1.45)	78.99 (1.44)	0.76 (0.05)	99.92 (0.01)	0.39 (0.01)	81.98 (1.18)	69.50 (3.28)	75.47 (0.09)	75.46 (0.08)	0.62 (0.03)	99.91 (0.01)	0.50 (0.00)	77.22 (1.83)
4	LASSO	71.94 (1.82)	79.45 (1.80)	79.43 (1.80)	0.77 (0.07)	99.92 (0.01)	0.40 (0.01)	81.86 (0.88)	65.62 (2.00)	80.02 (0.14)	79.99 (0.14)	0.72 (0.02)	99.90 (0.01)	0.50 (0.00)	78.32 (1.14)
5	LASSO	71.75 (1.66)	79.57 (1.49)	79.55 (1.48)	0.77 (0.04)	99.92 (0.00)	0.40 (0.00)	82.46 (1.14)	68.21 (3.39)	75.67 (0.12)	75.65 (0.12)	0.62 (0.03)	99.91 (0.01)	0.51 (0.00)	77.22 (1.81)
6	LASSO	72.26 (3.75)	78.40 (1.79)	78.39 (1.78)	0.74 (0.03)	99.92 (0.01)	0.40 (0.01)	81.81 (1.26)	69.30 (2.39)	77.12 (0.07)	77.10 (0.07)	0.67 (0.02)	99.91 (0.01)	0.52 (0.00)	79.30 (1.48)
7	Elastic Net	67.93 (6.35)	69.30 (3.09)	69.30 (3.07)	0.49 (0.04)	99.90 (0.02)	0.56 (0.01)	74.59 (1.75)	65.82 (2.18)	68.67 (0.11)	68.66 (0.11)	0.46 (0.02)	99.89 (0.01)	0.61 (0.00)	74.13 (1.38)
8	Elastic Net	71.76 (3.55)	79.49 (2.28)	79.48 (2.27)	0.77 (0.08)	99.92 (0.01)	0.42 (0.01)	81.82 (1.14)	70.75 (2.61)	76.55 (0.17)	76.54 (0.17)	0.66 (0.03)	99.92 (0.01)	0.49 (0.00)	79.58 (1.45)
9	Elastic Net‡,§	71.59 (3.52)	79.83 (2.04)	79.81 (2.03)	0.78 (0.07)	99.92 (0.01)	0.40 (0.01)	82.30 (0.86)	68.51 (3.26)	77.75 (0.18)	77.73 (0.18)	0.68 (0.03)	99.91 (0.01)	0.46 (0.00)	78.85 (1.73)
10	Elastic Net	72.91 (2.53)	77.86 (1.57)	77.85 (1.56)	0.72 (0.03)	99.92 (0.01)	0.40 (0.01)	82.04 (1.29)	68.56 (3.32)	79.36 (0.09)	79.34 (0.09)	0.73 (0.03)	99.91 (0.01)	0.48 (0.00)	78.87 (1.90)
11	Elastic Net§	71.26 (3.02)	80.08 (1.15)	80.06 (1.15)	0.79 (0.05)	99.92 (0.01)	0.40 (0.01)	82.11 (1.29)	67.76 (3.65)	79.80 (0.14)	79.77 (0.13)	0.74 (0.04)	99.91 (0.01)	0.48 (0.00)	79.03 (1.93)
12	Elastic Net	71.77 (1.98)	78.16 (1.22)	78.14 (1.21)	0.72 (0.03)	99.92 (0.00)	0.40 (0.01)	81.76 (0.72)	68.31 (1.86)	79.37 (0.16)	79.35 (0.17)	0.73 (0.02)	99.91 (0.01)	0.48 (0.00)	79.39 (1.11)
13	Ridge	70.43 (3.57)	73.70 (2.32)	73.69 (2.31)	0.59 (0.04)	99.91 (0.01)	0.50 (0.01)	78.41 (1.53)	69.00 (3.45)	75.17 (0.16)	75.15 (0.15)	0.61 (0.03)	99.91 (0.01)	0.54 (0.00)	76.42 (1.65)
14	Ridge	72.09 (1.83)	78.15 (1.32)	78.13 (1.31)	0.73 (0.03)	99.92 (0.00)	0.42 (0.01)	81.80 (1.15)	70.85 (3.41)	76.14 (0.22)	76.12 (0.22)	0.65 (0.03)	99.92 (0.01)	0.49 (0.00)	79.48 (1.47)
15	Ridge	70.60 (2.22)	79.15 (1.05)	79.13 (1.05)	0.74 (0.04)	99.92 (0.01)	0.41 (0.01)	82.36 (0.91)	69.90 (2.97)	76.77 (0.13)	76.76 (0.13)	0.66 (0.03)	99.91 (0.01)	0.49 (0.00)	78.31 (2.14)
16	Ridge	71.60 (2.48)	79.06 (2.22)	79.04 (2.21)	0.75 (0.06)	99.92 (0.01)	0.39 (0.01)	81.86 (1.08)	66.42 (3.17)	79.19 (0.12)	79.16 (0.12)	0.70 (0.03)	99.91 (0.01)	0.48 (0.00)	78.16 (2.11)
17	Ridge‡	72.25 (1.60)	79.22 (1.45)	79.20 (1.45)	0.77 (0.04)	99.92 (0.00)	0.40 (0.01)	82.13 (1.33)	68.61 (3.35)	77.32 (0.15)	77.30 (0.15)	0.67 (0.03)	99.91 (0.01)	0.48 (0.00)	77.25 (2.15)
18	Ridge	71.44 (3.26)	79.47 (1.77)	79.45 (1.76)	0.77 (0.04)	99.92 (0.01)	0.41 (0.02)	82.11 (0.86)	66.07 (4.81)	77.54 (0.15)	77.52 (0.15)	0.65 (0.05)	99.90 (0.01)	0.50 (0.00)	76.89 (2.49)
19	RF	68.27 (3.72)	79.66 (2.39)	79.64 (2.38)	0.74 (0.06)	99.91 (0.01)	0.37 (0.01)	81.19 (0.40)	68.96 (2.68)	76.50 (0.11)	76.49 (0.11)	0.65 (0.03)	99.91 (0.01)	0.48 (0.00)	79.11 (1.58)
20	RF§	68.61 (2.93)	81.23 (2.40)	81.20 (2.39)	0.81 (0.09)	99.91 (0.01)	0.37 (0.01)	81.79 (0.45)	62.29 (4.61)	76.35 (0.08)	76.32 (0.08)	0.58 (0.04)	99.89 (0.01)	0.54 (0.00)	78.42 (1.78)
21	RF	68.11 (1.85)	79.89 (1.63)	79.87 (1.63)	0.75 (0.07)	99.91 (0.01)	0.36 (0.01)	80.93 (1.03)	68.41 (2.84)	77.91 (0.09)	77.89 (0.09)	0.68 (0.03)	99.91 (0.01)	0.48 (0.00)	79.32 (1.78)
22	RF	69.24 (4.08)	79.83 (2.38)	79.81 (2.37)	0.76 (0.05)	99.92 (0.01)	0.37 (0.02)	81.58 (0.57)	67.16 (4.31)	79.23 (0.16)	79.20 (0.16)	0.71 (0.05)	99.91 (0.01)	0.51 (0.00)	78.68 (1.86)
23	RF‡	70.25 (4.33)	80.23 (0.34)	80.21 (0.34)	0.78 (0.05)	99.92 (0.01)	0.37 (0.01)	81.65 (0.75)	64.38 (3.99)	83.28 (0.14)	83.24 (0.14)	0.85 (0.05)	99.91 (0.01)	0.45 (0.00)	80.25 (1.78)
24	RF	70.60 (4.74)	79.29 (2.59)	79.27 (2.58)	0.75 (0.07)	99.92 (0.01)	0.36 (0.02)	81.94 (0.71)	64.53 (3.26)	78.82 (0.12)	78.79 (0.12)	0.67 (0.04)	99.90 (0.01)	0.49 (0.00)	78.05 (1.88)
25	RF	68.27 (1.67)	79.42 (1.05)	79.40 (1.04)	0.73 (0.04)	99.91 (0.00)	0.37 (0.00)	81.69 (0.75)	59.50 (3.87)	80.71 (0.12)	80.66 (0.12)	0.68 (0.05)	99.89 (0.01)	0.48 (0.00)	77.50 (1.97)
26	RF	71.59 (4.13)	78.60 (1.58)	78.58 (1.57)	0.74 (0.02)	99.92 (0.01)	0.36 (0.02)	81.88 (0.51)	69.55 (3.39)	78.45 (0.16)	78.43 (0.16)	0.71 (0.03)	99.91 (0.01)	0.45 (0.00)	79.80 (1.73)
27	XGBoost‡,§	67.92 (2.87)	80.95 (1.12)	80.93 (1.11)	0.78 (0.02)	99.91 (0.01)	0.45 (0.01)	80.74 (0.83)	66.72 (3.12)	80.30 (0.11)	80.27 (0.11)	0.74 (0.03)	99.91 (0.01)	0.52 (0.00)	78.95 (1.92)

(*Continued*)

Table 2. Continued

No.	Algorithm	Validation cross-validation folds derivation data set, mean (SD on CV folds)							Validation data set, mean (SD on 10 bootstrapped samples)						
		Sensitivity*	Specificity*	Accuracy*	PPV*	NPV*	Logloss†	AUC*	Sensitivity*	Specificity*	Accuracy*	PPV*	NPV*	Logloss†	AUC*
28	XGBoost	68.93 (2.76)	79.23 (1.23)	79.21 (1.22)	0.73 (0.04)	99.91 (0.01)	0.46 (0.01)	80.53 (1.21)	61.94 (5.41)	80.37 (0.14)	80.33 (0.15)	0.69 (0.06)	99.90 (0.01)	0.51 (0.00)	78.83 (1.73)
29	XGBoost§	69.77 (3.51)	80.33 (1.50)	80.31 (1.49)	0.78 (0.05)	99.92 (0.01)	0.41 (0.02)	81.58 (0.85)	61.14 (2.29)	80.70 (0.14)	80.65 (0.14)	0.70 (0.03)	99.89 (0.01)	0.47 (0.00)	78.28 (1.29)
30	XGBoost	71.43 (1.46)	79.14 (1.22)	79.13 (1.22)	0.75 (0.04)	99.92 (0.00)	0.41 (0.01)	81.59 (1.06)	70.70 (4.19)	76.10 (0.17)	76.08 (0.16)	0.65 (0.04)	99.91 (0.01)	0.49 (0.00)	79.54 (1.45)
31	ANN	70.09 (1.78)	79.45 (1.58)	79.43 (1.58)	0.75 (0.04)	99.92 (0.00)	0.41 (0.01)	81.58 (1.27)	68.06 (3.03)	77.68 (0.11)	77.65 (0.11)	0.67 (0.03)	99.91 (0.01)	0.51 (0.00)	78.99 (1.41)
32	ANN‡	71.76 (1.27)	79.26 (2.09)	79.25 (2.09)	0.77 (0.09)	99.92 (0.01)	0.40 (0.00)	81.90 (1.24)	63.73 (2.97)	78.05 (0.11)	78.02 (0.12)	0.64 (0.03)	99.90 (0.01)	0.48 (0.00)	77.21 (1.38)
33	ANN	72.43 (2.10)	77.53 (1.82)	77.52 (1.82)	0.71 (0.04)	99.92 (0.01)	0.39 (0.01)	82.06 (1.16)	67.46 (3.63)	76.92 (0.12)	76.89 (0.12)	0.64 (0.03)	99.91 (0.01)	0.48 (0.00)	78.13 (2.03)
34	ANN	71.41 (3.50)	78.49 (3.02)	78.47 (3.01)	0.74 (0.08)	99.92 (0.01)	0.43 (0.01)	81.97 (1.50)	67.36 (4.38)	79.72 (0.12)	79.69 (0.12)	0.73 (0.05)	99.91 (0.01)	0.49 (0.00)	79.59 (1.77)
35	ANN	70.58 (2.83)	78.76 (1.76)	78.74 (1.75)	0.73 (0.05)	99.92 (0.01)	0.40 (0.01)	81.69 (1.30)	69.70 (4.82)	76.97 (0.13)	76.95 (0.13)	0.67 (0.05)	99.91 (0.01)	0.49 (0.00)	78.45 (2.72)
36	ANN	71.09 (2.78)	79.22 (2.05)	79.20 (2.04)	0.75 (0.05)	99.92 (0.01)	0.40 (0.01)	81.89 (0.87)	69.50 (3.77)	77.74 (0.10)	77.72 (0.10)	0.69 (0.04)	99.91 (0.01)	0.49 (0.00)	79.38 (2.07)
37	SVM	70.92 (3.35)	77.83 (1.43)	77.81 (1.42)	0.70 (0.03)	99.92 (0.01)	0.41 (0.01)	81.30 (1.25)	66.87 (2.30)	74.01 (0.12)	73.99 (0.12)	0.57 (0.02)	99.90 (0.01)	0.49 (0.00)	78.41 (1.16)
38	SVM	71.76 (1.69)	78.74 (0.48)	78.73 (0.47)	0.74 (0.01)	99.92 (0.00)	0.41 (0.00)	81.87 (1.21)	62.74 (3.98)	78.86 (0.13)	78.83 (0.13)	0.65 (0.04)	99.90 (0.01)	0.50 (0.00)	78.21 (2.50)
39	SVM	71.94 (2.45)	78.37 (2.13)	78.36 (2.12)	0.74 (0.08)	99.92 (0.01)	0.41 (0.01)	82.13 (1.24)	60.25 (3.48)	81.12 (0.18)	81.08 (0.18)	0.70 (0.04)	99.89 (0.01)	0.49 (0.00)	77.65 (1.79)
40	SVM‡	71.10 (1.58)	79.80 (1.13)	79.78 (1.13)	0.77 (0.04)	99.92 (0.00)	0.39 (0.01)	82.27 (0.94)	64.08 (2.65)	77.55 (0.14)	77.52 (0.14)	0.63 (0.03)	99.90 (0.01)	0.49 (0.00)	78.35 (2.09)
41	Stacking 40 models	75.92 (2.88)	78.72 (3.85)	78.72 (3.84)	0.80 (0.12)	99.93 (0.01)	0.01 (0.00)	83.21 (0.81)	74.88 (2.94)	74.20 (0.15)	74.20 (0.15)	0.64 (0.03)	99.93 (0.01)	0.01 (0.00)	81.18 (1.59)
42	Stacking 5 models with the highest PPV overall (§)	75.58 (3.73)	79.02 (4.65)	79.01 (4.63)	0.82 (0.15)	99.93 (0.01)	0.01 (0.00)	83.16 (0.85)	74.93 (2.44)	72.80 (0.15)	72.80 (0.14)	0.61 (0.02)	99.92 (0.01)	0.01 (0.00)	79.12 (2.01)
43	Stacking models with highest PPV f (‡)	75.76 (2.60)	78.44 (3.95)	78.44 (3.94)	0.79 (0.10)	99.93 (0.00)	0.01 (0.00)	82.90 (0.97)	74.18 (2.64)	74.17 (0.13)	74.17 (0.13)	0.63 (0.02)	99.92 (0.01)	0.01 (0.00)	79.33 (1.66)

The chosen models in the validation folds of the derivation data set (ie, our "best models") are highlighted in bold font. The 5 best models in terms of PPV are in italics. AUC indicates area under the receiver operating characteristics curve; CV, cross-validation; LASSO, least absolute shrinkage operator; NPV, negative predictive value; PPV, positive predictive value; RF, random forest; SVM: support vector machine; and XGBoost, eXtreme Gradient Boosting.

*Expressed in percentage and SD using probability cutoff obtained using ROC curve analysis on the training (validation cross-validation folds derivation data set) and validation folds (validation data set).

†Expressed in terms of the natural logarithms of the losses.

‡Indicates chosen models in the validation folds of the derivation data set (ie, our "best models").

§Indicates the 5 best models in terms of PPV.

United Kingdom, South Africa, Italy, Sweden, and China. However, only 4 studies have explicitly focused on developing models for identifying FH within the general population.[20–23] In the United States, Banda et al[20] and Myers et al[21] derived models to identify FH among high-risk patients with ASCVD, based on unstructured and structured EHRs, respectively. However, these models require the identification of high-risk populations with established disease, that is, secondary prevention, where the prevalence of FH is much higher,[2] and are therefore less generalizable to primary prevention, where the aim is to identify individuals before a first ASCVD event. Furthermore, considering their definitions of the FH outcome incorporated both clinical and genetic diagnoses, these algorithms would be expected to be sensitive but less specific for the gold standard genetic diagnosis of FH. In the United Kingdom, Gratton et al[22] and Akyea et al[23] developed models specifically for identifying FH within primary care EHRs. However, Akyea and colleagues' definition of the FH outcome was not limited to the molecular diagnosis of FH, making it uncertain whether these were true cases of FH. The study by Gratton et al was conducted using an earlier and smaller UKB data release and focused on 1 type of model, namely, LR, least absolute shrinkage and selection operator, which our study suggests is not the best-performing algorithm. Moreover, their best-performing model

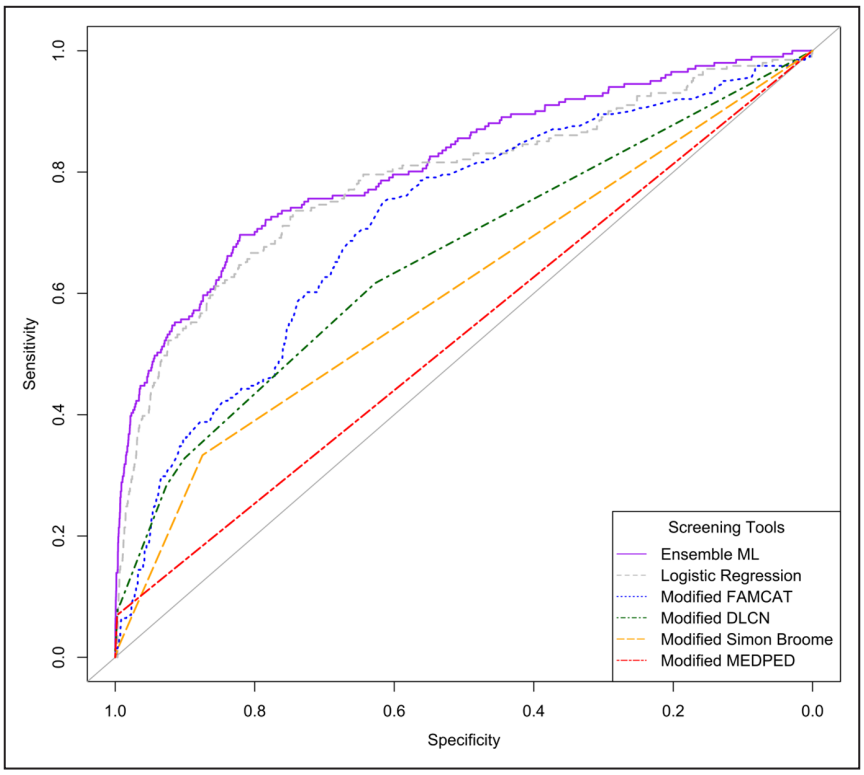

Figure 4. Receiver operating characteristics curve for our chosen model and modified versions of existing screening criteria and clinical diagnostic criteria.
DLCN indicates Dutch Lipid Clinic Network; FAMCAT, Familial Hypercholesterolemia Case Ascertainment Tool; MEDPED: Make Early Diagnosis to Prevent Early Death; and ML, machine learning.

requires the incorporation of a polygenic risk score as an input variable, which necessitates prior genetic testing and has a different aim than the present study, namely, to derive an algorithm to better identify "true" FH by reducing the NNS to ascertain a correct molecular diagnosis.

Table 3. Comparison of Performances Between Our Chosen Model and Existing Clinical Diagnostic Criteria Diagnostic Methods and the FAMACT Algorithm

Algorithm and clinical diagnostic criteria	Validation data set								Validation data set stratified by age terciles		
	Sensitivity*	Specificity*	Accuracy*	PPV*	NPV*	Logloss†	AUC*	NNS	NNS≤53	NNS>53≤61	NNS >61
Chosen "best" stacking model	74.93 (2.44)	72.80 (0.15)	72.80 (0.14)	0.61 (0.02)	99.92 (0.01)	0.01 (0.00)	79.12 (2.01)	164	111	203	187
Logistic regression	69.10 (3.96)	78.10 (0.17)	78.08 (0.17)	0.69 (0.04)	99.91 (0.01)	0.48 (0.00)	77.05 (2.74)	144	106	165	176
FAMCAT	69.55 (3.27)	65.43 (0.11)	65.43 (0.11)	0.44 (0.02)	99.90 (0.01)	0.14 (0.00)	71.12 (1.19)	227	117	237	325
DLCN	64.23 (3.49)	62.72 (0.17)	62.73 (0.17)	0.38 (0.02)	99.87 (0.01)	1.08 (0.00)	67.13 (2.23)	263	196	303	357
MEDPED	7.16 (2.06)	99.73 (0.02)	99.52 (0.02)	5.50 (1.58)	99.79 (0.00)	0.06 (0.00)	NA	18	10	31	35
Simon Broome	35.12 (2.46)	87.49 (0.18)	87.38 (0.18)	0.62 (0.04)	99.84 (0.01)	0.77 (0.00)	61.30 (1.23)	161	108	222	213

AUC indicates area under the receiver operating characteristics curve; DLCN, Dutch Lipid Clinic Network; FAMCAT, Familial Hypercholesterolemia Case Ascertainment Tool; LR, logistic regression; MEDPED, Make Early Diagnosis to Prevent Early Death; NNS, number needed to screen; NPV, negative predictive value; and PPV, positive predictive value.

*Expressed in percentage and SD using probability cutoff obtained using receiver operating characteristics curve analysis on the derivation dataset (cutoffs: Ensemble ML=0.0023, FAMCAT=0.0022, LR=0.0022, DLCN=0.6155, Simon Broome=0.6155, MEDPED=0.5).

†Expressed in terms of the natural logarithms of the losses.

Table 4. Projected Clinical Utility of Our ML-Derived Model Against DLCN and FAMCAT in a Population of 33 Million UK Individuals

Comparator	Differences in the number of patients identified (106 109 individuals with FH)	Differences in the number of genetic tests required to identify FH; based on sensitivity and NNS
ML vs DLCN	Median, 11 352 95% CI, 11 323 to 11 380 *P* value<0.0001	Median, −4 900 935 95% CI, −4 903 916 to −4 897 955 *P* value<0.0001
ML vs FAMCAT	Median, 5726 95% CI, 5698 to 5753 *P* value<0.0001	Median, −3 733 824 95% CI, −3 736 334 to −3 731 313 *P* value<0.0001
ML vs logistic regression	Median, 6181 95% CI, 6150 to 6212 *P* value<0.0001	Median, 2 403 587 95% CI, 2 401 470 to 2 405 705 *P* value<0.0001

DLCN indicates Dutch Lipid Clinic Network; FAMCAT, Familial Hypercholesterolemia Case Ascertainment Tool; FH, familial hypercholesterolemia; ML, machine learning; and NNS, number needed to screen. Corresponds to the number of individuals aged ≥40 y in the UK, and assuming an FH prevalence of 1:311 in the general population (as reported in Hu et al[2]), represents 106 109 individuals with FH.

While large-scale ML-powered screening of EHRs may not seemingly align with contemporary recommendations for universal pediatric screening and reverse child–parent cascade testing,[24,25] it still has a crucial role to play in addressing this genetic disorder. Calls for pediatric screening are driven by the compelling rationale that early detection and timely interventions can significantly reduce FH-associated morbidity and death.[2] However, the implementation of universal screening in children is much debated and rarely implemented globally. Moreover, it is likely that a 2-step approach will be needed with measuring LDL-C levels as a first step, which is currently not routinely undertaken during childhood.[24] Finally, and importantly, although pediatric screening could help identify younger affected parents, there will be many adult patients with FH with no children or whose children are aged >18 years (adults) who would not be identified through reverse cascade testing, unlike with ML-driven EHR screening. Furthermore, the identification of older patients with FH can considerably simplify the cascade genetic testing process, as unequivocal identification of a common ancestor reduces the number of necessary genetic tests. Several limitations merit consideration. While we believe our model may provide a practical tool for improving the identification of true FH cases in adults, which might be implemented into EHRs, we did not test it directly in an EHR system; instead, we evaluated it in the UKB, a large biorepository of volunteers that could have different characteristics. Similarly, our study used adapted versions of conventional FH clinical diagnostic criteria (DLCN, Simon Broome, Make Early Diagnosis to Prevent Early Death) and the FAMCAT algorithm, excluding variables not in the UKB.

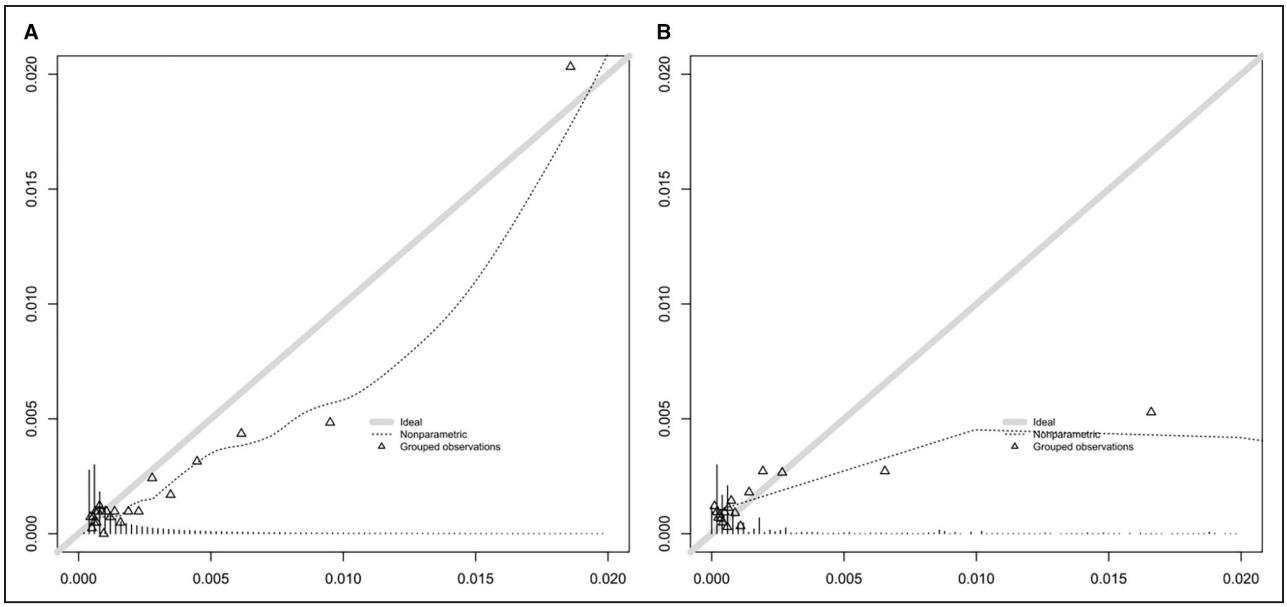

Figure 5. Calibration plots.
The (**A**) Stacking Ensemble ML models and (**B**) FAMCAT algorithm on the validation data set. FAMCAT indicates Familial Hypercholesterolemia Case Ascertainment Tool; and ML, machine learning.

However, this aligns with clinical practice where absent data impact risk prediction. We thus believe that our study reflects actual medical practice in primary care. The UKB recording of prescribed medications is often incomplete, for example, lacking dose details, making it frequently not possible to back-calculate untreated LDL-C levels. This limitation extends to the FAMCAT variables related to medication intensity, and it is also a usual case with EHRs and health databases, revealing constraints imposed by the unavailability of the necessary information in EHRs to back-calculate untreated LDL-C from treated LDL-C levels. This point, however, reflects real-world data recording and, because our model is intended to be applied for screening of EHRs and health databases, we had to consider the LDL-C as recorded. Excluding the large proportion of patients receiving LLMs, or those on LLMs for whom not all needed information is recorded to back-calculate the LDL-C levels, would substantially reduce the applicability and usefulness of the model and will not be appropriate (and bias the results) as treated and untreated patients or those with more or less detailed information recorded may differ between them. Additionally, and importantly, to minimize this issue, being or not being on an LLM was a variable that we also included in the models, thus accounting for whether patients (and so their LDL-C levels) were on treatment. Participants were aged >40 years, mostly White adults, thus limiting generalizability to other age groups and ethnicities who may have different phenotypes, genotypes, and prevalence associated with FH.[2,26] That said, our ML algorithm still provides lower NNS than FAMCAT across all terciles of age, and most adults do not have their first LDL-C measurement before the age of 40 years (Table 3).[7] Additionally, the age range of participants in this study aligns with the national UK National Health Service Health Check screening program. During this period, individuals without preexisting conditions undergo free National Health Service Health Checks every 5 years. This continuous screening provides a consistent flow of clinical information to EHRs, facilitating their potential identification through automated EHR screening for FH. While genetic mutations causing FH are present at birth, and thus the risk of having FH remains constant during one's lifetime, the likelihood for our model to identify individuals carrying an FH mutation may vary with age, as older individuals without FH may develop symptoms that resemble FH due to other health conditions (Table 3). Therefore, we believe that our model could be implemented as part of programs such as the National Health Service Health Check screening program (or similar country-specific programs), starting at age 40 years and repeated every 5 years to account for new data collection and the manifestation of FH-related symptoms. Also, we did not apply the full ClinGen FH Variant Curation Expert Panel variant classification criteria but solely used genetic markers and a database of genetic variants to avoid using the predictors to both define and predict the outcome, which would be a major issue and artificially increase our performance. Consequently, our achieved predictive performance should be interpreted as conservative, with the expectation that efficacy would likely be even more pronounced in populations characterized by higher FH prevalences. We should highlight the methodological limitation of the XGBoost feature importance algorithm, which may have inflated the importance of continuous variables or categorical variables with high cardinality (ie, variables with a large number of unique categories or groups). Importantly, the feature importance analysis was carried out in a subset of the data to reduce overfitting, but it should be acknowledged that the data in the validation data set may exhibit a different distribution compared with the derivation data set. Consequently, this discrepancy could potentially result in different features being identified as important when conducting the feature importance analysis. To address this, we compared the distribution of predictors across the various data sets and the statistically significant differences in proportions or medians observed were not deemed clinically relevant. Additionally, given the variations in how algorithms encode relationships between predictors and outcomes, we anticipate that the feature importance rankings in Figure 3 would differ had we conducted the analysis on our final ensemble model using Shapley methods, rather than using the XGBoost model. Finally, although we used a derivation and a validation data set, external validation is required. Unfortunately, genetically characterized records are not commonly available in current EHRs on a scale that enables swift validation. Nevertheless, before considering widespread deployment, it is crucial to prioritize external validation and substantiation of our model's clinical utility.

Future work should concentrate on implementation and validation within live EHR systems of our ML screening model. Considering the diverse landscape of EHR systems characterized by complex data standards ranging from proprietary to open standards like Fast Healthcare Interoperability Resources, alongside varied coding systems such as READ and Systematized Nomenclature of Medicine Clinical Terms codes, it becomes imperative to devise system-specific "wrapper" mapping functions. These functions will seamlessly transform EHR data into the requisite inputs for our model, enhancing its integration across a spectrum of EHR systems. In summary, the novel model developed through state-of-the-art ML methodologies has the potential to more accurately identify individuals carrying a pathogenic variant associated with FH than the current recommended screening tool. The ML-derived model has potential applications to EHRs in providing

large-scale automated screening, thus helping prioritizes individuals for genetic testing, reducing the number of required tests to identify more true cases of FH with potential efficiencies and savings for resource-stretched health systems.

ARTICLE INFORMATION

Received March 7, 2024; accepted May 15, 2024.

Affiliations

Department of Primary Care and Public Health, School of Public Health, Imperial College London, London, United Kingdom (C.A.S., A.J.V., F.B., J.B., M.T.S., K.K.R.); Department of Medicine, Faculty of Medicine, Universidad de Sevilla, Sevilla, Spain (A.J.V.); Clinical Epidemiology and Vascular Risk, Instituto de Biomedicina de Sevilla (IBiS), IBiS/Hospital Universitario Virgen del Rocío/Universidad de Sevilla/CSIC, Sevilla, Spain (A.J.V.); Centro de Investigación Biomédica en Red (CIBER) de Epidemiología y Salud Pública, Instituto de Salud Carlos III, Madrid, Spain (A.J.V.); Nacional Institute of Health Dr. Ricardo Jorge, Lisbon, Portugal (J.R.C.); BioISI—Biosystems and Integrative Sciences Institute, University of Lisbon, Portugal (J.R.C.); Department of Internal Medicine, Faculty of Medicine, School of Health Sciences, University of Ioannina, Greece (F.B.); Department of Medicine I, University Hospital Aachen, Aachen, Germany (J.B.); Quantitative Research, Davidson Kempner Capital Management, New York, NY (A.M.); and National Institute of Cancer, National Institute of Health, Rockville, MD (L.A.).

Acknowledgments

This research was conducted using the UK Biobank Resource under application number 67789. C.A.T. Stevens: conceptualization, methodology, software, formal analysis, investigation, data curation, writing–original draft, visualization, project administration, and funding acquisition; A.J. Vallejo-Vaz: conceptualization, methodology, investigation, data curation, writing–review and editing, and supervision; M.T.A. Sharabiani: conceptualization, methodology, investigation, writing–review and editing, and supervision; K.K. Ray: conceptualization, methodology, investigation, writing–original draft, writing–review and editing, supervision, and funding acquisition; F. Barkas: data curation, investigation, and writing–review and editing; Dr Brandts: writing–review and editing; J.R. Chora, A. Mahani, and L. Abar: methodology, investigation, writing–review: methodology, investigation, and writing–review and editing.

Sources of Funding

Data access was funded by the Familial Hypercholesterolemia Studies Collaboration. K.K. Ray acknowledges support from the National Institute for Health Research Imperial Biomedical Research Centre, UK. A.J. Vallejo-Vaz acknowledges support from the Programme "Beatriz Galindo" from the Ministry of Universities, Spain, and University of Seville, Spain. J.R. Chora acknowledges her PhD fellowship funded by FCT-Fundação para a Ciência e Tecnologia do Ministério da Ciência, Tecnologia e Ensino Superior (SFRH/BD/108503/2015).

Disclosures

C.A.T. Stevens reports grants from Pfizer, Amgen, Merck Sharp & Dohme, Novartis, Sanofi-Aventis, Daiichi Sankyo, and Regeneron, during the conduct of the study. K.K. Ray reports grants and personal fees from Amgen, Sanofi–Regeneron, Pfizer, Merck Sharp & Dohme, and Daiichi Sankyo and Ultragenyx; and personal fees from AstraZeneca, Kowa, Novartis, Sanofi, Amgen, Eli Lilly, Algorithm, Boehringer Ingelheim, Novo Nordisk, Silence Therapeutics, Bayer, Esperion, Daiichi Sankyo, Abbott, New Amsterdam, SCRIBE, CRISPR, VAXXINITY, EMENDOBIO, Cargene, Viatris, Amarin, Nodthera and Resverlogix, outside the submitted work. A.J. Vallejo-Vaz reports past or current participation in research grants to Imperial College London from Pfizer, Amgen, Merck Sharp & Dohme, Sanofi-Aventis, Daiichi Sankyo, and Regeneron, outside the submitted work; and personal fees for consulting from Bayer and Regeneron and honoraria for lectures from Amgen, Mylan, Akcea and Ferrer, all outside the submitted work. F. Barkas reports personal fees from Novartis, Amgen, Novo, Nordisk, Eli Lilly, and honoraria from Novartis, Viatris and Sanofi. Dr Brandts received speaker honoraria from Amgen and research grant support from AstraZeneca, outside the submitted work. The remaining authors have no disclosures to report.

Supplemental Material

Data S1
Tables S1–S6

REFERENCES

1. Nordestgaard BG, Chapman MJ, Humphries SE, Ginsberg HN, Masana L, Descamps OS, Wiklund O, Hegele RA, Raal FJ, Defesche JC, et al. Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: guidance for clinicians to prevent coronary heart disease: consensus statement of the European Atherosclerosis Society. *Eur Heart J*. 2013;34:3478–90a. doi: 10.1093/eurheartj/eht273
2. Hu P, Dharmayat KI, Stevens CAT, Sharabiani MTA, Jones RS, Watts GF, Genest J, Ray KK, Vallejo-Vaz AJ. Prevalence of familial hypercholesterolemia among the general population and patients with atherosclerotic cardiovascular disease: a systematic review and meta-analysis. *Circulation*. 2020;141:1742–1759. doi: 10.1161/CIRCULATIONAHA.119.044795
3. Vallejo-Vaz AJ, Ray KK. Epidemiology of familial hypercholesterolaemia: community and clinical. *Atherosclerosis*. 2018;277:289–297. doi: 10.1016/j.atherosclerosis.2018.06.855
4. Luirink IK, Wiegman A, Kusters DM, Hof MH, Groothoff JW, de Groot E, Kastelein JJP, Hutten BA. 20-year follow-up of statins in children with familial hypercholesterolemia. *N Engl J Med*. 2019;381:1547–1556. doi: 10.1056/NEJMoa1816454
5. W. H. O. Human Genetics Programme. World Health Organization.
6. Williams RR, Hunt SC, Schumacher MC, Hegele RA, Leppert MF, Ludwig EH, Hopkins PN. Diagnosing heterozygous familial hypercholesterolemia using new practical criteria validated by molecular genetics. *Am J Cardiol*. 1993;72:171–176. doi: 10.1016/0002-9149(93)90155-6
7. Simon Broome Register Group. Risk of fatal coronary heart disease in familial hypercholesterolaemia. Scientific steering committee on behalf of the Simon Broome register group. *BMJ*. 1991;303:893–896. doi: 10.1136/bmj.303.6807.893
8. Marco-Benedí V, Cenarro A, Vila À, Real JT, Tamarit JJ, Walther LAA-S, Diaz-Diaz JL, Perea V, Civeira F, Vaz AJV. Impact of conducting a genetic study on the management of familial hypercholesterolemia. *J Clin Lipidol*. 2023;17:717–731. doi: 10.1016/j.jacl.2023.08.008
9. Khera AV, Won HH, Peloso GM, Lawson KS, Bartz TM, Deng X, van Leeuwen EM, Natarajan P, Emdin CA, Bick AG, et al. Diagnostic yield and clinical utility of sequencing familial hypercholesterolemia genes in patients with severe hypercholesterolemia. *J Am Coll Cardiol*. 2016;67:2578–2589. doi: 10.1016/j.jacc.2016.03.520
10. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, et al. The UK biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562:203–209. doi: 10.1038/s41586-018-0579-z
11. Van Hout CV, Tachmazidou I, Backman JD, Hoffman JX, Ye B, Pandey AK, Gonzaga-Jauregui C, Khalid S, Liu D, Banerjee N, et al. Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK biobank. *bioRxiv*. 2019:572347.
12. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17:405–424. doi: 10.1038/gim.2015.30
13. Chora JR, Iacocca MA, Tichý L, Wand H, Kurtz CL, Zimmermann H, Leon A, Williams M, Humphries SE, Hooper AJ, et al. The clinical genome resource (ClinGen) familial hypercholesterolemia variant curation expert panel consensus guidelines for LDLR variant classification. *Genet Med*. 2022;24:293–306. doi: 10.1016/j.gim.2021.09.012
14. Cao YX, Sun D, Liu HH, Jin JL, Li S, Guo YL, Wu NQ, Zhu CG, Gao Y, Dong QT, et al. A novel modified system of simplified Chinese criteria for familial hypercholesterolemia (SCCFH). *Mol Diagn Ther*. 2019;23:547–553. doi: 10.1007/s40291-019-00405-1
15. Akyea RK, Qureshi N, Kai J, de Lusignan S, Sherlock J, McGee C, Weng S. Evaluating a clinical tool (FAMCAT) for identifying familial hypercholesterolaemia in primary care: a retrospective cohort study. *BJGP Open*. 2020;4:1–10. doi: 10.3399/bjgpopen20X101114
16. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis

(TRIPOD): the TRIPOD statement. The TRIPOD Group. *Circulation*. 2015;131:211–219. doi: 10.1161/CIRCULATIONAHA.114.014508
17. Vallejo-Vaz AJ, Stevens CAT, Lyons ARM, Dharmayat KI, Freiberger T, Hovingh GK, Mata P, Raal FJ, Santos RD, Soran H, et al. Global perspective of familial hypercholesterolaemia: a cross-sectional study from the EAS familial hypercholesterolaemia studies collaboration (FHSC). *Lancet*. 2021;398:1713–1725. doi: 10.1016/S0140-6736(21)01122-3
18. Vallejo-Vaz AJ, De Marco M, Stevens CAT, Akram A, Freiberger T, Hovingh GK, Kastelein JJP, Mata P, Raal FJ, Santos RD, et al. Overview of the current status of familial hypercholesterolaemia care in over 60 countries—the EAS familial hypercholesterolaemia studies collaboration (FHSC). *Atherosclerosis*. 2018;277:234–255. doi: 10.1016/j.atherosclerosis.2018.08.051
19. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Springer New York Inc.; 2001.
20. Banda JM, Sarraju A, Abbasi F, Parizo J, Pariani M, Ison H, Briskin E, Wand H, Dubois S, Jung K, et al. Finding missed cases of familial hypercholesterolemia in health systems using machine learning. *NPJ Digit Med*. 2019;2:23. doi: 10.1038/s41746-019-0101-5
21. Myers KD, Knowles JW, Staszak D, Shapiro MD, Howard W, Yadava M, Zuzick D, Williamson L, Shah NH, Banda JM, et al. Precision screening for familial hypercholesterolaemia: a machine learning study applied to electronic health encounter data. *Lancet Digit Health*. 2019;1:e393–e402. doi: 10.1016/S2589-7500(19)30150-5
22. Gratton J, Futema M, Humphries Steve E, Hingorani Aroon D, Finan C, Schmidt AF. A machine learning model to aid detection of familial hypercholesterolemia. *JACC: Adv*. 2023;2:100333.
23. Akyea RK, Qureshi N, Kai J, Weng SF. Performance and clinical utility of supervised machine-learning approaches in detecting familial hypercholesterolaemia in primary care. *NPJ Digit Med*. 2020;3:142. doi: 10.1038/s41746-020-00349-5
24. Groselj U, Wiegman A, Gidding SS. Screening in children for familial hypercholesterolaemia: start now. *Eur Heart J*. 2022;43:3209–3212. doi: 10.1093/eurheartj/ehac224
25. Gidding SS, Wiegman A, Groselj U, Freiberger T, Peretti N, Dharmayat KI, Daccord M, Bedlington N, Sikonja J, Ray KK, et al. Paediatric familial hypercholesterolaemia screening in Europe: public policy background and recommendations. *Eur J Prev Cardiol*. 2022;29:2301–2311. doi: 10.1093/eurjpc/zwac200
26. Toft-Nielsen F, Emanuelsson F, Benn M. Familial hypercholesterolemia prevalence among ethnicities-systematic review and meta-analysis. *Front Genet*. 2022;13:840797. doi: 10.3389/fgene.2022.840797