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