Atherosclerosis 277 (2018) 262–266

Contents lists available at ScienceDirect

Atherosclerosis

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

Identifying patients with familial hypercholesterolemia using data mining methods in the Northern Great Plain region of Hungary

György Paragh [a, *], Mariann Harangi [a], Zsolt Karányi [a], Bálint Daróczy [b], Ákos Németh [c], Péter Fülöp [a]

[a] Department of Internal Medicine, University of Debrecen Faculty of Medicine, Debrecen, Hungary
[b] Institute for Computer Science and Control, Hungarian Academy of Sciences, (MTA SZTAKI), Budapest, Hungary
[c] Aesculab Medical Solutions, Black Horse Group Ltd., Debrecen, Hungary

ARTICLE INFO

Article history:
Received 30 March 2018
Received in revised form
4 May 2018
Accepted 22 May 2018

Keywords:
Familial hypercholesterolemia
Screening
Low-density lipoprotein
Dutch Lipid Clinic Network criteria
Data mining
Deep learning

ABSTRACT

Background and aims: Familial hypercholesterolemia (FH) is one of the most frequent diseases with monogenic inheritance. Previous data indicated that the heterozygous form occurred in 1:250 people. Based on these reports, around 36,000–40,000 people are estimated to have FH in Hungary, however, there are no exact data about the frequency of the disease in our country. Therefore, we initiated a cooperation with a clinical site partner company that provides modern data mining methods, on the basis of medical and statistical records, and we applied them to two major hospitals in the Northern Great Plain region of Hungary to find patients with a possible diagnosis of FH.
Methods: Medical records of 1,342,124 patients were included in our study. From the mined data, we calculated Dutch Lipid Clinic Network (DLCN) scores for each patient and grouped them according to the criteria to assess the likelihood of the diagnosis of FH. We also calculated the mean lipid levels before the diagnosis and treatment.
Results: We identified 225 patients with a DLCN score of 6–8 (mean total cholesterol: 9.38 ± 3.0 mmol/L, mean LDL-C: 7.61 ± 2.4 mmol/L), and 11,706 patients with a DLCN score of 3–5 (mean total cholesterol: 7.34 ± 1.2 mmol/L, mean LDL-C: 5.26 ± 0.8 mmol/L).
Conclusions: The analysis of more regional and country-wide data and more frequent measurements of total cholesterol and LDL-C levels would increase the number of FH cases discovered. Data mining seems to be ideal for filtering and screening of FH in Hungary.

1. Introduction

Healthcare data indicate that cardiovascular diseases are the leading cause of death in Europe. Hungarian data are even less favorable since, compared to the EU-15 countries, the life expectancy is shorter by 6.8 years among men and by 4.8 years among women at the age of 40. Indeed, the risk of premature mortality caused by cardiovascular diseases (CVDs) is approximately three times higher in the Central Eastern European region than in the Western European countries. Additionally, only 5 years are expected to be spent in health after the age of 65 in Hungary, while this number is 12.5 years in the three best performing EU states, which shows a significant gap *versus* the developed Western European states [1,2].

Hyperlipidemia is a major risk factor of cardiovascular diseases. Increased blood cholesterol levels contribute significantly to atherosclerosis, therefore, diseases resulting in excessively elevated cholesterol concentrations lead to premature cardiovascular complications even at younger age [3]. Familial hypercholesterolemia (FH) is one of the most frequent diseases with monogenic inheritance caused by various mutations in the genes encoding the low-density lipoprotein (LDL) receptor, apolipoprotein (Apo) B100 and the proprotein convertase subtilisin/kexin type 9 (PCSK9) [4]. Previous data indicates that the heterozygous form of FH occurs in 1:500 subjects, while the homozygous form develops in 1:1,000,000 [5]. Recent studies also brought attention to certain populations where familial hypercholesterolaemia appears to be more frequent. In Holland, 1 person out of 200 was found to have heterozygous FH [6]; while a recent meta-analysis indicates the FH

* Corresponding author. Department of Internal Medicine, University of Debrecen Faculty of Medicine, Nagyerdei krt. 98, H-4032, Debrecen, Hungary.
E-mail address: paragh@belklinika.com (G. Paragh).

https://doi.org/10.1016/j.atherosclerosis.2018.05.039
0021-9150/

frequency of 1:250. FH prevalence appears to vary by age and geographical location [7]. 10–30 million people are estimated to have FH globally, although 80% of the cases are not diagnosed. It has to be mentioned that only 10% of the diagnosed patients reach the target LDL level and studies indicate that patients with FH die 15 years earlier compared to those without [8]. Other studies indicate a 3.5–16 times increased risk of coronary artery disease (CAD) and a 5–10 times increased risk of peripheral arterial disease (PAD) in heterozygous FH patients [9–11].

These data highlight that FH is a major challenge in cardiovascular disease prevention. Around 20,000–40,000 people are estimated to have FH in Hungary, however, there are no exact data about the frequency of the disease in our country. Therefore, to assess its real prevalence in Hungary, we created an online FH registry in 2016 (http://fhreg.hu/). The project started with three purposes: (1) to inform the broader (lay) population about the disease, (2) to provide information about FH to family doctors emphasizing the screening possibilities, (3) to have suspected FH patients registered by physicians. Our FH registry is based upon the Dutch Lipid Clinic Network (DLCN) criteria [12] and score is calculated using the clinical and laboratory data provided by the colleagues. Patients with a possible diagnosis of FH are registered to their regional lipid centers, where the final diagnosis is made, together with risk stratification, and therapy is initialized. Including the 2 national centers in Budapest and Debrecen, there are 18 regional lipid centers in Hungary. Based upon the data mentioned above, we estimated the number of patients expected to be registered in each center. Our primary goal was to find approximately 10% of the suspected FH patients in the first year after commencing the project. We also aimed to gather specific information about the disease and to start treatment as soon as possible to improve health statistics and life expectancy in the region. After running the project for two years, we found that patient enrollment was not satisfactory, thus we looked for other methods to find FH patients in Hungary.

We initiated a cooperation with a clinical site partner company to utilize their medical system framework, which provides modern data mining methods on the basis of medical and statistical records and we applied it to two major hospitals in the Northern Great Plain region of Hungary. We supposed that we could identify more FH patients and we also targeted to test the potential usage and scope of the software.

To identify patients with possible, probable and definite diagnosis of FH, we relied on the DLCN criteria, which are based on the family history of the patients, their own clinical history, physical signs, untreated LDL-C levels and DNA analysis [12]. The accessible data were poor in family history and DNA analysis, so we focused mainly on the other three criteria generated automatically from the databases. Most of the time was spent in the pre-processing phase to make data comparable from various sources.

2. Materials and methods

Two leading medical centers, University of Debrecen Clinical Center and County Hospital of Szabolcs-Szatmár Bereg, provided access to anonymous medical records for software development purposes from the Northern Great Plain region of Hungary. The data source contained all medical records from these two centers between January 1, 2007 and December 31, 2014. We set up a data mining cooperation with a partner company (Black Horse Group Ltd.) to utilize their medical system framework named “AescuLab” (www.aesculab.net). First, data were extracted from the clinical record systems after anonymization to protect patient privacy. The records included several tables with unique identifications per case and patient, but without the possibility to link them to the real patients. We used open source tools (http://pandas.pydata.org/, http://www.numpy.org/) as well as our self-developed scripts and solutions to clean the data and fill the missing or corrupt data parts. From all separated data, we built a complete concatenated data source containing laboratory cases, textual history data, diagnosis codes and patient statistic data. We also built special serializing and buffering methods to process data and avoid obvious memory problems of this massive data source.

Regular preprocessing steps of any textual information were parsing, stemming (http://hunspell.github.io/), bag-of-words (BOW) modelling and ranking of expressions with “Term Frequency - Inverse Document Frequency” (TF-IDF) [13] and word2vec (W2V) modelling performed in Keras (https://keras.io/) to identify important expressions [14]. The BOW models describe a document as a histogram of occurring terms or expressions without taking advantage of the sequential structure. This results in robustness in representation and invariance in case of comparable documents with different sequential structure. The W2V models describe a term or expression in a document with an element in a vector space defined by a neural network. These families of models are based on a simple language model where the contextual terms determine the actual elements in a sequence utilizing the sequential structure. Both models are suitable to identify the importance of the terms and expressions in a natural way by ranking them based on their IDF score [13] or their perplexity [14]. Besides, we collected a list of important expressions based on expert knowledge and utilized string matching algorithms to overcome regular misspelling and recover expressions based on partial information.

Since one of the extracted data incorporates regular, unprocessed anamnesis, additional pre-processing procedures were necessary, such as text extraction and content identification with regular expressions. The resulted data embrace a finite set of expressions with a simple indicator of occurrence per record with an additional value in case of medical examinations. The list of expressions initially included several million elements, which we reduced to 250 thousands with the above mentioned methods. Connecting the records of the patients, their cases and diagnoses, we described the medical history as a series of events in time associated with the patients. This format allowed us to identify potential patients with familial hypercholesterolemia and their medical history of hypercholesterolemia ranked by Dutch criteria.

From the mined data, we calculated DLCN scores for each patient and grouped them according to the criteria to assess the likelihood of the diagnosis of FH. We also calculated the mean lipid levels of the patients before the diagnosis and treatment.

3. Results

Medical records of 1,342,124 patients were included in our study: 44% of the records were retrieved from University of Debrecen Clinical Center and 56% of them were accessed from County Hospital of Szabolcs-Szatmár Bereg. First, we assigned patients into 9 separate groups as it is depicted in Table 1. Group 1 contains the number of patients with a diagnosis of FH, using mined textual history data. This group was really small and provided acceptable results only in Debrecen. Group 2 contains patients with a hypercholesterolemia diagnosis; groups 3,4,5 represent patients with CAD, cerebrovascular disease and PAD, respectively. Groups 6,7 are for those with tendinous xanthoma and corneal arcus diagnoses, respectively. We only used cases strongly supported by textual data to ensure likelihood of the diagnosis. Group 8 represents the set of those individuals with LDL-C levels above 3.4 mmol/L and triglyceride levels below 1.7 mmol/L (averages are calculated before statin treatment); while group 9 encompasses patients with total cholesterol levels above 5.2 mmol/L