Computational and Structural Biotechnology Journal 16 (2018) 121–130

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Prediction of dyslipidemia using gene mutations, family history of diseases and anthropometric indicators in children and adolescents: The CASPIAN-III study

Hamid R. Marateb [a,b], Mohammad Reza Mohebian [a], Shaghayegh Haghjooy Javanmard [c], Amir Ali Tavallaei [a], Mohammad Hasan Tajadini [d], Motahar Heidari-Beni [e], Miguel Angel Mañanas [b,f], Mohammad Esmaeil Motlagh [g], Ramin Heshmat [h], Marjan Mansourian [c,i,*], Roya Kelishadi [j]

[a] Department of Biomedical Engineering, Facultyof Engineering, University of Isfahan, Isfahan, Iran
[b] Department of Automatic Control, Biomedical Engineering Research Center, Universitat Politècnica de Catalunya, BarcelonaTech (UPC), Barcelona, Spain
[c] Applied physiology researchcenter, Isfahan cardiovascular research institute, Isfahan University of Medical Sciences, Isfahan, Iran
[d] Department of Clinical Biochemistry, Tarbiat Modares University, Tehran, Iran
[e] Nutrition Department, Child Growth and Development Research Center, Research Institute for Primordial Prevention of Non-Communicable Disease,Isfahan University of Medical Sciences, Isfahan, Iran
[f] Biomedical Research Networking Center in Bioengineering, Biomaterialsand Nanomedicine (CIBER-BBN), Barcelona, Spain
[g] Department of Pediatrics, Ahvaz Jundishapur University of MedicalSciences, Ahvaz, Iran
[h] Department of Epidemiology, Chronic Diseases Research Center, Endocrinology and MetabolismPopulation Sciences Institute, Tehran University of Medical Sciences, Tehran, Iran
[i] Biostatistics and Epidemiology Department, Faculty of Health, Isfahan University of Medical Sciences, Isfahan, Iran
[j] Pediatrics Department, Child Growth and Development Research Center, Research Institute for Primordial Prevention of Non-Communicable Disease, Isfahan University of Medical Sciences, Isfahan, Iran

ARTICLE INFO

Article history:
Received 28 August 2017
Received in revised form 27 February 2018
Accepted 27 February 2018
Available online 2 March 2018

Keywords:
Computer-assisted diagnosis
Deep learning
Dyslipidemia
Genomics
Health promotion
Machine learning

ABSTRACT

Dyslipidemia, the disorder of lipoprotein metabolism resulting in high lipid profile, is an important modifiable risk factor for coronary heart diseases. It is associated with more than four million worldwide deaths per year. Half of the children with dyslipidemia have hyperlipidemia during adulthood, and its prediction and screening are thus critical. We designed a new dyslipidemia diagnosis system. The sample size of 725 subjects (age 14.66 ± 2.61 years; 48% male; dyslipidemia prevalence of 42%) was selected by multistage random cluster sampling in Iran. Single nucleotide polymorphisms (rs1801177, rs708272, rs320, rs328, rs2066718, rs2230808, rs5880, rs5128, rs2893157, rs662799, and Apolipoprotein-E2/E3/E4), and anthropometric, life-style attributes, and family history of diseases were analyzed. A framework for classifying mixed-type data in imbalanced datasets was proposed. It included internal feature mapping and selection, re-sampling, optimized group method of data handling using convex and stochastic optimizations, a new cost function for imbalanced data and an internal validation. Its performance was assessed using hold-out and 4-foldcross-validation. Four other classifiers namely as supported vector machines, decision tree, and multilayer perceptron neural network and multiple logistic regression were also used. The average sensitivity, specificity, precision and accuracy of the proposed system were 93%, 94%, 94% and 92%, respectively in cross validation. It significantly outperformed the other classifiers and also showed excellent agreement and high correlation with the gold standard. A non-invasive economical version of the algorithm was also implemented suitable for low- and middle-income countries. It is thus a promising new tool for the prediction of dyslipidemia.

1. Introduction

Strengthening the capacity of the entire countries, for early warning, and health risk reduction is one of the targets of the Sustainable Development Goal (SDG) #3. Non-communicable diseases (NCDs) have adverse human, social and economic consequences in all societies. Also, the first global NCD Action Plan is "A 25% relative reduction in the overall mortality from cardiovascular diseases, cancer, diabetes, or chronic respiratory diseases" [1]. Coronary heart diseases (CHDs), are the number 1 source of death and disability in countries including Iran [1,2]. Dyslipidemia, the disorder of lipoprotein metabolism resulting in high lipid profile, is a major risk factor of CHD [3]. It is related to more

* Corresponding author at: Department of biostatistics and Epidemiology, Health School, Isfahan University of Medical Sciences, Isfahan, Iran.
E-mail address: j_mansourian@hlth.mui.ac.ir (M. Mansourian).

https://doi.org/10.1016/j.csbj.2018.02.009

than four million deaths per year [4]. The accurate and reliable prediction of dyslipidemia is thus important in targeting SDG #3 and NCD Action Plan #1.

Metabolic risk factors including dyslipidemia are the most important determinants of emerging NCDs worldwide [5,6]. Dyslipidemia is, in fact, an important modifiable risk factor for CHD [7]. Although significant adverse health outcomes in childhood are not associated with dyslipidemia, it was shown in the literature that there is a link between childhood dyslipidemia and occurrence of atherosclerosis and its follow-up in adulthood [8,9]. Not only 40–55% of children with dyslipidemia will have hyperlipidemia during adulthood [10], but also subclinical atherosclerotic abnormalities, resulting in cardiovascular disease (CVD) events, occur in childhood [11]. Prediction and screening dyslipidemia, an important CVD risk factor, in children and adolescents is thus critical [12].

Some studies were performed in the literature to assess the genetic risk for dyslipidemia [13,14]. In such studies, statistically significant dyslipidemia predictors were identified, and no actual prediction (or classification) was performed. CAD (Computer-aided diagnosis), on the other hand, could use risk factors and predict if a subject is at high risk or not. CAD, which is using data mining to interpret medical information, could improve the diagnosis accuracy [15]. CAD is in fact used as a second opinion by the physicians to make the final diagnosis or prognosis decision [16–18].

Two methods were proposed in the literature to predict dyslipidemia in adults [19,20]. Wang et al. [19] analyzed 8914 subjects aged 35–78 years (with the prevalence of dyslipidemia about 46%). The predictors' age, gender, occupation, education, marital status, physical activity, individual income, waist circumference, smoking, family history of dyslipidemia, and diet were used to predict dyslipidemia (High TC, or TG or low HDL-C [21]). Artificial neural network (ANN) and Multiple Logistic Regression (MLR) models were used and the sensitivity, specificity, and precision of 90%, 77%, and 76% were obtained in the hold-out (75%) internal validation.

Costanza and Paccaud [20], analyzed 2549 subjects aged 35–64 years (the prevalence of dyslipidemia about 43%). The predictors waist-to-hip circumference ratio (WHR), body mass index (BMI), gender, age, current cigarette Smoking, and high blood pressure were used and dyslipidemia (total serum cholesterol to high-density lipoprotein cholesterol (TC/HDL-C) ratio ≥5.0) was predicted using different data mining methods, namely as the linear and logistic regressions, regression and classification trees. The sensitivity, specificity, and precision of 70%, 77%, and 69% were obtained in the hold-out external validation.

Although the prediction methods proposed in [19,20], are simple and effective and thus worthwhile for the identification of high risk people for having dyslipidemia based on the demographic, dietary and life-style, and anthropometric data, an optimal prediction is still required. Genome-based prediction of diseases has been recently focused in bioinformatics [22]. Identifying genetic mutations could assist in choosing optimal patient treatment. In fact, a lot of methods exist to reveal such mutations, including next-generation sequencing and future commercially available kits [23]. Moreover, in reliable clinical systems, critical criteria regarding statistical errors, precision, and DOR (Diagnosis Odds Ratio) must be met [24]. Moreover, considering ethnic differences in life-style, environmental factors and genetic background, examining gene polymorphisms associated with dyslipidemia in each ethnic group is important [13].

The purpose of our work is thus to design an *accurate* and *reliable* system for the prediction of dyslipidemia using gene mutations, family history of diseases and anthropometric indicators in a nationally-representative sample of the pediatric population in the Middle East and North Africa (MENA). To the best of our knowledge, this is the first study of its kind for genome-based dyslipidemia prediction using data mining.

2. Material and methods

2.1. Study population

The third study of a school-based surveillance system known as the childhood and adolescence surveillance and prevention of Adult Noncommunicable disease (CASPIAN) was conducted in Iran as the national survey of school students with high-risk behaviors (2009–2010) [25]. The description of the CASPIAN-III study was provided elsewhere in details [25]. Here, it is briefly described.

Among the youngsters, long-term changes in disease patterns are following rapid modifications in lifestyle, nutrition, and physical activity. Iranian youths are experiencing such lifestyle changes, making them prone to risk factors of chronic diseases such as NCDs. Surveilling such factors is important for long-term national planning based on monitoring NCD-related risk factors from childhood to adulthood. A school-based surveillance system entitled as CASPIAN Study was implemented in IRAN from 2003–2004. The surveys have been repeated every 2 years, with blood sampling for biochemical factors every 4 years.

This study was performed among 5570 students, sampled from 27 provinces of Iran. The entire students and their parents gave informed consent to the experimental procedure. It was approved by Isfahan University of Medical Sciences Panel on Medical Human Subjects and conformed to the Declaration of Helsinki.

According to the US National Institutes of Health Heart, Lung, and Blood Institute (NHLBI) guideline, which is one the acceptable criteria, dyslipidemia was defined for children and Adolescents (age ≤ 19 years) as having at least one of the following: TC (total cholesterol) ≥5.17 mmol/L (≥200 mg/dL), LDL-C (low-density lipoprotein cholesterol) ≥3.36 mmol/L (≥130 mg/dL), HDL-C (high-density lipoprotein cholesterol) levels <1.04 mmol/L (<40 mg/dL), TG (triglyceride) ≥1.13 mmol/L (≥100 mg/dL) when age is between zero and nine years and TG ≥1.47 mmol/L (≥130 mg/dL) when age is between 10 and 19 years, and finally non-HDL-C (subtracting HDL-C from TC) ≥3.75 mmol/L (≥145 mg/dL) [7,26].

We randomly selected 725 frozen whole blood samples for genome analysis from children and adolescents (48% male, 42% prevalence of dyslipidemia) taken from CASPIAN-III study. Such a sample size was estimated based on the sample-size estimation method proposed by Hajian-Tilaki [27]. Total required sample size (N) could be estimated based on the target sensitivity (Se_e) and Specificity (Sp_e) using Eq.(1):

$$N = \max\left(\frac{z^2_{\alpha/2} \times Se_e \times (1-Se_e)}{d^2 \times \text{Prev}}, \frac{z^2_{\alpha/2} \times Sp_e \times (1-Sp_e)}{d^2 \times (1-\text{Prev})}\right) \quad (1)$$

where α is the significance level, Prev is the prevalence of the disease in the population and d is the precision of estimate (i.e.,the maximum marginal error). The number of subjects in the case (n_{case}) and control ($n_{controls}$) categories could be then estimated using Eq.(2):

$$n_{controls} = N \times (1-\text{Prev}); \quad n_{case} = N - n_{controls} \quad (2)$$

The parameters Se_e and Sp_e were set to 70% and 77%, respectively based on the literature [20]. The prevalence of dyslipidemia in Iranian population was hypothesized as about 42% [6,28] and parameters α and d were both set to 0.05 [29]. Thus, the sample size of 725 ($n_{controls} = 418$, $n_{case} = 307$), sufficed.

2.2. Procedure and measurements

2.2.1. DNA extraction

Single nucleotide polymorphisms (SNPs) of lipoprotein lipase LPL (D9N [rs1801177]), cholesteryl ester transfer protein CETP (TaqIB [rs708272]) [30], LPL (HindIII [rs320]), LPL (S447X [rs328]) [31], ATP-binding cassette transporter-1 ABCA1 (V771M [rs2066718]), ABCA1

(R1587K [rs2230808]) [32], CETP (A373P [rs5880]) [33,34], apolipoprotein C-3 APOC3 (*Sst*I [rs5128]) [35], apolipoprotein A-1 APOA1 (MspI [rs2893157]) [36], apolipoprotein A-5 APOA5 (C-1131T [rs662799]) [37] and apolipoprotein-E ApoE genes [38,39], appearing to relate to lipid profile disorders and (or) cardiovascular diseases, were investigated [3,40].

Subjects' peripheral blood was analyzed using the QIAamp DNA Blood Mini kit (Qiagen, Germany) and DNA was extracted following the manufacturer's protocol [41]. Corbett rotor-gene 6000 instruments (Corbett Research Pty Ltd, Sydney Australia) were used for Real-time PCR and high- resolution melt analysis [42]. The details of later analysis were mentioned in the Supplementary material S1.

Alleles of the genotypes were analyzed. Typically, only two out of the four possible nucleotides occur, and each sample contains a pair of every autosome. Alternatively, the carrier and non-carrier genes were represented as a binary variable for each genotype. For example, for the SNP rs320, nucleotide pairs GG, and TG/GT with the minority nucleotide G were considered as 'carrier' while the TT pair was set to 'non-carrier'. Thus, two feature sets (nucleotide pairs, and carrier/non-carriervariables) were considered for further analysis.

2.2.2. Other analyzed features

The Anthropometric information was recorded by a team of trained health care professionals and the examinations were conducted under standard protocol by using calibrated instruments. Weight was measured to the nearest 200g in barefoot and lightly dressed condition. BMI was calculated as weight (kg) divided by height squared (m^2). The parameter weight circumference (WC) was measured using a non-elastic tape to the nearest 0.2 cm at the end of expiration at the midpoint between the top of iliac crest and the lowest rib in standing position [25].

The anthropometric and life-style attributes such as age, sex, hypertension (either high systolic blood pressure (SBP) (≥90th percentile for age, sex and height) or high diastolic blood pressure (DBP) (≥90th percentile for age, sex and height) [43]), abdominal obesity (defined as waist-to-height ratio (WHtR) equal or more than 0.5 [44]), BMI categories (underweight, normal, overweight and obese defined using WHO growth curves [45]) and physical activity (low, moderate, and severe categories [46]), as well as the family history of diabetes, obesity, CVD, cancer, and birth weight (<2500 g (low), 2500 g–4000 g (medium), and >4000 g (high) categories) were also included.

2.3. The proposed diagnosis system

2.3.1. Pre-processing

The dataset was split into the estimation, validation (overall known as the training set) and test sets (40%, 10%, and 50% respectively in a hold-out validation setting). The input variables were grouped based on their interval or categorical measurement scales [47]. The categorical group consisted of nominal (such as sex) and ordinal (such as birth order) variables. The interval features were then transferred using robust Z-score measure [48,49]. In this transformation, the median and MAD (median absolute deviation) of each feature was estimated, and the median was then reduced from each feature and then normalized by the MAD value. Such features were then normalized between zero and one for further processing.

For each categorical feature, the indicator variable was estimated. It takes the value 0 or 1 to indicate the absence or presence of each category. Logit transformation was performed on each indicator variable whose intercept and slope parameters were estimated using maximum Likelihood Estimating (MLE) on the training set [50]. Thus, each indicator variable was expressed as a continuous value between zero and one. Such processed features are entitled as "predictors" from now on. The number of predictors was N_p.

2.3.2. Optimized inductive learning

Group Method of Data Handling (GMDH), first proposed by Ivakhnenko [51,52], has been applied in many areas for data mining [53]. Inductive GMDH algorithms find interactions in data, select an optimal network structure and thus improve the performance of current algorithms [54]. Here we proposed an optimized GMDH method to predict dyslipidemia using mixed-type data.

Feature selection was performed by iteratively estimating their weights based on their capability to discriminate between neighboring patterns in the framework of the Expectation-Maximization algorithm using I-RELIEF algorithm [55]. Moreover, the parallel selective sampling (PSS) method was used to select data from the majority class as to reduce the problems in the imbalanced datasets [56].

Multilayered induction for the gradual increase of complexity was performed using different layers. Instead of the fixed regression polynomial, the nonlinear regression matrix (X) was formed between any pairs (i,j) of predictors at the first layer that has N_n nonlinear regression functions:

$$X_{N_n\times N_1}=\begin{bmatrix} a_1\times ones(1,N_1) \\ a_2\times(1+x(i,:))^{a_3} \\ a_4\times x(j,:)^{a_5} \\ a_6\times \sin(a_7\times x(i,:)+a_8\times x(j,:)) \\ a_9\times x(i,:)\odot(1+x(j,:))^{a_{10}} \\ a_{11}\times \log_2\left(1+|1+x(i,:)|^{a_{12}}\right) \\ a_{13}\times \log_2\left(1+|1+x(j,:)|^{a_{14}}\right) \\ \cdots \end{bmatrix} \tag{3}$$

where $\odot$ is the element-by-element multiplication, a_i is the regression coefficients and N_1 is the number of samples in the training set. If we fix the regression coefficients, the Regularized Least Squares (RLS) solution to $X^T\times W\approx B$ (B is a column vector with the class label of the analyzed samples) could be estimated as below:

$$W_{N_n\times 1}=\left(X\times X^T+\lambda\times I_{N_n\times N_n}\right)^{-1}\times X\times B \tag{4}$$

where λ is the regularization parameter (set to 0.1 in our study), I is the identity matrix, and T is the matrix transpose operator. It could be easily shown that the optimal solution is the global minimum point of the RLS optimization [57]. In principle, it is possible to tune polynomial regression coefficients using a stochastic optimization [58]. Instead, we tune the regression coefficients used in the matrix X, using Particle Swarm Optimization (PSO). PSO is a meta-heuristicspopulation-based method inspired by flocking birds [59]. The topology and the internal parameters of PSO were the same as Mohebian et al. [15] except that the maximum number of iterations was set to 10 and the PSO fitness function was defined differently. At each PSO iteration, the random regression coefficients are used to calculate the matrix X for a predictor pair. Then, the parameter W is estimated on the training set. To avoid overfitting, the estimated weight W is used on the validation set to estimate the output of the analyzed pair in the validation set. The cut-off of 0.5 was then used to estimate the parameters of signal detection theory such as True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Then, parameters Sensitivity ($=\frac{TP}{TP+FN}$), Specificity ($=\frac{TN}{TN+FP}$) and Precision ($=\frac{TP}{TP+FP}$) are estimated, and their average is used as the fitness function. The PSO method usually converged at few iterations due to the internal RLS optimization.

The selection pressure of the network was set to 0.7, in our study. Thus, 70% of the best pairs were selected for each layer. The approximating function of each selected pair was used as new features at the next layer [54]. The number of layers was estimated based on the required number of interactions. In a case of N_i interval features and N_d indicator variables, it was hypothesized as $1+round(\log_2(N_i+N_d))$. At the last layer, the best approximation function was used as the output of the classification system. the overall structure of the proposed prediction system was shown in the Supplementary material S2.

2.4. State-of-the-art

In our study, other classification methods namely as multilayer perceptron (MLP), MLR and decision tree (DT), as proposed in other studies [19,20], were used for comparison. Supported vector machines (SVM) was also used for comparison. MLP, a feed-forward artificial neural network (ANN) model mapping sets of inputs onto a set of outputs [60], with one hidden layer with ten neurons and the sigmoid activation function [61] was used. SVM, constructing a hyper plane in a high-dimensional space [62], with the radial basis function (RBF) kernels were used. The soft-margin parameter and the radius of the RBF kernel were tuned using the method proposed by Wu and Wang [63]. DT, building classification models in the form of a tree structure [64], uses entropy to calculate the homogeneity of samples to build the tree. The statistical classifier C4.5 with pruning (i.e., removing redundant subtrees) was used in our study [65]. The best splitting attribute is determined at each node. MLR uses the linear regression model with the Logit link function for the prediction.

After fitting the model [66], by estimating the model parameters, each case with the estimated class probability higher than 50%was classified as having dyslipidemia, or normal otherwise. In fact, DT and MLR could select relevant features because of the internal statistical validation. For MLP and SVM, Sequential Forward Selection (SFS) method, a bottom-up search procedure [67], was used for feature selection.

2.5. Validation

2.5.1. The performance indices for each classifier

The performance of the classifiers was determined using the holdout method, where the dataset was split into two mutually exclusive sets (50% training and 50% test). The classifiers were then trained on the training set and tested on the test set [68]. Moreover, 4-foldcross-validation (60% estimation, 15% validation, 25% test in each analysis fold) was used to test the best classifiers to control a possible biased error estimate [67]. A variety of performance indices [15,69,70] were reported for the analyzed classifiers. Such indices along with their definitions were shown in Table1, among which, MCC is a single unbiased performance measure in balanced as well as imbalanced datasets [71]. It is related to chi-square statistics, also known as phi-coefficient, a measure of association for two binary variables (predicted versus observed gold-standard class) that could be interpreted as the correlation coefficient between those binary variables [72]. The interpretation of the reference intervals of the indices AUC ROC [73], Kappa [74], MCC [75] and DP [69,76] was listed in Supplementary material S3.

A diagnosis system was considered as clinically reliable based on its Type I and II statistical errors [77], False Discovery Rate (FDR = 1-Precision) [78], and DOR [79] as to fulfill –all– the following conditions: the minimum Sensitivity, Specificity, Precision and DOR of 80%, 95%, 95% and 100, respectively.

2.5.2. Comparison between different classifiers

When different classifiers are compared with the gold standard, the superiority of one method to another must be presented using a proper statistical test. Otherwise, insignificant improvements might be erroneously reported as important [70]. McNemar's test, also known as the Gillick test, was used to compare the performance of two classifiers [67,80].

2.6. Statistical analysis

Results are reported as mean ± standard deviation (for interval variables) and frequencies (for categorical variables).The pairwise χ^2 analysis was used to test for allele frequency differences (and nominal features) between dyslipidemia and normal groups and when the Cochran conditions were not met, the Fisher exact test was used. The χ^2 analysis was used to test genotype frequency deviations from what predicted by the Hardy Weinberg equation. P-values less than 0.05 were considered significant. The entire data processing was performed off-line using Matlab version 8.6 (The MathWorks Inc., Natick, MA, USA). The statistical analysis and calculations were performed using the SPSS statistical package, version 16.0 (SPSS Inc., Chicago, IL, USA).

3. Results

The average age of the participants was 14.66 ± 2.61 years. Among the number of 725 patients participated in our study, 42.34% had dyslipidemia. Characteristics of the participants, grouped by their classification with/without dyslipidemia, are depicted in Table2. SNP genotype and allele frequencies in the study population were shown in Table3. None of the SNP distributions showed the deviation from Hardy-Weinberg equilibrium. Moreover, nucleotide pairs (Table3) showed better discrimination compared with carrier/non-carrier variables. Thus, nucleotide pairs, were used for prediction.

Three feature subsets were considered for prediction. Set 1 included sex, analyzed SNPs and family history of diseases: sex, LPL D9N [rs1801177], ABCAI V771M [rs2066718], LPL LPL HindIII [rs320], LPL S447X [rs328], ABCAI R1587K [rs2230808], CETP TaqIB [rs708272], APOC3 SstI [rs5128], CETP A373P [rs5880], APOA1 MspI [rs2893157], APOA5 C-1131T [rs662799], ApoE, Family history of diabetes, obesity, cancer, and CVD. Set 2 included Set 1 and birth weight, age, and physical activity. We also considered set 3 in which easily-measured features were analyzed, i.e., sex, age, physical activity, birth weight, BMI category, abdominal obesity, family history of diabetes, obesity, cancer, and CVD. The hold-out (50%) validation of the proposed method as well as the base learners DT, MLP, MLR, and SVM were performed in each feature subset, and the results of the classifiers on the test set were shown in Table4.

In each feature subset, the proposed method significantly outperformed the base learners (DT, MLP, MLR, and SVM) (P-value < 0.05). In the third subset, the entire base learners did not reject the NULL hypothesis of an accidental agreement. Moreover, in such classifiers, the AUC ROC was not significant (P-value < 0.05) showing

Table1
The classification performance measures used in our study.

$Se = Rl = \frac{TP}{TP+FN}$	$Sp = \frac{TN}{TN+FP}$	$Acc = \frac{TP+TN}{TP+TN+FP+FN}$
$\Pr = \frac{TP}{TP+FP}$	$FA = \alpha = 1 - Sp$	$Power = 1 - \beta = Se$
$F_1S = \frac{2\times \Pr\times Rl}{\Pr+Rl}$	$AUC = \frac{Se+Sp}{2}$	$LR^+ = \frac{Se}{1-Sp}$
$LR^- = \frac{1-Se}{Sp}$	$DOR = \frac{LR^+}{LR^-}$	$DP = (\sqrt{\frac{3}{\pi}}) \times \log(DOR)$
$MCC = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}}$	Kappa = agreement rate	

True positive (TP): subjects with dyslipidemia, correctly identified; false positive (FP): subjects without dyslipidemia, incorrectly identified; true negative (TN): subjects without dyslipidemia, correctly identified; false negative (FN): subjects with dyslipidemia, incorrectly identified; Se: sensitivity; Rl: recall; Sp: specificity; FA: false alarm; Acc: accuracy; Pr: precision; F_1S: F1-Score; AUC: area under the receiver operating characteristic (ROC) curve; LR: likelihood ratio; DOR: diagnosis odds ratio; MCC: Matthews correlation coefficient; DP: discriminant power; Kappa: Cohen's kappa coefficient defined as the agreement rate between the predicted class labels and the gold standard.

Table2
Characteristics of the participants in the dyslipidemia and normal groups.

Predictors	Categories	Dyslipidemia*		OR [CI 95%]	P-value
		No	Yes		
Age (years)		14.28 ± 2.26	14.64 ± 2.39	–	0.058
Sex	Male	49.28	46.61	0.90 [0.67,1.21]	0.477
	Female			–	
Region	Urban	64.80	71.71	1.38 [1.01,1.89]	0.049
	Rural			–	
Family history of diabetes	No	70.54	66.14	–	0.207
	Yes			1.23 [0.89,1.68]	
Family history of obesity	No	68.32	70.12	–	0.604
	Yes			0.92 [0.67,1.27]	
Family history of cancer	No	83.23	78.88	–	0.137
	Yes			1.33[0.91,1.93]	
Family history of CVD	No	87.16	92.43	–	0.023
	Yes			0.55 [0.33,0.93]	
Abdominal obesity	No	88.41	61.59	–	<0.001
	Yes			4.76 [3.26,6.94]	
BMI category (WHO criteria)	Under weight	25.85	19.52	0.76 [0.52,1.09]	0.007
	Normal	58.22	58.17	-	
	Over weight	8.36	10.76	1.29 [0.77,2.15]	
	Obese	7.57	11.55	1.53 [0.91,2.56]	
Physical activity	Mild	25.47	45.82	2.03 [1.43,2.87]	<0.001
	Moderate	40.37	35.86	–	
	High	34.16	18.32	0.60 [0.40,0.89]	
Birth weight	Low	11.67	16.73	1.54 [1.0,2.34]	0.249
	Normal	79.58	74.10	–	
	High	8.75	9.17	1.13 [0.67,1.89]	
Systolic blood pressure (mm Hg)		101.87 ± 13.16	104.16 ± 13.09	–	0.025
Diastolic blood pressure (mm Hg)		65.89 ± 10.74	66.69 ± 10.61	–	0.338
Fast blood sugar (mg/dL)		87.6 ± 11.85	84.32 ± 11.85	–	0.002
HDL-C (mg/dL)		59.95 ± 18.22	29.40 ± 12.37	–	<0.001
LDL-C (mg/dL)		75.43 ± 28.35	92.55 ± 38.09	–	<0.001
Total cholesterol (mg/dL)		149.66 ± 29.50	154.46 ± 30.20	–	0.061
Triglyceride (mg/dL)		86.06 ± 33.08	93.35 ± 34.35	–	<0.001

*: Results are reported as mean ± standard deviation (for interval variables) and percentage (for categorical variables). CVD: cardio-vascular disease; BMI: body mass index; WHO: world health organization; HDL-C: high-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol; OR: Odds ratio (a categorical level was set to reference for each categorical variable); CI: confidence interval. In each dyslipidemia group, the frequency percentage of one of the categories in binary variables was shown.

that none of them performed properly on the third subset. The proposed classifier on the set 2 significantly outperformed than sets 1 and 3 (P-value < 0.05). Also, the results of Set 1 was significantly better than those of Set 3 (P-value < 0.05).

The selected features of the proposed classifier on the Set 1 were CETP TaqIB [rs708272], CETP A373P [rs5880], LPL D9N [rs1801177], ApoE, ABCAI R1587K [rs2230808], APOA5 C-1131T [rs662799], LPL HindIII [rs320], APOC3 SstI [rs5128], family history of obesity, and diabetes, and APOA1 MspI [rs2893157]. Such features for Set 2 were CETP TaqIB [rs708272], ApoE, LPL D9N [rs1801177], ABCAI R1587K [rs2230808], age, birth weight, family history of obesity and for Set 3 were abdominal obesity, birth weight, physical activity, family history of diabetes, and BMI category. The performance of the best classifiers in each subset (i.e. the proposed classifier) was further assessed using 4-fold cross validation (Table5).

The proposed prediction system showed limited discriminant power (DP = 1.3), excellent diagnosis accuracy (AUC ROC = 0.94), excellent agreement with the gold standard (Kappa = 0.87) and high correlation with the gold standard (MCC=0.87) on the second subset (Table4). The average statistical power and Type I error (α) were 93 % and 0.07, respectively based on the cross-validation on the second subset (Table5). The training time of the proposed system was 26.1 ± 2.2 (s), 33.6 ± 3.0 (s) and 20.5 ± 3.1 (s)in the first, second, and third subsets, respectively. The average running time was the average of 3 runs over 363 subjects in the training set (hold-out 50%) on an Intel Core i7-6500uCPU with 8 GB of RAM.

4. Discussion

Identifying high-risk children based on gene polymorphisms (sets 1, and 2), at the first place, is useful for further dietary, and life-styletreatments and screening. Using life-style, anthropometric indicators and family history of diseases (set 3), on the other hand, could identify the high-risk population in low-income countries.

4.1. The risk factors of dyslipidemia

Although the environment is very important in the development of dyslipidemia, genetic components are also critical [81]. CETP TaqIB [rs708272] was selected by the proposed dyslipidemia prediction system in both sets 1 and 2. In the literature, Genome wide association studies (GWAS) in adults showed a high correlation between CETP and plasma lipid concentrations [82]. However, such an association is less distinct in children [33,83]. It was shown in the literature that such a mutation has the protective effect on dyslipidemia [33] and Myocardial Infarction (MI) [84]. This was in agreement with our findings, where the OR of CT/TT vs. CC was 0.15 (P-value<0.001) (Table3).

ApoE was also selected in both sets. ApoE, playing an important function in lipid metabolism, has three isoforms, Apo-e2, Apo-e3, and Apo-e4. They are in fact translated into three alleles of the gene. It was shown in the literature that ApoE , and particularly, its e4 isoform, is associated with plasma lipid parameters and CVD risks [85,86]. Similarly, in our study, the prevalence of dyslipidemia was 85% in subjects with ApoE-e4 isoforms. Moreover, the OR of e2/e4 vs. e3 was 1.73 (P-value < 0.001) (Table3).

ABCAI R1587K [rs2230808] was the other selected feature in both sets 1 and 2. Several ABCA1 gene polymorphisms including R1587K [rs2230808], were identified. Dean et al. showed that this SNP is associated with the HDL-C concentration [87], thus affecting dyslipidemia. In our study, the OR of AG/GG vs. AA was 2.21 (P-value < 0.001) (Table3). Thus, such polymorphisms increased the risk of dyslipidemia.

Table3
SNP genotype and allele frequencies (in percentage) of the participants in the dyslipidemia and normal groups.

Polymorphism	Genotype and allele*	Dyslipidemia		OR [CI 95%]	P-value
		No	Yes		
LPL D9N [rs1801177]	AA	96.4	91.2	–	0.003
	AG			2.59 [1.35–4.96]	
ABCAI V771M [rs2066718]	GG	94.0	98.7	-	0.002
	GA			0.21 [0.07–0.60]	
LPL HindIII [rs320]	GG	24.4	50.8	–	<0.001
	GT	48.6	42.0	0.31 [0.23–0.43]	
	TT	27.0	7.2		
LPL S447X [rs328]	CC	72.7	88.6	–	<0.001
	CG	24.6	10.4	0.34 [0.23–0.52]	
	GG	2.6	1.0		
ABCAI R1587K [rs2230808]	AA	66.7	47.6	–	<0.001
	AG	29.9	39.4	2.21 [1.64–3.00]	
	GG	3.3	13.0		
CETP TaqIB [rs708272]	CC	19.1	60.6	–	<0.001
	CT	61.7	35.5	0.15 [0.11–0.22]	
	TT	19.1	3.9		
APOC3 *SstI* [rs5128]	CC	83.0	83.7	–	0.371
	CG	16.7	15.3	0.95 [0.64–1.41]	
	GG	0.2	1.0		
CETP A373P [rs5880]	CC	93.5	77.9	–	<0.001
	CG	6.5	20.8	4.12 [2.56–6.62]	
	GG	0.0	1.3		
APOA1 MspI [rs2893157]	GG	69.4	74.3	–	0.119
	GA	27.8	24.8	0.79 [0.56–1.09]	
	AA	2.9	1.0		
APOA5 C-1131T [rs662799]	CC	98.8	97.7	–	0.525
	CT	0.5	1.0	1.93 [0.61–6.13]	
	TT	0.7	1.3		
ApoE	e2	6.9	0.7	1.73 [1.08–2.76]	<0.001
	e4	1.7	13.4		
	e3	91.4	86.0	–	

*: The alleles GG (SNP rs1801177) and CC (SNP rs2066718) had zero frequency in both normal and dyslipidemia groups and thus not shown in the results. OR: Odds ratio (a categorical level was set to reference for each categorical variable); CI: confidence interval. In each dyslipidemia group, the frequency percentage of one of the categories in binary variables was shown.

D9N [rs1801177] was the other commonly selected SNP in our study. Corsetti et al. showed that D9N is as a predictor of CVD risk directly and through its interaction with TaqIB [30]. In fact, LPL is involved with triglyceride-rich lipoprotein metabolism and lipoprotein remodeling including HDL [88,89]. Similarly in our study, the OR of (AG/GG vs. AA was 2.59 (P-value = 0.003) (Table3).

The family history of obesity was another common feature. Valdez et al. indicated that people who have one or more relatives with diabetes or CVD have a high risk of such problems [90]. Such diseases have common risk factors such as obesity and dyslipidemia sharing etiology [91]. FH of obesity, however, had poor agreement rate with FH of diabetes in our database (Cohen's Kappa = 0.24; P-value < 0.05). FH of diabetes was selected in the first and third subset, though. The prevalence of dyslipidemia in subjects without FH of obesity and diabetes were 43% and 41%, respectively.

Birth weight was a selected feature for the subsets 2 and 3. Rodríguez Vargas et al. showed that high birth weight is not a risk factor for hypercholesterolemia or HDL and LDL-cholesterol esters, but is positive for TG [92]. In our study the ORs of the low and high birth weight categories were more than one, but not significant (Table3). The prevalence of dyslipidemia in the abnormal and normal birth weight groups were 45% and 41%, respectively.

CETP A373P [rs5880] was selected in the first set. Agerholm-Larsen et al. indicated that such a polymorphism is associated with decreased HDL-C [93]. Heidari-Beni et al. showed that HDL-C levels were significantly lower among those with CETP A373P [rs5880] polymorphism [33]. In our study, the OR of CG/GG vs. CC was 4.12 (P-value < 0.001) (Table3).

APOA5 C-1131T [rs662799] was another selected SNP in the first set. Wang et al. indicated that this polymorphism is associated with dyslipidemia and the severity of CHD [94]. In our dataset, the OR of AG/GG vs. AA was 1.93, but it was not significant due to the small sample size of carrier genotypes (P-value = 0.525) (Table3).

Radha et al. found an association between LPL HindIII [rs320] SNP with low HDL-C and elevated TG levels [95]. Song et al. indicated a significant association between the APOC3 SstI [rs5128] polymorphism and higher levels of TG, TC, and LDL-C [35]. Albahrani et al. showed that APOA1 MspI [rs2893157] polymorphism is associated with CVD risk [36]. We did not find such an increased risk of dyslipidemia for LPL HindIII [rs320], APOC3 SstI [rs5128] and APOA1 MspI [rs2893157] SNPs. However, Odds (dyslipidemia| GG) was 1.5 in LPL HindIII [rs320] showing that this was possibly a good feature for the proposed classifier. Due to the small sample size of AA alleles in APOA1 MspI [rs2893157] and GG alleles in APOC3 SstI [rs5128] (Table3), no significant association between such polymorphisms and the risk of dyslipidemia was found.

Anthropometric indices such as BMI and WHtR were shown to be associated with dyslipidemia in children and adolescents in the literature [96]. In our study, people with abdominal obesity had 4.76 times risk of dyslipidemia (OR = 4.76; P-value < 0.001) compared with those without such an obesity (Table2). Moreover, overweight and obese subjects had a higher risk of dyslipidemia compared with normal BMI subjects (Table2). In fact, WHtR and BMI were moderately correlated (r = 0.737; P-value < 0.001). WHtR was poorly correlated with TG (r = 0.257; P-value < 0.001) while BMI was poorly correlated with SBP (r = 0.248; P-value < 0.001) and TG (r (Pearson's correlation) = 0.293; P-value < 0.001). They could be the reason why BMI and WHtR were selected by the proposed classifier on the third set.

Panagiotakos et al. showed that lipid profile disorders are correlated with physical activity [97]. In our dataset, the ORs of high and low physical activity compared with moderate activity were 0.60 (P-value < 0.001) and 2.03 (P-value < 0.001), respectively (Table2). It was poorly correlated with HDL levels (ρ (Spearman's correlation)

Table4
The hold-out (50%) validation of the classifiers.

Feature subset	Classifier	Se %	Sp %	Acc %	F_1S %	Pr %	FA	AUC	MCC	DOR	DP	Kappa
1	Proposed	85	91	88	86	87	0.09	0.88	0.76	57	1.0	0.76
	DT	69	80	75	70	72	0.20	0.75	0.47	9	0.5	0.46
	MLP	67	88	79	73	80	0.12	0.78	0.56	15	0.6	0.56
	MLR	61	86	75	68	76	0.14	0.74	0.49	10	0.5	0.49
	SVM	71	78	75	70	70	0.22	0.75	0.45	9	0.5	0.44
2	Proposed	93	95	94	93	93	0.05	0.94	0.87	252	1.3	0.87
	DT	71	81	77	72	73	0.19	0.76	0.50	10	0.6	0.50
	MLP	70	86	79	74	79	0.14	0.78	0.57	14	0.6	0.57
	MLR	59	87	75	67	77	0.13	0.73	0.48	10	0.5	0.47
	SVM	71	82	77	72	74	0.18	0.77	0.52	11	0.6	0.52
3	Proposed	82	84	83	80	79	0.16	0.83	0.64	24	0.8	0.64
	DT	48	68	60	50	52	0.32	0.58*	0.12	2	0.2	0.10*
	MLP	17	93	61	27	64	0.07	0.55*	0.16	3	0.2	0.13*
	MLR	17	94	61	27	68	0.06	0.56*	0.18	3	0.3	0.14*
	SVM	61	68	65	59	58	0.32	0.65*	0.17	3	0.3	0.12*

Set 1 included sex, analyzed SNPs and family history of diseases: sex, LPL D9N [rs1801177], ABCAI V771M [rs2066718], LPL HindIII [rs320], LPL S447X [rs328], ABCAI R1587K [rs2230808], CETP TaqIB [rs708272], APOC3 SstI [rs5128], CETP A373P [rs5880], APOA1 MspI [rs2893157], APOA5 C-1131T [rs662799], ApoE, Family history of diabetes, obesity, cancer, and CVD. Set 2 included Set 1 and birth weight, age, and physical activity. Set 3 included sex, age, physical activity, birth weight, BMI category, abdominal obesity, family history of diabetes, obesity, cancer, and CVD. The classifiers were trained on the same training set and then validated on the test set and the results of the classifiers on the test set were shown.

* Non-significant (P-value > 0.05).

= 0.252; P-value < 0.001). That could support its selection on the third set. Age was selected in the second set. Age was shown to be an independent predictor of dyslipidemia in children and adolescents [26]. Although age was directly used in the second set, age and sex are indirectly required for dyslipidemia prediction on the thirst set. The identification of BMI category in children and adolescents is dependent on the growth-curve charts that are gender and age specific [45].

4.2. Application in health policy making

The proposed automatic diagnosis of dyslipidemia on the third set is indeed an effective screening system. It used the input features of abdominal obesity, birth weight, physical activity, family history of diabetes, and BMI category. It includes therapeutic life-style change (e.g., dietary therapy, and increased physical activity), before necessary pharmacologic interventions [98]. In fact, the primary treatment for dyslipidemia in children and adolescents is such a life-style change [26].

Although the proposed system on the set 3 it is not a fully clinically reliable system (Type I error of 16% and FDR of 21%), it could be possibly used in low- and middle- income countries where genomics is not possible for a large population. Moreover, embedding the prediction system into a public online web-interface is useful in health promotion programs [15,99] that will be the focus in our future work.

4.3. The Properties and Performance of the proposed system

The proposed system for dyslipidemia prediction in the subset 2, showed promising results regarding variety of performance indices (Tables4 and 5). The statistical power, Type I error, FDR and DOR of the proposed system were 93%, 0.05, 7%, 252 (Table4). Thus, the proposed system fulfilled the criteria of a clinically reliable system except that it surpassed the minimum required FDR of 5% by 2%. We considered a variety of performance indices introduced in the literature (Tables1 and 2), and also the Standards for Reporting Diagnostic Accuracy (STARD 2015) and its extensions [70,100] in reporting the results. Guarding against testing hypotheses suggested by the data (Type III errors [101]) done by cross-validation and the low variation (high consistency) of the performance indices in different folds (Table5), excellent balanced diagnosis accuracy (AUC ROC = 0.94), excellent class labeling agreement rate (Kappa = 0.87), high correlation between predicted and observed class labels (MCC = 0.87), limited discriminant power (DP = 1.3) (Tables2 and 4), it is promising for clinical diagnosis tests. It significantly outperformed the other systems namely as DT, MLP, MLR, and SVM (McNemar's test; P-value<0.05).

Selecting only one kind of lipid disorder such as high total cholesterol/HDL-C ratio rather than dyslipidemia, could facilitate the interpretation of the results [20]. However, dyslipidemia contributes to cardio-metabolic risks in children and adolescents [102]. Moreover, In addition to cholesterol and HDL-C [103], triglyceride [104] and LDL-C [105] were shown to be important CVD risk factors. Thus, the outcome of the proposed system was dyslipidemia. We also considered high total cholesterol/HDL-C ratio outcome in our study and the selected features in the feature set 1 were ABCA1 (R1587K [rs2230808]), CETP (A373P [rs5880]), LPL (HindIII [rs320]), LPL (D9N [rs1801177]), and CETP (TaqIB [rs708272]). The AUC of this model was 0.82 in the hold-out validation.

Table5
The four-fold cross validation results of the proposed prediction system in MEAN ± SD.

Feature subset	Se %	Sp %	Acc %	Pr %
1	87± 2	90 ± 1	89 ± 1	86 ± 1
2	93± 2	94 ± 1	94 ± 1	92 ± 1
3	83± 2	84 ± 2	84 ± 1	79 ± 2

Se: sensitivity; Sp: specificity; Acc: accuracy; Pr: precision.

4.4. Further application of the proposed classification system

The proposed dyslipidemia prediction system made use of the following properties: I) mapping the mixed-data types to interval data using Logit function, II) RELIEF feature selection, III) PSS random sampling for imbalanced datasets, IV) the involvement of feature interactions proposed by GMDH, V) using the nonlinear regression matrix instead of a fixed regression polynomial, VI) using inner-loopRLS instead of LS, VII) using outer-loopPSO for stochastic optimization, VIII) using estimation, validation and test sets to avoid over-fitting, IX) internal cross validation on the training set (estimation plus validation set) to improve generalization capability, and X) proper cost function as the mean of Se, Sp, and Pr suitable for imbalanced data sets.

In fact, the proposed system could be regarded as a general framework for two-class classification of imbalanced mixed-type data given that it is successfully tested on different datasets. The following datasets were used for validation of the proposed framework: Wisconsin breast

Cancer (BCW), Pima Indian Diabetes (PIM), Glass [106], and Hepatitis [107]. The performance of the proposed framework on such datasets was shown in Supplementary material S4.

4.5. Final considerations

The limitation of the current study is that it was a retrospective study. More sources of error are more common in such studies compared with prospective studies because of bias and possible confounders [108]. Also, the sample size must be increased as to improve the statistical power in our diagnosis system [109]. Moreover, instead of testing a small number of pre-specified genetic regions, performing GWAS could be used in the examination of a genome-wide set of genetic variants in the entire genome in different individuals. For instance, more-prevalent mutations in LDL receptor (LDLR) gene were associated with dyslipidemia such as familial hypercholesterolemia, which is associated with early severe atherosclerosis and CAD [110]. In our study, NHLBI guideline was used to define dyslipidemia in children and adolescents. However, other standards such as American Heart Association (AHA) guideline [111] exist. The AHA guideline has different cut-points for TG and HDL-C. It also does not have a non-HDL-C criterion. Using AHA guideline, the class labels might change; thus affecting the proposed classification system. Finally, external validation (i.e. assessing the performance of the model on datasets from different institutions) is required in addition to an internal validation (i.e. hold-out and cross-validation) [112]. Unlike Costanza and Paccaud who rightfully used external validation in assessing their proposed lipid-disorder prediction model [20], other studies such as Wang et al. [19] and our study in this field and many studies in the other data mining areas in the literature do have only traditional internal cross-validation. This is the other limitation of our study.

5. Conclusions

In conclusion, we proposed a computer-aided diagnosis system to predict dyslipidemia whose performance was assessed using different criteria and in different validation frameworks. It is accurate and precise and could be possibly used for screening and risk assessment in the health promotion programs for children and adolescents. The developed framework is available to interested readers upon request.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2018.02.009.

Acknowledgements

This work was supported by the People Programme (Marie Curie Actions) of the European Union Seventh Framework Programme(FP7/2007–2013) under REA grant agreement no. 600388 (TECNIOspring Programme), from the Agency for Business Competitiveness of the Government of Catalonia, ACCIÓ and from Spanish Ministry of Economy and Competitiveness-Spain (project DPI2014-59049-R). This work was supported in part by the University of Isfahan (7109) and Isfahan University of Medical Sciences (194030).

References

[1] W.H. Organization. Global status report on noncommunicable diseases 2014. World Health Organization; 2014.
[2] Roger VL. Epidemiology of myocardial infarction. Med Clin North Am 2007;91:537–52.
[3] Kwiterovich P. The Johns Hopkins textbook of dyslipidemia. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2010.
[4] W.H. Organization. Quantifying selected major risks to health. World Health Rep 2002;2002:47–97.
[5] Stevens GA, Singh GM, Lu Y, Danaei G, Lin JK, Finucane MM, et al. National, regional, and global trends in adult overweight and obesity prevalences. Popul Health Metr 2012;10:22.
[6] Hovsepian S, Kelishadi R, Djalalinia S, Farzadfar F, Naderimagham S, Qorbani M. Prevalence of dyslipidemia in Iranian children and adolescents: a systematic review. J Res Med Sci 2015;20:503–21.
[7] Expert panel on integrated guidelines for cardiovascular health and risk reduction in children and adolescents: summary report, Pediatrics 2011;128(Suppl. 5):S213–256.
[8] Daniels SR. Screening and treatment of dyslipidemias in children and adolescents. Horm Res Paediatr 2011;76(Suppl. 1):47–51.
[9] Psaty BM, Rivara FP. Universal screening and drug treatment of dyslipidemia in children and adolescents. JAMA 2012;307:257–8.
[10] Hatami M, Tohidi M, Mohebi R, Khalili D, Azizi F, Hadaegh F. Adolescent lipoprotein classifications according to National Health and Nutrition Examination Survey (NHANES) vs. National Cholesterol Education Program (NCEP) for predicting abnormal lipid levels in adulthood in a Middle East population. Lipids Health Dis 2012;11:107.
[11] Zachariah JP, de Ferranti SD. NHLBI integrated pediatric guidelines: battle for a future free of cardiovascular disease. Future Cardiol 2013;9:13–22.
[12] Weintraub WS, Daniels SR, Burke LE, Franklin BA, Goff Jr DC, Hayman LL, et al. Value of primordial and primary prevention for cardiovascular disease: a policy statement from the American Heart Association. Circulation 2011;124:967–90.
[13] Yamada Y, Matsuo H, Warita S, Watanabe S, Kato K, Oguri M, et al. Prediction of genetic risk for dyslipidemia. Genomics 2007;90:551–8.
[14] Kelishadi R, Haghjooy Javanmard S, Tajadini MH, Mansourian M, Motlagh ME, Ardalan G, et al. Genetic association with low concentrations of high density lipoprotein-cholesterol in a pediatric population of the Middle East and North Africa: the CASPIAN-III study. Atherosclerosis 2014;237:273–8.
[15] Mohebian MR, Marateb HR, Mansourian M, Mañanas MA, Mokarian F. A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) using optimized ensemble learning. Comput Struct Biotechnol J 2017;15:75–85.
[16] Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput Med Imaging Graph 2007;31:198–211.
[17] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8–17.
[18] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 2017;15:104–16.
[19] Wang C-J, Li Y-Q, Wang L, Li L-L, Guo Y-R, Zhang L-Y, et al. Development and evaluation of a simple and effective prediction approach for identifying those at high risk of dyslipidemia in rural adult residents. PLoS One 2012;7:e43834.
[20] Costanza MC, Paccaud F. Binary classification of dyslipidemia from the waist-to-hip ratio and body mass index: a comparison of linear, logistic, and CART models. BMC Med Res Methodol 2004;4:7.
[21] Committee CADP. China adult dyslipidemia prevention guide. Beijing, China: People's Health Publishing House; 2007; 390–419.
[22] Mooney SD, Krishnan VG, Evani US. Bioinformatic tools for identifying disease gene and SNP candidates. Methods Mol Biol 2010;628:307–19.
[23] Bai J, Gao J, Mao Z, Wang J, Li J, Li W, et al. Genetic mutations in human rectal cancers detected by targeted sequencing. J Hum Genet 2015;60:589.
[24] Guilherme HMO, Nesbitt GC, Murphy JG, Habermann TM. Mayo Clinic medical manual and Mayo Clinic internal medicine review. 7th ed. Rochester; MN, USA: CRC Press; 2007.
[25] Kelishadi R, Heshmat R, Motlagh ME, Majdzadeh R, Keramatian K, Qorbani M, et al. Methodology and early findings of the third survey of CASPIAN study: a National School-based Surveillance of Students' High Risk Behaviors. Int J Prev Med 2012;3:394–401.
[26] Yoon JM. Dyslipidemia in children and adolescents: when and how to diagnose and treat? Pediatr Gastroenterol Hepatol Nutr 2014;17:85–92.
[27] Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. J Biomed Inform 2014;48:193–204.
[28] Tabatabaei-Malazy O, Qorbani M, Samavat T, Sharifi F, Larijani B, Fakhrzadeh H. Prevalence of dyslipidemia in Iran: a systematic review and meta-analysis study. Int J Prev Med 2014;5:373–93.
[29] Machin D. Sample size tables for clinical studies. 3rd Ed. Chichester, West Sussex, UK; Hoboken, NJ: Wiley-Blackwell; 2008.
[30] Corsetti JP, Gansevoort RT, Navis G, Sparks CE, Dullaart RP. LPL polymorphism (D9N) predicts cardiovascular disease risk directly and through interaction with CETP polymorphism (TaqIB) in women with high HDL cholesterol and CRP. Atherosclerosis 2011;214:373–6.
[31] Peacock RE, Hamsten A, Nilsson-Ehle P, Humphries SE. Associations between lipoprotein lipase gene polymorphisms and plasma correlations of lipids, lipoproteins and lipase activities in young myocardial infarction survivors and age-matched healthy individuals from Sweden. Atherosclerosis 1992;97:171–85.
[32] Frikke-Schmidt R, Nordestgaard BG, Jensen GB, Steffensen R, Tybjaerg-Hansen A. Genetic variation in ABCA1 predicts ischemic heart disease in the general population. Arterioscler Thromb Vasc Biol 2008;28:180–6.
[33] Heidari-Beni M, Kelishadi R, Mansourian M, Askari G. Interaction of cholesterol ester transfer protein polymorphisms, body mass index, and birth weight with the risk of dyslipidemia in children and adolescents: the CASPIAN-III study. IranJ Basic Med Sci 2015;18:1079–85.
[34] Kontush A, Chapman MJ. High-density lipoproteins: structure, metabolism, function, and therapeutics. Hoboken, N.J.: John Wiley & Sons, Inc.; 2012
[35] Song Y, Zhu L, Richa M, Li P, Yang Y, Li S. Associations of the APOC3 rs5128 polymorphism with plasma APOC3 and lipid levels: a meta-analysis. Lipids Health Dis 2015;14:32.

[36] Albahrani AI, Usher JJ, Alkindi M, Marks E, Ranganath L, Al-yahyaee S. ApolipoproteinA1-75 G/A (M1-) polymorphism and lipoprotein(a); anti- vs. pro-atherogenic properties. Lipids Health Dis 2007;6:19.
[37] Xu C, Bai R, Zhang D, Li Z, Zhu H, Lai M, et al. Effects of APOA5 −1131T>C (rs662799) on fasting plasma lipids and risk of metabolic syndrome: evidence from a case-control study in China and a meta-analysis. PLoS One 2013;8:e56216.
[38] Yin Y-W, Sun Q-Q, Zhang B-B, Hu A-M, Liu H-L, Wang Q, et al. Association between apolipoprotein E gene polymorphism and the risk of coronary artery disease in Chinese population: evidence from a meta-analysis of 40 studies. PLoS One 2013;8:e66924.
[39] Tudorache IF, Trusca VG, Gafencu AV. Apolipoprotein E- a multifunctional protein with implications in various pathologies as a result of its structural features. Comput Struct Biotechnol J 2017;15:359–65.
[40] Brunham LR, Hayden MR. Human genetics of HDL: insight into particle metabolism and function. Prog Lipid Res 2015;58:14–25.
[41] QIAamp D. Mini and Blood Mini Handbook. Qiagen; 2016.
[42] Askari G, Heidari-Beni M, Mansourian M, Esmaeil-Motlagh M, Kelishadi R. Interaction of lipoprotein lipase polymorphisms with body mass index and birth weight to modulate lipid profiles in children and adolescents: the CASPIAN-III study. Sao Paulo Med J 2016;134:121–9.
[43] Daniels SR. How to Define Hypertension in Children and Adolescents. Circulation 2016;133:350–1.
[44] Li C, Ford ES, Mokdad AH, Cook S. Recent trends in waist circumference and waist-height ratio among US children and adolescents. Pediatrics 2006;118:e1390-398.
[45] Mansourian M, Marateb HR, Kelishadi R, Motlagh ME, Aminaee T, Taslimi M, et al. First growth curves based on the World Health Organization reference in a nationally-representative sample of pediatric population in the Middle East and North Africa (MENA): the CASPIAN-III study. BMC Pediatr 2012;12:149.
[46] Kelishadi R, Majdzadeh R, Motlagh ME, Heshmat R, Aminaee T, Ardalan G, et al. Development and evaluation of a questionnaire for assessment of determinants of weight disorders among children and adolescents: the Caspian-IV study. Int J Prev Med 2012;3:699–705.
[47] Marateb HR, Mansourian M, Adibi P, Farina D. Manipulating measurement scales in medical statistical analysis and data mining: a review of methodologies. J Res Med Sci 2014;19:47–56.
[48] Birmingham A, Selfors LM, Forster T, Wrobel D, Kennedy CJ, Shanks E, et al. Statistical methods for analysis of high-throughputRNA interference screens. Nat Methods 2009;6:569–75.
[49] Liu W, Ju Z, Lu Y, Mills GB, Akbani R. A comprehensive comparison of normalization methods for loading control and variance stabilization of reverse-phase protein array data. Cancer Inform 2014;13:109–17.
[50] Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006.
[51] Ivakhnenko AG. Heuristic self-organization in problems of engineering cybernetics. Automatica 1970;6:207–19.
[52] Ivakhnenko AG. Polynomial theory of complex systems. IEEE Trans Syst Man Cybern 1971:364–78 SMC-1.
[53] Bozdogan H. Statistical data mining and knowledge discovery. Boca Raton, FL: Chapman & Hall/CRC; 2004.
[54] Madala HR, Ivakhnenko AGe. Inductive learning algorithms for complex systems modeling. Boca Raton: CRC Press; 1994.
[55] Sun Y. Iterative RELIEF for feature weighting: algorithms, theories, and applications. IEEE Trans Pattern Anal Mach Intell 2007;29:1035–51.
[56] D'Addabbo A, Maglietta R. Parallel selective sampling method for imbalanced and large data classification. Pattern Recognit Lett 2015;62:61–7.
[57] Beck A. Introduction to nonlinear optimization: theory, algorithms, and applications with MATLAB, 282, Philadelphia: Society for Industrial and Applied Mathematics: Mathematical Optimization Society; 2014 (ISBN: 978-1-61197-364-8e, ISBN: 978-1-61197-365-5).
[58] Onwubolu GC. Design of hybrid differential evolution and group method of data handling networks for modeling and prediction. Inform Sci 2008;178:3616–34.
[59] Eberhart R, Kennedy J. A new optimizer using particle swarm theory, micro machine and human science, 1995MHS '95., Proceedings of the Sixth International Symposium; 1995. p. 39–43.
[60] Wasserman PD, Schwartz T. Neural networks. II. What are they and why is everybody so interested in them now? IEEE Expert 1988;3:10–5.
[61] Isa IS, Saad Z, Omar S, Osman MK, Ahmad KA, Sakim HAM. Suitable MLP network activation functions for breast cancer and thyroid disease detection2010 Second International Conference on Computational Intelligence, Modelling and Simulation; 2010. p. 39–44.
[62] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97.
[63] Wu K-P, Wang S-D. Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space. Pattern Recognit 2009;42:710–7.
[64] Quinlan JR. Induction of decision trees. Mach Learn 1986;1:81–106.
[65] Salzberg SL. C4.5: programs for machine learning by J. Ross Quinlan. Machine Learning, 16. Morgan Kaufmann Publishers, Inc., 1993; 1994; 235–40.
[66] Freedman D. Statistical models: theory and practice. Cambridge; New York: Cambridge University Press; 2009.
[67] Webb AR, Copsey KD. Statistical pattern recognition. 3rd Ed. Hoboken: Wiley; 2011.
[68] Sammut Claude, Webb Geoffrey I, editors. Encyclopedia of machine learning and data mining. New York, NY: Springer Berlin Heidelberg; 2016. p. 1335.
[69] Sokolova M, Japkowicz N, Szpakowicz S. In: Sattar A, Kang B-h, editors. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. AI 2006: Advances in Artificial Intelligence: 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, December 4-8, 2006. Proceedings. Berlin Heidelberg, Berlin, Heidelberg: Springer; 2006. p. 1015–21.
[70] Marateb HR, Mansourian M, Mañanas MA. Re: STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2017 [http://www.bmj.com/content/351/bmj.h5527/rr-1].
[71] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–51.
[72] Ernest J, Davenport C, El-Sanhurry NA. Phi/Phimax: review and synthesis. Educ Psychol Meas 1991;51:821–8.
[73] Simundic AM. Measures of diagnostic accuracy: basic definitions. Ejifcc 2009;19:203–11.
[74] Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. 3rd Ed. Hoboken, N.J.: J. Wiley; 2003
[75] Mukaka MM. Statistics corner: a guide to appropriate use of correlation coefficient in medical research. Malawi Med J 2012;24:69–71.
[76] Mert A, Kilic N, Bilgili E, Akan A. Breast cancer detection with reduced feature set. Comput Math Methods Med 2015;2015:265138.
[77] Ellis PD. The essential guide to effect sizes: statistical power, meta-analysis, and the interpretation of research results. Cambridge; New York: Cambridge University Press; 2010.
[78] Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci 2014;1.
[79] Ghosh AK, Wittich CM, Rhodes DJ, Beckman TJ, Edson RS. Mayo clinic internal medicine review. Rochester, MN: Informa Healthcare; 2008.
[80] Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 1998;10:1895–923.
[81] Nock NL. Genetics of lipid disorders. In: Ahima RS, editor. Metabolic syndrome: a comprehensive Textbook. Cham: Springer International Publishing; 2016. p. 159–93.
[82] Kathiresan S, Manning AK, Demissie S, D'Agostino RB, Surti A, Guiducci C, et al. A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. BMC Med Genet 2007;8(Suppl. 1):S17.
[83] Smart MC, Dedoussis G, Louizou E, Yannakoulia M, Drenos F, Papoutsakis C, et al. APOE, CETP and LPL genes show strong association with lipid levels in Greek children. Nutr Metab Cardiovasc Dis 2010;20:26–33.
[84] Cao M, Zhou ZW, Fang BJ, Zhao CG, Zhou D. Meta-analysis of cholesteryl ester transfer protein TaqIB polymorphism and risk of myocardial infarction. Medicine 2014;93:e160.
[85] Wilson PW, Myers RH, Larson MG, Ordovas JM, Wolf PA, Schaefer EJ. Apolipoprotein E alleles, dyslipidemia, and coronary heart disease. The Framingham offspring study. JAMA 1994;272:1666–71.
[86] Zende PD, Bankar MP, Kamble PS, Momin AA. Apolipoprotein e gene polymorphism and its effect on plasma lipids in arteriosclerosis. J Clin Diagn Res 2013;7:2149–52.
[87] Dean M, Hamon Y, Chimini G. The human ATP-binding cassette (ABC) transporter superfamily. J Lipid Res 2001;42:1007–17.
[88] Stein Y, Stein O. Lipoprotein lipase and atherosclerosis. Atherosclerosis 2003;170:1–9.
[89] Murdoch SJ, Breckenridge WC. Influence of lipoprotein lipase and hepatic lipase on the transformation of VLDL and HDL during lipolysis of VLDL. Atherosclerosis 1995;118:193–212.
[90] Valdez R, Greenlund KJ, Khoury MJ, Yoon PW. Is family history a useful tool for detecting children at risk for diabetes and cardiovascular diseases? A public health perspective. Pediatrics 2007;120(Suppl. 2):S78–86.
[91] Stern MP. Do non-insulin-dependent diabetes mellitus and cardiovascular disease share common antecedents? Ann Intern Med 1996;124:110–6.
[92] Rodriguez Vargas N, Martinez Perez TP, Martinez Garcia R, Garriga Reyes M, Ortega Soto M, Rojas T. Dyslipidemia in schoolchildren with a history of a high birth weight. Clin Investig Arterioscler 2014;26:224–8.
[93] Agerholm-Larsen B, Tybjærg-Hansen A, Schnohr P, Steffensen R, Nordestgaard BG. Common cholesteryl ester transfer protein mutations, decreased HDL cholesterol, and possible decreased risk of ischemic heart disease, the Copenhagen City Heart Study. 2000;102:2197–203.
[94] Wang Y, Lu Z, Zhang J, Yang Y, Shen J, Zhang X, et al. The APOA5 rs662799 polymorphism is associated with dyslipidemia and the severity of coronary heart disease in Chinese women. Lipids Health Dis 2016;15:170.
[95] Radha V, Mohan V, Vidya R, Ashok AK, Deepa R, Mathias RA. Association of lipoprotein lipase Hind III and Ser 447 Ter polymorphisms with dyslipidemia in Asian Indians. Am J Cardiol 2006;97:1337–42.
[96] Hashemipour M, Soghrati M, Malek Ahmadi M. Anthropometric indices associated with dyslipidemia in obese children and adolescents: a retrospective study in isfahan. ARYA Atheroscler 2011;7:31–9.
[97] Panagiotakos DB, Pitsavos C, Chrysohoou C, Skoumas J, Zeimbekis A, Papaioannou I, et al. Effect of leisure time physical activity on blood lipid levels: the ATTICA study. Coron Artery Dis 2003;14:533–9.
[98] Daniels SR, Pratt CA, Hayman LL. Reduction of risk for cardiovascular disease in children and adolescents. Circulation 2011;124:1673–86.
[99] Safran Naimark J, Madar Z, Shahar DR. The impact of a web-based app (eBalance) in promoting healthy lifestyles: randomized controlled trial. J Med Internet Res 2015;17:e56.
[100] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351.
[101] Mosteller F. A k-sample slippage test for an extreme population. Ann Math Stat 1948;19:58–65.
[102] Kelishadi R, Heidari-Beni M, Qorbani M, Motamed-Gorji N, Motlagh ME, Ziaodini H, et al. Association between neck and wrist circumferences and cardiometabolic risk in children and adolescents: The CASPIAN-V study. Nutrition 2017;43-44:32–8.

[103] Sarrafzadegan N, Hassannejad R, Marateb HR, Talaei M, Sadeghi M, Roohafza HR, et al. PARS risk charts: a 10-year study of risk assessment for cardiovascular diseases in Eastern Mediterranean Region. PLoS One 2017;12:e0189389.
[104] Reiner Z. Hypertriglyceridaemia and risk of coronary artery disease. Nat Rev Cardiol 2017;14:401–11.
[105] Klag MJ, Ford DE, Mead LA, He J, Whelton PK, Liang KY, et al. Serum cholesterol in young men and subsequent cardiovascular disease. N Engl J Med 1993;328:313–8.
[106] Bonissone P, Cadenas JM, Carmen Garrido M, Andrés Díaz-Valladares R. A fuzzy random forest. Int J Approx Reason 2010;51:729–47.
[107] Raymer ML, Doom TE, Kuhn LA, Punch WF. Knowledge discovery in medical and biological datasets using a hybrid Bayes classifier/evolutionary algorithm. IEEE Trans Syst Man Cybern B Cybern 2003;33:802–13.
[108] Spark A. Nutrition in public health: principles, policies, and practice. Boca Raton: CRC Press; 2007.
[109] Rubin A. Statistics for evidence-based practice and evaluation. 3rd Ed. Cengage Learning, Belmont, CA: Brooks/Cole; 2013.
[110] Guardamagna O, Restagno G, Rolfo E, Pederiva C, Martini S, Abello F, et al. The type of LDLR gene mutation predicts cardiovascular risk in children with familial hypercholesterolemia. J Pediatr 2009;155:199–204 [.e192].
[111] Kavey R-EW, Daniels SR, Lauer RM, Atkins DL, Hayman LL, Taubert K. American Heart Association Guidelines for primary prevention of atherosclerotic cardiovascular disease beginning in childhood. Circulation 2003;107:1562–6.
[112] Taylor JMG, Ankerst DP, Andridge RR. Validation of biomarker-based risk prediction models. Clin Cancer Res 2008;14:5977–83.