

https://dspace.org/

Familial hypercholesterolemia identification by machine learning using lipid profile data performs as well as clinical diagnostic criteria

Hesse, Reinhardt

2022

https://hdl.handle.net/10539/34436

Downloaded from a DSpace repository. https://dspace.org/

MMed Research Report

Degree: MMed (Chemical Pathology)
Candidate: **Reinhardt Hesse**
Student number: **2041612**

Date: 10 December 2021

American Heart Association.
Circulation: Genomic and Precision Medicine
Manuscript Submission and Peer Review System

Title page

Full title: Familial hypercholesterolemia identification by machine learning using lipid profile data performs as well as clinical diagnostic criteria.

First author surname: Hesse

Short title: Machine learning for familial hypercholesterolemia

Authors details:

- Reinhardt Hesse[1]; MBChB;
  - ORCID iD: https://orcid.org/0000-0003-2440-7609
- Frederick J. Raal[2]; MBBCh, FRCP, FRCPC, FCP(SA), Cert Endo, MMed, PhD;
  - ORCID iD: https: //orcid.org/0000-0002-9170-7938
- Dirk Blom[3]; MBChB, MMed, FCP (SA), PhD;
  - ORCID iD: https://orcid.org/0000-0003-3965-5912
- Jaya A. George[1]; MBBS, DTM&H, MSc, FCPath(SA)Clin; PhD
  - ORCID iD: https://orcid.org/0000-0002-8741-8746

1. Department of Chemical Pathology; National Health Laboratory Service and University of the Witwatersrand; Johannesburg; South Africa
2. Division of Endocrinology and Metabolism; Department of Internal Medicine; University of the Witwatersrand; Johannesburg; South Africa
3. Division of Lipidology and Hatter Institute for Cardiovascular Research in Southern Africa; Department of Medicine; University of Cape Town; Cape Town; South Africa

Corresponding author: Reinhardt Hesse; +27 82 338 4024; reinhardt.hesse@nhls.ac.za; Department of Chemical Pathology; University of the Witwatersrand Health Sciences Campus; 7 York Road; Park Town; 2193; Johannesburg; South Africa

Abstract

Background

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

Methods

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

Results

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

Conclusions

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

Introduction

Familial hypercholesterolemia (FH) is one of the most common monogenic disorders and, if untreated, increases the risk of premature atherosclerotic cardiovascular disease (ASCVD) due to lifelong exposure to increased levels of low-density lipoprotein cholesterol (LDL-C).[1-3] Early diagnosis and treatment have been proven to reduce both ASCVD events and CVD-related mortality.[2,3]

Worldwide, FH is estimated to affect approximately 1 in 250 individuals in the general population.[4-7] The prevalence of FH in the general population of South Africa is unclear, but due to founder effects in three distinct ethnic groups (White Afrikaners, Jewish South Africans of Lithuanian descent, and Indian South Africans of Gujarati descent) the prevalence in these groups has been reported to be as high as 1 in 70 individuals.[8,9]

Despite the benefits of early intervention, the majority of FH patients remain undiagnosed and untreated. Worldwide, more than 99% of individuals with FH are undiagnosed and it is estimated that only approximately 3% of individuals with FH in South Africa have been diagnosed.[10] Only six countries have identified more than 5% of their estimated FH population, but even in these countries the identification rate is also generally well below 50%, except for the Netherlands where an estimated 71% of individuals with FH have been identified by a national screening program.[10]

Current screening practices for FH, other than the investigation of individuals with established ASCVD and their family, rely on population-wide LDL-C concentration testing and further investigation of individuals with an LDL-C concentration above a predetermined cut-off.[10] However, in the general population, less than 3% of individuals flagged with an LDL-C above 4.9 mmol/L have genetically confirmed FH.[11,12]

Individuals with high LDL-C should be evaluated further by ruling out secondary causes of hyperlipidemia and applying a clinical scoring tool that combines data obtained by a clinician (consisting of patient history, physical examination, and family history) with the results of biochemical and genetic investigations. Notably, amongst laboratory analytes, only the LDL-C and/or total cholesterol (TC) concentration form part of the scoring criteria, and not the other components of a standard lipid profile, namely, high-density lipoprotein cholesterol (HDL-C) or triglyceride (TG) concentration.[11] At least three validated clinical scoring tools exist, with the Dutch Lipid Clinic Network (DLCN) scoring criteria being the most widely accepted and commonly used.[1,10,11] However, genetic evidence of a FH-causing

mutation is considered to be the gold standard and can be used in isolation or combination with clinical scoring tools.[1,11]

Even in high prevalence cohorts, such as lipid clinics, a FH causing gene variant is found in only 60-80% of individuals diagnosed as 'definite FH' by one of the various clinical scoring criteria.[12-14] Because the reliability of genetic screening depends on the genetic tools used (i.e., some techniques may miss large deletions or insertions), some older cohort studies may have slightly underestimated the proportion of individuals with FH-causing variants.[12] Subsequently, genome wide association studies in patients with clinical FH but without identifiable monogenic mutations have found an accumulation of multiple 'small effect size allelic variants' responsible for the increase in LDL-C.[12,15] These patients are sometimes labelled as having 'polygenic FH', although the term 'polygenic hypercholesterolemia' is a better description as 'familial hypercholesterolemia' is generally accepted to refer to the monogenic entity.[12] This is an important distinction as monogenic FH confers significantly higher ASCVD risk and a different inheritance pattern, with cascade screening implications, versus polygenic hypercholesterolemia.[12,16-21] Despite the diagnostic superiority of genetic testing, due to cost and availability it is currently not being used for widespread population level screening and diagnosis but has been shown to be of value in cascade screening once an index case has been diagnosed.[10,21,22]

Other than the suboptimal diagnostic accuracy of the current screening and diagnostic approach to FH, other barriers such as the high reliance on human decision making also exist. Human decision making is prone to bias and misinterpretation and depends on experienced health care workers.[23] These barriers are evident across all countries but are more relevant in resource-constrained countries such as South Africa.

Technological advances, such as the advent of artificial intelligence (AI), provide an opportunity for solving old problems with new tools. Machine learning (ML), a subdivision of AI that utilizes computer powered statistical analysis of large datasets to find patterns, is one such tool. Supervised ML, a subdivision of ML, allows for multiple numerical and categorical variables to be associated with a specific outcome.[24] When many such instances are presented, a pattern is recognized, and an algorithm can be constructed which can be used to predict future outcomes when faced with new data. It is important that the outcome used to train and evaluate the ML model on be as objective as possible (i.e., diagnosis as determined by the gold standard).[25]

Various forms of supervised ML exist employing different statistical tools such as logistic regression, linear regression, deep neural network, and random decision forests.[24,26]

Although relatively new to the medical world, ML has already demonstrated its usefulness in various prognostic and predictive applications across many medical disciplines.[27,28] The field of laboratory medicine is especially well suited to integration with ML-based tools due to the great quantity of numeric and categorical data already available on a typical laboratory information system (LIS) and, as a result, various promising tools have already been developed in this setting.[29]

The aim of our study was to create a ML model, from data available on a basic LIS, capable of achieving improved screening accuracy when compared to the current screening tool of LDL-C cut-off and non-inferior diagnostic performance when compared to the DLCN criteria as measured by the gold standard of genetic FH mutation confirmation on difference prevalence cohorts.

Methods

Study design

We used electronic databases from the lipid clinics of two large academic hospitals to build, train and assess the performance of the ML model. Data from the Charlotte Maxeke Johannesburg Academic Hospital's (CMJAH) lipid clinic, Johannesburg, South Africa, was used to develop and conduct internal evaluation of the ML model. Data from the Groote Schuur Hospital (GSH) lipid clinic, Cape Town, South Africa, was used to perform external evaluation of the ML model developed using the CMJAH data. Both datasets consisted of clinically diagnosed FH individuals, and as a result were considered to be a 'high prevalence' cohort in terms of FH genetic mutation presence. The DLCN FH diagnostic criteria and LDL-C cut-off tool for FH screening were applied to both these internal and external datasets to serve as comparison to the ML model's performance.

Additionally, another dataset from the GSH lipid clinic was used which consisted of individuals referred to the clinic for evaluation of severe hypercholesterolemia. We expected that the FH mutation presence in this group would be lower than the diagnosed FH cohort, but still higher than the general population - thus a 'medium prevalence' FH mutation presence cohort. This database was also used to manually construct a third external dataset which served as a 'low prevalence' FH mutations presence cohort intended to mimic the general population prevalence of FH mutation. This was done because

disease prevalence and differences in cohort characteristics may have a marked effect on the performance of classification tools.[30,31]

Descriptive statistics

The descriptive statistics of each dataset were calculated for comparison within and between datasets by calculating the percentage presence of categorical data (FH genetic status; sex; ethnicity; lipid lowering therapy use and previous cardiovascular events) as well as the mean and standard deviation of numerical data (age; TC; HDL-C; LDL-C; and TG concentration).

Training dataset

Supplementary figure 1 provides an overview of the data cleaning, model development, internal and external evaluation workflow.

A database of 678 patients attending CMJAH lipid clinic, clinically suspected of having FH, was obtained from the University of the Witwatersrand database of the FIND-FH project. This part of the international FIND-FH project aims to identify patients with FH in all population groups in South Africa.[9,32] This database contained demographic information (age, sex and race) as well as biochemical (including full lipid profile), clinical information (history, medication and clinical examination findings, including an assessment of skin and tendon xanthoma and corneal arcus) and genetic analysis by targeted next generation sequencing for the coding regions of the four genes definitively linked to FH thus far: low density lipoprotein receptor (LDLR), apolipoprotein B (APOB), proprotein convertase subtilisin/kexin type 9 (PCSK9) and low density lipoprotein receptor adaptor protein 1 (LDLRAP1).

Patients younger than 20 years, which is the recommended age of lipid profile screening in South Africa, those with genetic mutation of uncertain significance and patients without at least one LDL-C measurement were removed from the dataset. More information on the genetic techniques used in the evaluation of this cohort can be found in the genetic sequencing approach section of the data supplementary.

The dataset was then randomized and split into a 70% training dataset and a 30% testing set. The training dataset was further cleaned to remove all clinical and laboratory results apart from the lipid profile (TC; HDL-C; LDL-C; and TG concentration) and basic demographics (age and sex) and the FH-mutation status. This was done as the intention of the ML model is for utilization on a LIS database with availability of only the specified information (except the FH-mutation status). The sex and age were included not because it is expected that they will have an influence on the FH-mutation status, but rather because age

and sex itself has an influence on the components of the lipid profile.

Machine learning model design and development

The training dataset was used to create various candidate ML algorithms based on different supervised learning methods with BigML software (Winter 2019 release; BigML Inc.; Corvallis, Oregon, USA). The objective identifier used for training was the FH mutation status (either "positive" or "negative"). The variables used to train the algorithm were: age, sex, TC, HDL-C, LDL-C, and TG concentration. Where more than one set of biochemical results were available per individual, preference was given to results obtained before initiation of lipid-lowering therapy. To ensure applicability of the ML model to its intended use (to identify FH patients on LIS data), neither lipid lowering therapy details nor imputed LDL-C values were used for ML because real-world application of the ML model would not be able to benefit from such information as prescribing information is not routinely available on a LIS. LDL-C baseline imputation is a validated method to calculate the probable pre-treatment LDL-C concentration by applying a correction factor based on the specific type and dose of lipid-lowering medication.[33,34] Imputation for the two most common classes of lipid-lowering treatment, namely statins and ezetimibe, was done where applicable.

Multiple candidate models with variable configurations were created via Bayesian parameter optimization, with the area under the receiver operator characteristic (AUROC) curve selected as target performance metric. The AUROC curve was chosen as target metric because of the need for threshold adjustment and the ranked nature of the outcomes, due to the importance of correctly identifying both the positive and negative individuals. Cross-validation was performed with Monte Carlo cross-validation (5 -fold with 0.8 sample rate). The top four best performing models, as determined by achieving an AUROC curve of 0.700 or greater, consisted of one logistic regression, one deep learning, and two decision tree ensemble classifiers. The AUROC curves achieved during cross-validation of these four models were 0.767, 0.779, 0.722 and 0.700, respectively.

The logistic regression model was constructed with L1 regularization (c=7.78009637445), inclusion of the bias term, with auto-scaling, missing values prediction, and eps=0.0755009108922. The deep learning model was created by employing automatic network topology search and parameter optimization (128 networks evaluated), missing value prediction, and a maximum training time of 9000 seconds. On the first decision tree ensemble classifier, gradient boosted trees were employed consisting of 328 nodes, 4 boosting iterations,

deterministic order model construction, and proportional class balancing. The other decision tree ensemble classifier was constructed without boosting, using 619 nodes, a random candidate ratio of 0.71, in a 49-model, pruned deterministic order.

These four models were then selected to form a 'fusion' model, which combines the output of the four evenly weighted models to produce a single output prediction. The underlying models differed in their findings on the order of importance of the variables, but TG concentration was in the top two most important variables in all the models. More information regarding the programming inputs can be found in the supplementary .json files.

Model performance and comparison with existing FH classification tools

The combined FH model was then used to predict the FH-status of the 30% testing set, which was kept separate from the start. The same 30% testing dataset was used to assess the performance of the DLCN criteria as well as the LDL-C cut-off approach of identifying FH. In contrast to the data used for the ML training and testing, both the DLCN criteria and the LDL-C cut-off tools benefited from LDL-C concentration imputation when a baseline LDL-C concentration was not available, as this is the recommendation when using these tools.[33,34] Additionally, subgroup analysis was done to evaluate the performance of the ML model on "on-treatment" vs "treatment-naïve" individuals to assess if this had a major impact on the model's performance. The classification performance of all three tools (ML model; DLCN criteria; LDL-C cut-off) were evaluated by applying the following performance metrics for each tool: AUROC curve and sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), overall accuracy and F-score for pre-specified cut-offs. The F-score is the harmonic mean of the PPV and sensitivity and useful in assessing the over-all performance of binary classification tools at a specific cut-off.[36] The cut-off for the LDL-C tool was chosen as ≥4.5 mmol/L, as this was shown to be the recommended optimal cut-off point for screening purposes to detect a FH-mutation.[4] The cut-off for the DLCN criteria was chosen as ≥8 points, correlating with the category of 'definite FH'. The cut-off for the ML model was 60%, which was the percentage of probability that achieved the best AUROC during the internal evaluation. For the AUROC curves, the thresholds for the ML model were increments of 5% probability (0 - 100%), the thresholds for the DLCN criteria were the numeric point value awarded by the criteria (0 - 20+) and the thresholds for the LDL-C cut-off were increments of 0.5 mmol/L (0.5 - 10.5+ mmol/L).

External evaluation

The three external datasets derived from a single database from the GSH lipid clinic were then used to evaluate the performance of the ML model when presented with data from a different center and different levels of FH-prevalence. The first dataset was similar to the training dataset from CMJAH in that it consisted of a cohort of patients diagnosed as FH on clinical grounds, and all had limited genetic analysis for FH-mutation detection done (more information available in the genetic sequencing approach section of the data supplementary), as well as clinical and laboratory findings. The second dataset was from a cohort of patients who had FH as part of their differential diagnosis at first referral to the lipid clinic. As this dataset was from the same lipid clinic as the first external dataset, all individuals with clinically confirmed FH would also have been represented in this dataset, along with individuals with other causes of dyslipidaemia. This dataset had demographic and laboratory results but lacked detailed clinical information (i.e., history and clinical findings), and genetic analysis had only been performed in patients with a high clinical suspicion of FH. The third GSH dataset was constructed to mimic the suspected prevalence of ~1% (1:100) in the South African community.[35] The intention of the third dataset was to assess the ML model's ability to detect low prevalence true FH individuals from individuals without FH. This was done by using a randomized subset (n=16) of known FH positive individuals from the second GSH dataset and an additional 1600 individuals from the greater GSH lipid clinic database that were not clinically suspected of having FH, but that had at least one other CVD-associated comorbidity (namely hypertension, diabetes mellitus or obesity). The randomization of the 16 mutation positive FH individuals was repeated three times and the ML model's performance was evaluated on each to ensure that chance sampling did not produce an inaccurate representation.

All three FH classification tools (ML model; DLCN criteria; LDL-C cut-off) used in the original CMJAH 30% testing dataset were again applied to the first database from GSH and their performances were evaluated by applying the same performance metrics for each tool as was done for the CMJAH dataset.

The DLCN criteria could not be applied to the low and medium FH prevalence GSH datasets and no pre-treatment LDL-C imputation could be done due to lack of sufficient clinical information. The same performance metrics as described previously were used to assess the performance of the ML model and LDL-C cut-off on these two cohorts.

Statistical analysis

Microsoft Excel (Version 2002, Redmond, USA) was used to calculate

the standard deviation (SD), mean, and percentage for the descriptive statistics. The AUROC curve and the 95% confidence interval (CI) for the AUROC curve was done via the Hanley and McNeil method. The F-score, the ROC curve graphs, and all tables were calculated using Excel. The work-flow overview schematic was made with diagrams.net (version 13.8.1). The various performance metrics as well as their respective 95% CI's were calculated with MedCalc (Version 15.1, Ostend, Belgium). The intervals for the predictive values are the standard logit confidence intervals as given by Mercaldo et al. 2007, and the intervals for the sensitivity, specificity and accuracy are 'exact' Clopper-Pearson CI's.

Ethics approval and data availability

The study was approved by the institutional review committee of the University of the Witwatersrand as well as the human ethics committees of the respective centers (University of the Witwatersrand and University of Cape Town Human Research Ethics Committees) and is compliant to the ethics standards of the Declaration of Helsinki (2013).

Due to the South African Protection of Personal Information Act the respective datasets can only be made available to researchers trained in human subject confidentiality protocols.[36] Requests can be done at https://www.witsethics.co.za/Wits/index.aspx and hrec-submissions@uct.ac.za for consideration by the relevant ethics committees.

Results

Descriptive statistics

As seen in table 1, the FH mutation prevalence was similar between the training and testing dataset of the internal database (57% vs 58%). Among individuals with a FH mutation, only 2.2% had homozygous FH. The FH mutation prevalence in the clinically diagnosed FH ("high prevalence") external testing database was slightly higher at 64%, but none of these individuals had homozygous FH as it was an exclusion criteria for that particular database. As could be expected, the medium prevalence group of the external database which consisted of individuals with various forms of hypercholesterolemia, had a much lower FH mutation prevalence of 20%. The low prevalence group was purposefully selected to have a prevalence of 1% as explained in the methods section.

In terms of ethnicity, the internal dataset consisted of 68% White individuals, 26% Indian individuals and only 3.2% Black African individuals. The ethnic information on the external datasets were too limited to calculate the prevalence of ethnic groups.

	Internal dataset				External dataset					
	Training Data		Testing Data		High Prevalence		Medium Prevalence		Low Prevalence	
	FH+	FH-	FH+	FH-	FH+	FH-	FH+	FH-	FH+	FH-
N	223 (57%)	166 (43%)	96 (58%)	70 (42%)	887 (64%)	489 (36%)	649 (20%)	2655 (80%)	16 (1%)	1600 (99%)
Age in years	46 (17)	50 (15)	47 (17)	47 (15)	43 (13)	52 (12)	43 (13)	53 (13)	42 (14)	54 (12)
Female sex	125 (56%)	82 (49%)	60 (63%)	37 (53%)	486 (55%)	291 (60%)	348 (54%)	-	-	-
Ethnicity										
-White	164 (73.5%)	101 (60.8%)	69 (71.9%)	43 (61.4%)	-	-	-	-	-	-
-Indian	47 (21.1%)	57 (34.3%)	19 (19.8%)	23 (32.9%)	-	-	-	-	-	-
-Black	9 (4.0%)	5 (3.0%)	2 (2.1%)	2 (2.9%)	-	-	-	-	-	-
-Asian	3 (1.3%)	3 (1.8%)	6 (6.3%)	2 (2.9%)	-	-	-	-	-	-
Lipid Rx										
-Statin	155 (70%)	90 (54%)	73 (76%)	33 (47%)	516 (58%)	141 (29%)	-	-	-	-
-Ezetimibe	42 (18.8%)	6 (3.6%)	17 (17.7%)	1 (1.4%)	-	-	-	-	-	-
-Omega-3	8 (3.6%)	11 (6.6%)	3 (3.1%)	3 (4.3%)	-	-	-	-	-	-
-Fibrate	5 (2.2%)	2 (1.2%)	2 (2.1%)	1 (1.4%)	-	-	-	-	-	-
CVD event	54 (24%)	10 (6%)	30 (31%)	9 (13%)	207 (23%)	87 (18%)	-	-	-	-
TC in mmol/L	6.8 (2.1)	6.0 (1.7)	6.8 (2.3)	6.1 (1.4)	8.6 (1.7)	8.4 (1.2)	8.9 (1.9)	7.7 (1.6)	8.9 (1.8)	7.5 1.3)
HDL-C in mmol/L	1.3 (0.4)	1.4 (0.5)	1.3 (0.3)	1.4 (0.5)	1.2 (0.4)	1.2 (0.4)	1.2 (0.4)	1.4 (0.5)	1.2 (0.2)	1.3 (0.4)
LDL-C in mmol/L	6.3 (3.6)	4.3 (2.0)	6.7 (4.0)	4.3 (1.6)	7.1 (1.9)	6.2 (1.2)	7.2 (1.9)	5.5 (1.6)	7.0 (1.7)	5.3 (1.2)
TG in mmol/L	1.6 (0.9)	2.2 (1.4)	1.8 (1.2)	2.1 (1.3)	1.5 (0.9)	2.0 (1.0)	1.5 (0.9)	1.9 (1.0)	1.5 (0.7)	1.9 (0.9)

Table 1: ***Descriptive statistics. Data expressed as n (%) or mean (SD). FH+ denotes genetically proven familial hypercholesterolemia patients and FH- denotes patients that either had no FH-associated mutation detected or did not have genetic analysis performed due to low clinical suspicion of FH. The dash denotes no analysis due to limited clinical information. N: number of individuals; Rx: Treatment; CVD: cardiovascular disease; TC: total cholesterol; HDL-C: high-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol. Where baseline LDL-C values were not available, imputation of the LDL-C was done based on prescription information, but TC values were not imputed.***

Generally, individuals with a confirmed FH mutation were slightly younger, but the average ages of all groups were between 42 and 54 years. All groups, except for the internal training set of individuals without mutations, had a slightly higher prevalence of females, ranging from 49 to 62%.

Both lipid lowering therapy use and CVD events were more prevalent in the FH individuals with confirmed mutations than in those without. The average TC and LDL-C were also higher in individuals with confirmed mutations, while the HDL-C and TG concentrations were generally lower in these groups. The internal dataset had higher statin coverage (ranging from 47 to 70%) than the external dataset (29-58%) and as the LDL-C concentrations were adjusted for statin use, but not the TC concentrations, the internal dataset's TC was lower than that of the external dataset (averages range 6.1 - 6.8 mmol/L vs 8.4 - 8.6 mmol/L). On average, the external dataset also had more severe lipid profiles and a higher FH mutation prevalence. The decrease in FH prevalence in the medium and low prevalence datasets were accompanied by an increase in heterogeneity of lipid profile results between the FH positive and FH negative groups.

Performance on internal data

As is evident in figure 1 (A), the DLCN criteria and the ML model had similar success in their ability to detect genetically proven FH as per the AUROC curves (0.755 and 0.754 respectively). The LDL-C cut-off performed worst, with an AUROC curve of 0.682. However, the 95% CI of all three tools were wide, resulting in an overlap.

The subgroup analysis to evaluate the ML model's performance on treatment-naïve versus individuals on lipid lowering therapy did not reveal a major difference (AUROC curves 0.821 vs 0.813).

When assessing the specific cut-offs of the various tools, the ML model numerically achieved the best overall accuracy, specificity, PPV, NPV and F-score. However, an overlap in the 95% CI was again evident in all the metrics, as seen in table 2.

	Internal dataset		
Tool	**LDL-C cut-off**	**DLCN definite FH**	**ML model**
Cut-off	4.5 mmol/L	8 points	60% probability
Accuracy	0.651 (0.573-0.723)	0.687 (0.610-0.756)	0.693 (0.617-0.762)
Sensitivity	0.667 (0.563-0.760)	0.583 (0.478-0.682)	0.594 (0.489-0.691)
Specificity	0.629 (0.505-0.741)	0.829 (0.716-0.905)	0.829 (0.716-0.905)
PPV	0.711 (0.638-0.775)	0.824 (0.708-0.902)	0.826 (0.712-0.903)
NPV	0.579 (0.496-0.658)	0.592 (0.488-0.689)	0.598 (0.493-0.695)
F-score	0.688	0.683	0.691

***Table 2:** Performance metrics of the different FH tools at their respective cut-off points on the internal dataset. Values expressed as mean and 95% confidence interval in brackets. PPV: positive predictive value; NPV: negative predictive value; LDL-C: Low-density lipoprotein cholesterol; DLCN: Dutch Lipid Clinic Network criteria; FH: familial hypercholesterolemia; ML: machine learning*

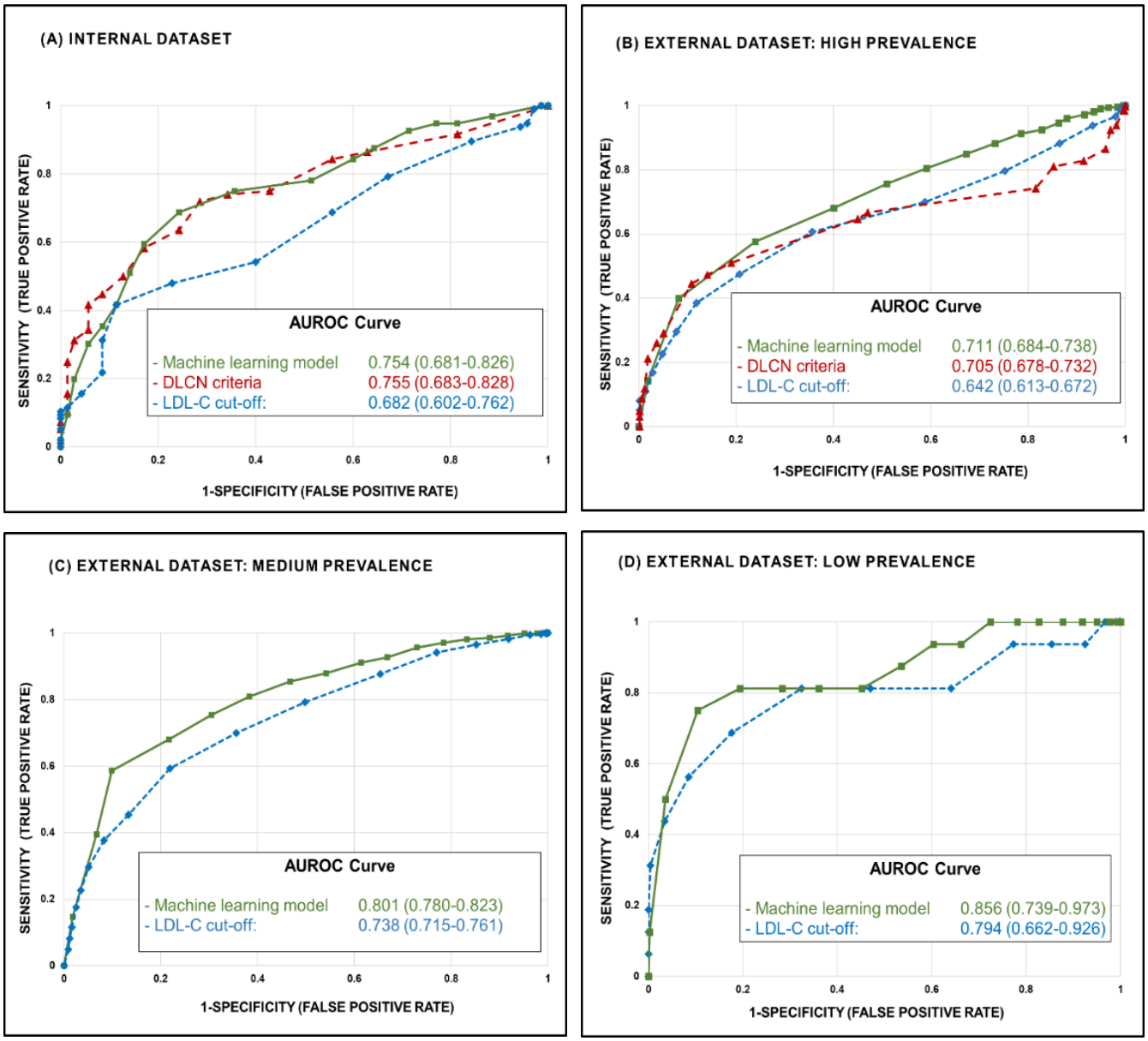

Figure 1: ***Receiver operator characteristic (ROC) curve analysis for familial hypercholesterolemia identification tools on the different datasets. AUROC: Area under the ROC; DLCN: Dutch Lipid Clinic Network criteria; LDL-C: Low-density lipoprotein***

Despite the low number of Black African individuals in both the training and testing dataset, all four of the individuals in the internal testing dataset were correctly classified by the ML model

Performance on external data

On the high prevalence external dataset, the ML model achieved the best AUROC curve (0.711), but the 95% CI overlapped with that of the DLCN criteria (AUROC=0.705) as seen in figure 1 (B). Both the ML model and the DLCN criteria performed better than the LDL-C cut-off (AUROC=0.642). The ML model also demonstrated a better accuracy, specificity, NPV and F-score versus both the DLCN criteria and LDL-C tools when the respective cut-offs were compared on the high prevalence dataset as seen in table 3. Importantly, this was achieved without sacrificing sensitivity and PPV as demonstrated by the non-inferior performance of these two metrics when the ML model was compared to the other two tools. On both the medium and low prevalence datasets, the AUROC curve of the ML model was better than the LDL-C cut-off (fig 1 C and D). The AUROC curves on these two datasets were better than what was achieved by the ML model on the high prevalence dataset, indicating that the model performs better when there is greater difference between the lipid profile results of individuals with and without FH. This is also reflected by the various performance metrics (i.e., accuracy, specificity, PPV, and F-score) in table 3 where the ML model performed better than the LDL-C cut-off. The LDL-C cut-off, which benefitted from LDL- C imputation based on lipid-lowering treatment information, did however display slightly better sensitivity (0.941 vs 0.878) on the medium prevalence

	External dataset: HP			External dataset: MP		External dataset: LP	
Tool	**LDL-C cut-off**	**DLCN definite FH**	**ML model**	**LDL-C cut-off**	**ML model**	**LDL-C cut-off**	**ML model**
Cut-off	4.5 mmol/L	8 points	60% probability	4.5 mmol/L	60% probability	4.5 mmol/L	60% probability
Accuracy	0.616 (0.590-0.642)	0.566 (0.540-0.593)	0.664 (0.638-0.689)	0.370 (0.353-0.386)	0.541 (0.524-0.558)	0.233 (0.212-0.254)	0.468 (0.444-0.493)
Sensitivity	0.882 (0.858-0.902)	0.788 (0.759-0.814)	0.882 (0.858-0.902)	0.941 (0.920-0.958)	0.878 (0.851-0.902)	0.813 (0.544-0.960)	0.875 (0.617-0.985)
Specificity	0.135 (0.106-0.169)	0.164 (0.133-0.200)	0.268 (0.230-0.310)	0.230 (0.214-0.246)	0.458 (0.439-0.478)	0.227 (0.210-0.248)	0.464 (0.440-0.489)
PPV	0.649 (0.639-0.659)	0.631 (0.602-0.660)	0.686 (0.658-0.713)	0.230 (0.225-0.235)	0.384 (0.275-0.293)	0.010 (0.008-0.013)	0.016 (0.013-0.019)
NPV	0.386 (0.321-0.456)	0.300 (0.245-0.358)	0.555 (0.489-0.619)	0.941 (0.921-0.957)	0.939 (0.926-0.950)	0.992 (0.978-0.997)	0.997 (0.990-0.999)
F-score	0.748	0.701	0.772	0.370	0.429	0.020	0.032

***Table 3:** Performance metrics of the different FH tools at their respective cut-off points on the external datasets. Values expressed as mean and 95% confidence interval in brackets. HP: high prevalence; MP: medium prevalence; LP: low prevalence; PPV: positive predictive value; NPV: negative predictive value; LDL-C: Low-density lipoprotein cholesterol; DLCN: Dutch Lipid Clinic Network criteria; FH: familial hypercholesterolemia; ML: machine learning*

dataset, but this was at the cost of specificity and PPV where it performed significantly worse than the ML model.

Discussion

Our ML model outperformed the LDL-C cut-off tool for FH screening and was as good as the DLCN criteria when the AUROC curves for the external dataset were assessed. All three tools were compared to the gold standard FH diagnosis identification of a pathogenic mutation. The ML model used only limited information from an LIS (age, sex and lipid profile results), while the LDL-C cut-off (LDL-C imputation) and the DLCN criteria (LDL-C imputation, family history, physical examination) benefited from the addition of clinical information.

At the selected probability cut-off, the ML model also had the best accuracy and F-score in each of the different datasets. The F-score, as a metric of the harmonized mean between the PPV and the sensitivity, is especially important in classification models as it provides a measure of the tool's ability to correctly identify individuals with FH (sensitivity) without including too many false positives (PPV).[37]

The application of ML for the identification of FH has been previously demonstrated on electronic health record (EHR) data.[22,38,39]. Despite only using basic lipid profile data, the AUROC curve (0.86) that our model achieved on the low prevalence dataset was equivalent to what was achieved by Banda et. al (0.94) and Myers et. al. (0.89).[22,38,39] As part of the FH Foundation's "Flag, Identify, Network, Deliver" ("FIND") FH initiative, Banda et. al. used a combination of clinical notes, biochemical results, pharmacy records and diagnostic codes to identify potential phenotypic FH individuals.[38,40] Their algorithm's performance was evaluated by testing its ability to correctly flag known FH patients on an external dataset, where it achieved a PPV of 0.85. The algorithm was also used to identify 56 patients with a 'high probability' of FH. Chart review of those 56 patients were done and it was established that 84% of the patients met the DLCN criteria of at least 'possible FH', but only 2% met the criteria for 'definite FH'.

In another FIND FH initiative study, Myers et. al. used EHR data consisting of procedural and diagnostic codes, prescription information and laboratory findings to create a ML algorithm for the purpose of precision screening for FH.[39] The algorithm was applied to an internal dataset where it achieved a PPV of 0.85 and a sensitivity of 0.45. The algorithm was then applied to two external databases where, respectively, 78% and 73% of identified individuals met the DLCN criteria for at least 'possible FH'. However, only 13% and 10% of the

screened individuals met the DLCN criteria for 'definite FH' on the two respective external databases.

Of note, these investigators used broader criteria for FH diagnosis (any one of the MEDPED, DLCN 'possible FH' or the opinion of an attending physician) while we diagnosed FH by the gold-standard of identification of a pathogenic mutation.

The distinction between individuals with high LDL-C due to genetically proven FH versus lifestyle or polygenic factors is important as FH-mutation have been associated with a 22-fold higher risk of CVD.[16] In our study it was also evident that individuals with a mutation had significantly higher rates of CVD events, despite being younger than the individuals without mutations and benefitting from a greater coverage of lipid lowering treatment. The lipid profile of the individuals with a mutation also differed from those without, as demonstrated by the higher average TC and LDL-C and lower TG and HDL-C, further supporting the impression that patients with an FH mutation are distinct from those without. This finding also supports similar findings in previous research.[13,14]

The previous ML studies mentioned not only employed different approaches to FH confirmation, but also used a more comprehensive EHR which provided rich clinical information. Recently, a study by Pina et. al. described three different ML algorithms that were based on genetic FH status and only used lipid profile data and age.[41] Their two datasets had high prevalence of FH positive mutations (45% and 84%), similar to our high prevalence internal and external datasets. When assessed on the external dataset, all three of their ML algorithms outperformed the DLCN criteria when the respective AUROC curves were compared (0.70, 0.78 and 0.76 for the three ML algorithms vs 0.64 for the DLCN criteria). This was very comparable to the performance of our ML model which achieved an AUROC curve of 0.71.

Despite using a smaller dataset (n=74) for testing than the internal testing dataset used in our study (n=166), Pina et. al. showed very narrow and non-overlapping 95% CI's when the DLCN criteria was compared to their three ML models. They however used a bootstrap method with 10 000-time resampling to calculate the 95% CI where we chose to use the traditional standard error approach outlined by Hanley and McNeil.[42] The Hanley and McNeil method is known for mildly overestimating the standard error, whereas bootstrapping may under-estimate the tails of the CI, especially in a small sample.[43,44]

In both our study and the study from Pina et. al. TG and HDL-C were important variables in addition to TC and LDL-C in identifying patients with FH. Previously we developed another ML model on the

same internal dataset that incorporated biochemical and clinical information and achieved an AUROC curve of 0.87.[45,46] In this model the variables of tendon xanthomata presence and lipoprotein(a) concentration, in addition to the discussed lipid profile analytes, were also shown to be of high value in identifying mutation positive FH, whereas the addition of apolipoprotein B did not make a significant difference. Lipoprotein(a) was not included in the ML model of the current study because it is not routinely requested in our setting. Nonetheless, it is evident that in addition to its intended purpose of detecting individuals with FH, the application of ML can also be used to identify useful variables outside of those used in traditional scoring systems.

Despite the obvious value that extensive EHR's and the addition of extra clinical and biochemical data has on the identification of FH individuals, this approach would not be feasible in settings such as ours due to limited resources. The current approach of cascade screening of an index patient's family has been demonstrated to be cost effective but may miss patients outside of the known ethnic clusters.[22] As mentioned, FH is very common in South Africa amongst certain White and Indian ethnicities, but the prevalence amongst Black South Africans has not been extensively studied.[8,9] The recent study by Raal et. al. highlighted this disparity, where only ~4% of individuals diagnosed with a FH mutation at a large South African academic hospital were of Black African ethnicity, despite contributing ~81% of the country's population.[47,48] This is probably not due to targeting the wrong mutations in these individuals, as the same study demonstrated that the 'clinical FH' group with no positive mutation consisted of only ~2% black individuals. Dyslipidaemia is however not an uncommon finding amongst Black South Africans, as a recent retrospective study done at another large South African academic hospital showed that 5.8% of Black patients with a lipid profile had a 'cholesterol-predominant' profile (TC >5 mmol/L, LDL-C >3 mmol/L and normal range TG), although specific investigations to diagnose FH were not done.[49] These two studies, taken together, highlight the possibility that FH is severely underdiagnosed in the Black South African population. The application of ML could prove very valuable in identifying previously undiagnosed individuals with FH, especially those outside of the traditional screening approaches.

The ability to easily adjust the probability thresholds is a feature of our ML model enabling it to be used in both high prevalence settings, such as a lipid clinic, and lower prevalence settings, such as individuals undergoing screening at their local clinic. We demonstrated the usefulness of this feature by selecting various probability thresholds for the

different cohorts and reported the performance metrics at each of the thresholds (Data Supplementary tables 2-5). This feature can be used to tailor the ML model to suite the intended purpose, depending on the required performance needed. As an example, applying the ML model at an 85% probability threshold for screening purposes to a low-prevalence cohort (such as the external low prevalence cohort used in this study; Data Supplementary table 5), significantly improves detection of FH affected individuals in comparison to an LDL-C cut-off approach (PPV of 0.067 vs 0.013) without increasing the risk of missing individuals with FH (sensitivities of 0.750 vs 0.813, with overlapping confidence intervals; and false omission rates of 0.3% vs 0.8%). Although the difference in detection might seem small, in a group of 1000 individuals this equates to 54 more true positive diagnosis's (67 versus 13). If every individual selected with the respective screening tools were offered further evaluation in the form a specialist clinical consultation (at a cost of ~$70 each) this would result in a cost of $1045 per true positive FH diagnosis in the ML model example versus $5385 in the LDL-C cut-off example. If every individual selected by the respective screening tools were offered limited FH genetic testing (at a cost of ~$170 each), instead of a clinical consultation, the cost difference in making a true positive diagnosis would be even greater at $2537 for the ML model approach versus $13077 for the LDL-C cut-off approach. The current practice of performing cascade screening on the family of FH diagnosed individuals as well as the down-stream savings in preventing major cardiovascular events would further compound this saving, not to mention the unmeasurable emotional and social impact that could be prevented.

The performance of the ML model on the lower prevalence FH cohorts are encouraging for possible screening use, however, the true performance on the low prevalence dataset needs to be evaluated in further research as the low-prevalence dataset in this study was not taken from a community cohort but rather selected from individuals attending the lipid clinic. Furthermore, both the relatively small size of our datasets and the low prevalence Black African individuals in our datasets is another limitation of our study, especially given the demographics of our setting, although the correct classification of the limited number of Black African individuals included in our study is reason for optimism. All of the datasets used to evaluate the ML model had high prevalence of hypercholesterolemia in both the mutation positive and negative groups. The effect of differing FH prevalence and lipid lowering treatment usage in various populations may result in different performance of the ML model. In this study the sensitivities and

specificities varied widely between the internal and external datasets across all three of the FH detection tools, highlighting the need for recalibration of the ML model before applying it to different cohorts. Lastly, as we did not apply the ML model on undiagnosed individuals on our LIS, this study can only serve as proof of concept. We hope to address these limitations in future research by applying the ML model described in this study to our LIS and fully investigating the individuals identified. Despite the promise that ML holds, we believe that this tool should only serve as decision support and not fully replace the current proven methods of screening and diagnosing FH, until more research, including prospective studies, are done to ensure its value and safety.

In summary, using only a basic lipid profile, age and sex, our ML model was better at correctly identifying genetically confirmed FH in a cohort of individuals clinically suspected of having FH than the LDL-C cut-off tool and comparable to the DLCN criteria. The same ML model performed even better when tested on two other cohorts with lower FH prevalence. The application of machine learning is therefore a promising tool in both the screening for, and diagnosis of, individuals with familial hyper-cholesterolemia.

List of abbreviations

AI:	artificial intelligence
APOB:	apolipoprotein B
ASCVD:	atherosclerotic cardiovascular disease
AUROC:	area under the receiver operator characteristic
CI:	confidence interval
CMJAH:	Charlotte Maxeke Johannesburg Academic Hospital
CVD	cardiovascular disease
DLCN:	Dutch Lipid Clinic Network
EHR:	electronic health record
FH:	familial hypercholesterolaemia
FIND FH	Flag, Identify, Network, Deliver Familial Hypercholesterolemia
GSH:	Groote Schuur Hospital
HDL-C:	high-density lipoprotein cholesterol
LDL-C:	low-density lipoprotein cholesterol
LDLR:	low density lipoprotein receptor
LDLRAP1:	low density lipoprotein receptor adaptor protein 1
LIS:	laboratory information system
ML:	machine learning
NPV:	negative predictive value
PCSK9:	proprotein convertase subtilisin/kexin type 9
PPV:	positive predictive value
SD:	standard deviation
TC:	total cholesterol
TG:	triglyceride

Acknowledgments

We would like to thank Prof. Evan Stein who made the FIND FH database for CMJAH available to us and Ms. Belinda Stevens for her help in acquiring the database.
We would also like to thank Prof. David Marais for his contribution to the GSH database which we used for external evaluation for our ML model.

Sources of Funding

Part of this study was funded by the Evan Stein Centre for Familial Hypercholesterolemia. The authors would also like to thank Medpace for database management and Medpace Reference Laboratories, Cincinnati, OH for funding and performing the biochemistry, lipid and apolipoprotein analysis, and next-generation sequencing for the FIND-FH cohort used in this study.

Disclosures

- R Hesse: None
- F.J. Raal has received research grants, honoraria, or consulting fees for professional input and/or delivered lectures from Sanofi, Amgen, Regeneron, Novartis and The Medicines Company.
- D. Blom has received research grants, honoraria, or consulting fees for professional input and/or delivered lectures from Sanofi, Amgen, Amryt, Regeneron, Abbott and MSD.
- J.A. George: None

Supplementary Materials

- Data Supplement:
  - Supplementary figure 1: Workflow in creating and evaluating the ML model
  - Supplementary table 1: Dutch Lipid Clinic Network (DLCN) Diagnostic Criteria for Familial Hypercholesterolemia
  - Supplementary table 2: Performance of the machine learning model at different probability thresholds on the internal dataset
  - Supplementary table 3: Performance of the machine learning model at different probability thresholds on the high prevalence external dataset
  - Supplementary table 4: Performance of the machine learning model at different probability thresholds on the medium prevalence external dataset
  - Supplementary table 5: Performance of the machine learning model at different probability thresholds on the low prevalence external dataset
- Machine learning workflow .json file
- Machine learning .json file actionable model reports x4
- Machine learning .json file fusion report

References

1. Alonso R, Perez de Isla L, Muñiz-Grijalvo O, Diaz-Diaz JL, Mata P. Familial Hypercholesterolaemia Diagnosis and Management. Eur Cardiol Rev [Internet]. 2018 [cited 2019 Sep 3];13(1):14. Available from: https://www.ecrjournal.com/articles/familial-hypercholesterolaemia-diagnosis-and-management

2. Luirink IK, Wiegman A, Kusters DM, Hof MH, Groothoff JW, De Groot E, et al. 20-Year follow-up of statins in children with familial hypercholesterolemia. N Engl J Med. 2019 Oct 17;381(16):1547–56.

3. Sijbrands EJG, Westendorp RGJ, Defesche JC, De Meier PHEM, Smelt AHM, Kastelein JJP. Mortality over two centuries in large pedigree with familial hypercholesterolaemia: Family tree mortality study. Br Med J [Internet]. 2001 Apr 28 [cited 2020 Oct 26];322(7293):1019–22

4. Benn M, Watts GF, Tybjærg-Hansen A, Nordestgaard BG. Mutations causative of familial hypercholesterolaemia: Screening of 98 098 individuals from the Copenhagen General Population Study estimated a prevalence of 1 in 217. Eur Heart J. 2016 May 1;37(17):1384–94.

5. Akioyamen LE, Genest J, Shan SD, Reel RL, Albaum JM, Chu A, et al. Estimating the prevalence of heterozygous familial hypercholesterolaemia: a systematic review and meta-analysis. BMJ Open [Internet]. 2017 Sep 1 [cited 2019 Sep 4];7(9):e016461. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28864697

6. Hu P, Dharmayat KI, Stevens CAT, Sharabiani MTA, Jones RS, Watts GF, et al. Prevalence of Familial Hypercholesterolemia Among the General Population and Patients With Atherosclerotic Cardiovascular Disease. Circulation [Internet]. 2020 Jun 2 [cited 2020 Oct 26];141(22):1742–59. Available from: https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.119.044795

7. Beheshti SO, Madsen CM, Varbo A, Nordestgaard BG. Worldwide Prevalence of Familial Hypercholesterolemia: Meta-Analyses of 11 Million Subjects. J Am Coll Cardiol [Internet]. 2020 May 26 [cited 2020 Oct 26];75(20):2553–66. Available from: https://pubmed.ncbi.nlm.nih.gov/32439005/

8. Blom D. Familial hypercholesterolaemia. JEMDSA. 2011;16(1):17–24.

9. Smyth N, Ramsay M, Raal FJ. Population specific genetic heterogeneity of familial hypercholesterolemia in South Africa. Vol. 29, Current Opinion in Lipidology. Lippincott Williams and Wilkins; 2018. p. 72–9.

10. Nordestgaard BG, Chapman MJ, Humphries SE, Ginsberg HN, Masana L, Descamps OS, et al. Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: Guidance for clinicians to prevent coronary heart disease. Eur Heart J. 2013 Dec 1;34(45).

11. McGowan MP, Hosseini Dehkordi SH, Moriarty PM, Duell PB. Diagnosis and Treatment of Heterozygous Familial Hypercholesterolemia. J Am Heart Assoc [Internet]. 2019 Dec 17 [cited 2020 Oct 9];8(24):e013225. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6951065/

12. Brandts J, Dharmayat KI, Ray KK, Vallejo-Vaz AJ. Familial hypercholesterolemia: is it time to separate monogenic from polygenic familial hypercholesterolemia? Curr Opin Lipidol [Internet]. 2020 Jun 1 [cited 2020 Oct 26];31(3):111–8. Available from: http://journals.lww.com/10.1097/MOL.0000000000000675

13. Civeira F, Ros E, Jarauta E, Plana N, Zambon D, Puzo J, et al. Comparison of genetic versus clinical diagnosis in familial hypercholesterolemia. Am J Cardiol [Internet]. 2008 Nov 1 [cited 2019 Oct 20];102(9):1187–93, 1193.e1. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18940289

14. Damgaard D, Larsen ML, Nissen PH, Jensen JM, Jensen HK, Soerensen VR, et al. The relationship of molecular genetic to clinical diagnosis of familial hypercholesterolemia in a Danish population. Atherosclerosis [Internet]. 2005 [cited 2020 Jul 1];180(1):155–60. Available from: https://pubmed.ncbi.nlm.nih.gov/15823288/

15. Sarraju A, Knowles JW. Genetic Testing and Risk Scores: Impact on Familial Hypercholesterolemia. Front Cardiovasc Med [Internet]. 2019 [cited 2019 Oct 20];6:5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30761309

16. Sharifi M, Futema M, Nair D, Humphries SE. Genetic Architecture of Familial Hypercholesterolaemia [Internet]. Vol. 19, Current Cardiology Reports. Current Medicine Group LLC 1; 2017 [cited 2020 Oct 26]. Available from: /pmc/articles/PMC5389990/?report=abstract

17. Séguro F, Rabès JP, Taraszkiewicz D, Ruidavets JB, Bongard V, Ferrières J. Genetic diagnosis of familial hypercholesterolemia is associated with a premature and high coronary heart disease risk. Clin Cardiol [Internet]. 2018 Mar 1 [cited 2020 Oct 9];41(3):385–91. Available from: /pmc/articles/PMC6489920/?report=abstract

18. Trinder M, Francis GA, Brunham LR. Association of Monogenic vs Polygenic Hypercholesterolemia with Risk of Atherosclerotic Cardiovascular Disease. JAMA Cardiol. 2020 Apr 1;5(4):390–9.

19. Migliara G, Baccolini V, Rosso A, D'Andrea E, Massimi A, Villari P, et al. Familial Hypercholesterolemia: A Systematic Review of Guidelines on Genetic Testing and Patient Management. Front public Heal [Internet]. 2017 [cited 2019 Sep 2];5:252. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28993804

20. Khera A V., Chaffin M, Zekavat SM, Collins RL, Roselli C, Natarajan P, et al. Whole-Genome Sequencing to Characterize Monogenic and Polygenic Contributions in Patients Hospitalized With Early-Onset Myocardial Infarction. Circulation [Internet]. 2019 Mar 26 [cited 2020 Oct 26];139(13):1593–602. Available from: https://pubmed.ncbi.nlm.nih.gov/30586733/

21. Henderson R, O'Kane M, McGilligan V, Watterson S. The genetics and screening of familial hypercholesterolaemia [Internet]. Vol. 23, Journal of Biomedical Science. BioMed Central Ltd.; 2016 [cited 2020 Jul 2]. Available from: /pmc/articles/PMC4833930/?report=abstract

22. Ibrahim S, Reeskamp LF, Stroes ESG, Watts GF. Advances, gaps and opportunities in the detection of familial hypercholesterolemia: overview of current and future screening and detection methods. Current Opinion in Lipidology. 2020 Dec;31(6):347-355. DOI: 10.1097/mol.0000000000000714

23. Alonso R, de Isla LP, Muñiz-Grijalvo O, Mata P. Barriers to early diagnosis and treatment of familial hypercholesterolemia: Current perspectives on improving patient care [Internet]. Vol. 16, Vascular Health and Risk Management. Dove Medical Press Ltd.; 2020 [cited 2020 Oct 9]. p. 11–25. Available from: /pmc/articles/PMC6957097/?report=abstract

24. Machine Learning —Fundamentals. Basic theory underlying the field of… | by Javaid Nabi | Towards Data Science [Internet]. [cited 2020 Oct 9]. Available from: https://towardsdatascience.com/machine-learning-basics-part-1-a36d38c7916

25. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ [Internet]. 2020 Mar 20 [cited 2020 Oct 9];368. Available from: http://dx.doi.org/10.1136/bmj.l6927

26. Classification and Regresion | BigML.com [Internet]. [cited 2020 Oct 9]. Available from: https://bigml.com/features/classification-regression

27. Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning and health systems. J Glob Health [Internet]. 2018 [cited 2020 Oct 9];8(2). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6199467/

28. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, Mahendiran T, Moraes G, Shamdas M, Kern C, et. al. (2019) A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health , 1 (6) e271-e297. 10.1016/S2589-7500(19)30123-2

29. Cabitza F, Banfi G. Machine learning in laboratory medicine: waiting for the flood? Clin Chem Lab Med [Internet]. 2018 Mar 28 [cited 2019 Mar 25];56(4):516–24. Available from: http://www.degruyter.com/view/j/cclm.2018.56.issue-4/cclm-2017-0287/cclm-2017-0287.xml

30. Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. Radiology [Internet]. 2018 Mar 1 [cited 2020 Oct 10];286(3):800–9. Available from: http://pubs.rsna.org/doi/10.1148/radiol.2017171920

31. Yu AC, Eng J. One Algorithm May Not Fit All: How Selection Bias Affects Machine Learning Performance. RadioGraphics [Internet]. 2020 Sep 25 [cited 2020 Oct 10];200040. Available from: https://pubs.rsna.org/doi/abs/10.1148/rg.2020200040

32. Raal FJ, Bahassi EM, Stevens B, Turner TA, Stein EA. Cascade Screening for Familial Hypercholesterolemia in South Africa. Arterioscler Thromb Vasc Biol [Internet]. 2020 Sep 3 [cited 2020 Nov 2];40:2747–55. Available from: https://www.ahajournals.org/doi/suppl/10.1161/ATVBAHA.120.315040.

33. Besseling J, Kindt I, Hof M, Kastelein JJP, Hutten BA, Hovingh GK. Severe heterozygous familial hypercholesterolemia and risk for cardiovascular disease: A study of a cohort of 14,000 mutation carriers. Atherosclerosis [Internet]. 2014 Mar [cited 2021 Mar 24];233(1):219–23. Available from: https://pubmed.ncbi.nlm.nih.gov/24529147/

34. Ruel I, Aljenedil S, Sadri I, de Varennes É, Hegele RA, Couture P, et al. Imputation of Baseline LDL Cholesterol Concentration in Patients with Familial Hypercholesterolemia on Statins or Ezetimibe. Clin Chem [Internet]. 2018 Feb 1 [cited 2020 Oct 11];64(2):355–62. Available from: https://academic.oup.com/clinchem/article/64/2/355/5608669

35. Klug, EQ. South African Dyslipidaemia Guideline Consensus Statement. South African Medical Journal, [S.l.], v. 102, n. 3, p. 178-187, feb. 2012. ISSN 2078-5135. Available at: http://www.samj.org.za/index.php/samj/article/view/5502/3929

36. Protection of Personal Information Act (POPI Act) - POPIA [Internet]. [cited 2021 Mar 25]. Available from: https://popia.co.za/

37. F-Score Definition | DeepAI [Internet]. [cited 2020 Oct 21]. Available from: https://deepai.org/machine-learning-glossary-and-terms/f-score

38. Banda JM, Sarraju A, Abbasi F, Parizo J, Pariani M, Ison H, et al. Finding missed cases of familial hypercholesterolemia in health systems using machine learning. npj Digit Med [Internet]. 2019 Dec 11 [cited 2020 Jun 2];2(1):23. Available from: http://www.nature.com/articles/s41746-019-0101-5

39. Myers KD, Knowles JW, Staszak D, Shapiro MD, Howard W, Yadava M, et al. Precision screening for familial hypercholesterolaemia: a machine learning study applied to electronic health encounter data. Lancet Digit Heal. 2019;

40. FIND FH® | The FH Foundation [Internet]. [cited 2020 Oct 10]. Available from: https://thefhfoundation.org/find-fh

41. Pina A, Helgadottir S, Mancina RM, Pavanello C, Pirazzi C, Montalcini T, et al. Virtual genetic diagnosis for familial hypercholesterolemia powered by machine learning. Eur J Prev Cardiol [Internet]. 2020 Oct 4 [cited 2020 Oct 11];27(15):1639–46. Available from: http://journals.sagepub.com/doi/10.1177/2047487319898951

42. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36.

43. Ruscio J, Mullen T. Confidence Intervals for the Probability of Superiority Effect Size Measure and the Area Under a Receiver Operating Characteristic Curve. Multivariate Behav Res [Internet]. 2012 Mar [cited 2020 Nov 8];47(2):201–23. Available from: https://www.tandfonline.com/doi/abs/10.1080/00273171.2012.658329

44. Haukoos JS, Lewis RJ. Advanced statistics: Bootstrapping confidence intervals for statistics with “difficult” distributions. Acad Emerg Med [Internet]. 2005 [cited 2020 Nov 8];12(4):360–5. Available from: https://pubmed.ncbi.nlm.nih.gov/15805329/

45. Hesse R, Raal FJ, George JA. Machine learning outperforms traditional screening and diagnostic tools for the detection of Familial Hypercholesterolaemia. American Association for Clinical Chemistry Annual Scientific Meeting 2020. Chicago; 2020. Available from: https://www.researchgate.net/publication/350384030_Machine_learning_outperforms_traditional_screening_and_diagnostic_tools_for_the_detection_of_Familial_Hypercholesterolemia

46. Hesse R, Raal FJ, George JA. Can machine learning improve on our current approaches of screening for, and diagnoses of, Familial Hypercholesterolemia? AACC Acad Sci Shorts [Internet]. 2021; Available from: https://www.aacc.org/science-and-research/scientific-shorts/2021/can-machine-learning-improve-screening-diagnosis-of-familial-hypercholesterolemia

47. Raal FJ, Bahassi EM, Stevens B, Turner TA, Stein EA. Cascade Screening for Familial Hypercholesterolemia in South Africa. Arterioscler Thromb Vasc Biol [Internet]. 2020 Sep 3 [cited 2020 Nov 2];40:2747–55. Available from: https://www.ahajournals.org/doi/suppl/10.1161/ATVBAHA.120.315040.

48. Stats SA. Mid-year population estimates 2019 [Internet]. 2019 Jul [cited 2020 Nov 2]. Available from: www.statssa.gov.za,info@statssa.gov.za

49. Khine AA, Marais AD. High prevalence of primary dyslipidaemia in black South African patients at a tertiary hospital in northern Gauteng, South Africa. South African Med J. 2016 Jul 1;106(7):724–9.