Elsevier Editorial System(tm) for Atherosclerosis

Manuscript Draft

Manuscript Number: ATH-D-18-00393R1

Title: Identifying patients with familial hypercholesterolemia using data mining methods in the Northern Great Plain region of Hungary "FH Special issue"

Article Type: Research paper

Section/Category: Clinical & Population Research

Keywords: Familial hypercholesterolemia; screening; low-density lipoprotein; Dutch Lipid Clinic Network Criteria; data mining; deep learning

Corresponding Author: Professor György Paragh,

Corresponding Author's Institution: University of Debrecen

First Author: György Paragh

Order of Authors: György Paragh; Mariann Harangi; Zsolt Karányi; Bálint Daróczy; Ákos Németh; Péter Fülöp

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 on 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 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 that were taken 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 patents 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: Analyzing more regional and country-wide data and more frequent measurements of total cholesterol and LDL-C levels would increase the number of the discovered FH cases. Data mining seems to be ideal for filtering and screening for FH in Hungary.

Highlights

- There are not exact data about the frequency of familial hypercholesterolemia (FH) in Hungary.
- We aimed to identify patients with FH using data mining methods.
- Medical records of 1,342,124 patients were included.
- We calculated Dutch Lipid Clinic Network (DLCN) scores and lipid levels.
- We identified 11,937 patients with a DLCN score of 3-8.

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

György Paragh[1], Mariann Harangi[1], Zsolt Karányi[1], Bálint Daróczy[2], Ákos Németh[3], Péter Fülöp[1]

[1]Department of Internal Medicine, University of Debrecen Faculty of Medicine, Debrecen, Hungary

[2] Institute for Computer Science and Control, Hungarian Academy of Sciences (MTA SZTAKI), Budapest, Hungary

[3]Aesculab Medical Solutions, Black Horse Group Ltd. Debrecen, Hungary

*Corresponding author:

György Paragh

Department of Internal Medicine, University of Debrecen Faculty of Medicine

Nagyerdei krt. 98, H-4032 Debrecen, Hungary.

E-mail: paragh@belklinika.com

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.

Keywords: Familial hypercholesterolemia, screening, low-density lipoprotein, Dutch Lipid Clinic Network Criteria, data mining, deep learning

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 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 [9,10,11].

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.

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.

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 and triglyceride concentrations below 1.7 mmol/L triglyceride level (averages are calculated before statin treatment).

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 (**Table 2**). Our data indicate that 0.89% of the hospital patients might be affected by FH in the Northern Great Plain region of Hungary. We also assessed the prevalence of other CVD risk factors in the patients, including current smoking, hypertension, any type of diabetes mellitus, chronic

kidney disease and obesity (defined as body mass index over 30 kg/m^2), low HDL-C levels and hypothyroidism **(Table 3).** However, this data might not cover the full population, due to the potentially increased hospitalization in the elderly. Indeed, the number of the yearly medical cases is lower at younger ages, as it is 7.1 below 18 years (cases also include births); 6.99 between years 19-30; with an increasing number of 12.09 yearly cases between years 31-60; topping it with 19.96 hospital cases above 61 years. It has to be noted that the age distribution of the hospital patients is very similar to that of the regional population according to the 2011 Census in Hungary (data provided by the Hungarian Central Statistical Office) (**Fig.1**).

We were unable to assess data of family history and DNA analysis, but we assume that it would not have altered the number of patients with 3+ DLCN score significantly. Hungarian Central Statistical Office data indicates that 76% of the total population appears at least once at the hospital each year. In a previous study, patients fulfilling the strict criterion of clinical definite, probable or possible FH according to the Dutch criteria were offered molecular genetic analysis, but only 33% of patients have been identified as mutation carriers [15]. Thus one might suspect that 1 every 340 subjects might be affected with familial hypercholesterolemia in our region.

Discussion

We report the results of the first Hungarian FH screening project utilizing modern data mining methods, on the basis of medical and statistical records, at two major hospitals in the Northern Great Plain region of Hungary. 225 patients with probable diagnosis of FH, and 11,706 patients with possible diagnosis using the calculated DLCN scores. Although this study is not eligible to provide exact data on FH prevalence in Hungary, an estimated prevalence was calculated and found to be 1:340, which is in line with the prevalence data of

some other European countries [16]. Being one of the most frequent monogenic disorders, familial hypercholesterolemia represents a major challenge in cardiovascular disease prevention. Acknowledging its significance, the United States (US) Make Early Diagnosis to Prevent Early Deaths (MEDPED) diagnostic criteria were elaborated [17]. Hungary also joined this program, however, the general opinion was that diagnosing FH would not significantly alter its treatment due to the lack of effective therapeutic tools at the time. Some previous studies drew the attention on the fact that the prevalence of FH might be higher in Europe, compared to the previous estimates [18, 19]. Besides these results, newly developed drugs have also contributed to boost screening efforts and more effective therapies in FH. Indeed, the discovery of PCSK9 protein and its role in lipid metabolism provided a new treatment target. Subsequently, PCSK9 inhibitors were widely found to be effective in lowering cholesterol levels and to promote increasing efforts to identify FH patients who would benefit from more effective lipid lowering [20].

Hungary has become a member of the EAS-FH Studies Collaboration (FHSC) [21] and takes part in the ScreenPro FH program, as well [22]. Aiming at improving FH awareness in Hungary, we have initiated a nationally coordinated FH screening program with the help of regional and national lipid centers in Hungary. The leaders of the regional centers are responsible for providing extensive information about FH to the lay population and for educating medical staff in the region, including nurses, assistants and physicians. National coordinators oversee the efforts of regional centers, provide effective media appearance and focus on building international relations. In the frame of this program, we joined the FH Week 2017 in Hungary and participated in several radio, television as well as print and online appearances.

Despite the efforts, we realized that registering patients was slow; therefore, to identify more FH patients in our region, we analyzed the cases of two leading medical centers

and estimated the number of patients with a potential diagnosis of FH. Sampling was not representative, although it covered a large number of individuals (about three fourths of the population). Obesity, diabetes mellitus and low HDL-C levels tended to be more frequent among patients with DLCN scores above 8, while other CVD risk factors as smoking, hypertension chronic kidney disease were found to be less prevalent in these patients. We also tried to assess the prevalence of familial hypercholesterolemia with the help of data provided by the national statistical office. Although there is no data about the exact number of the FH prevalence in Hungary, we suspected to find more patients with a potential diagnosis of FH analyzing these cases.

Weaknesses of the study may answer, at least in part, these results. Although it is suggested to adjust LDL-C levels to calculate DLCN scores in patients receiving lipid lowering treatment [18], we were unaware of whether general practitioners (GPs) had initiated therapy or not before patient referral; therefore, we calculated the mean lipid levels that were taken prior to the diagnosis and treatment proposed in the medical centers. Recent reports indicate that GPs are able to accurately identify FH patients using DLCN scores [23,24]; though, as a result of the regulations of the Hungarian national health system, financing issues and the general lack of LDL-C measurements in the community laboratories, general practitioners are disposed to refer hyperlipidemic patients early to the hospital, also to organize risk stratification.

We were unable to assess data of family history and genetic data, moreover, it has to be mentioned that not 100% of the population goes to hospital each year. In addition, hospital goers tended to be older and those who visited a hospital tended to be checked more frequently. On the contrary, younger patients usually had less thorough laboratory examinations and their history was asked less frequently. These tendencies mean that identifying FH patients is biased towards the elderly.

On the other hand, frequent hospital visits in our region can be considered as a strength of our study as it might help increase the chance of finding FH patients. Analyzing more regional and country-wide data and a higher prevalence of total cholesterol and LDL-C measurements would increase the number of discovered FH cases. Additionally, better interoperability of electronic health records are recommended, since data comparability is of major interest to effectively utilize software applications to identify potential FH patients [25]. Family history and genetic data, as structured data elements in health records as well as mapping family networks with the consent of the hyperlipidemic patients and FH probands, would improve disease awareness and detection.

Data mining seems to be ideal for filtering and screening both single and mass cases, though valid diagnoses of FH require thorough medical workup. As a next step, we would like to apply modern machine learning (Gradient Boosting Liu [26], Support Vector Machines [27] or recurrent and sequential Artificial Neural Networks [28]) methods to better understand the connections between expressions while predicting risk of FH based on medical time-series.

For the treatment, the 6th Hungarian Cardiovascular Consensus Conference stratified FH into the very-high risk category with a target LDL-C level of 1.8 mmol/L. The latest, 7th Hungarian Cardiovascular Consensus Conference kept FH as an option in the very-high risk category since the 2016 EAS/ESC Guideline [29] considered FH without CVD as only high-risk. FH is treated with statins in the first line, while statin + ezetimibe combination and LDL apheresis are for those not reaching LDL-C target levels. To date, PCSK9 inhibitors are not subsidized in Hungary, therefore, their wider availability is ponderous.

Our data, though, might help increase FH screening efforts in our region, thus improving the therapeutic opportunities and life expectancy of our patients. Local activities including better cooperation between GPs, laboratories and medical centers [30], uniform and comparable electronic health records, as well as wider regional and international cooperation,

such as EAS-FHSC [21] and ScreenPro FH [22] could contribute to these efforts. Effective screening and early treatment of FH might also improve the miserable cardiovascular mortality data in Hungary.

Conflict of interest

The authors declared they do not have anything to disclose regarding conflict of interest with respect to this manuscript.

Financial support

This research was supported by the GINOP-2.3.2-15-2016-00005 project. The project is co-financed by the European Union under the European Regional Development Fund.

Author contributions

Study design: G. Paragh, Z. Karányi.

Development of methodology: Á. Németh, B. Daróczy.

Collection of data: Á. Németh, B. Daróczy.

Analysis and/or interpretation of data: Z. Karányi, M. Harangi, P. Fülöp.

Writing (not revising) all or sections of the manuscript: P. Fülöp, M. Harangi, G. Paragh.

Manuscript review: G. Paragh.

Table 1. Number and lipid parameters of patients assigned to groups by age, clinical and laboratory parameters

		Age (years)					Lipid parameters		
	Grouping	**+18**	**19-30**	**♀ 31-60 ♂ 31-55**	**♀ 61+ ♂ 56+**	**Total**	**Total cholesterol (mmol/L)**	**LDL-C (mmol/L)**	**HDL-C (mmol/L)**
Group 1	With a diagnosis of FH, history data	11	10	69	45	135	7.90 ± 1.66	5.56 ± 1.55	1.41 ± 0.29
Group 2	Diagnosis of hypercholesterolemia	80	357	13428	13552	27387	6.69 ± 1.61	4.71 ± 1.37	1.44 ± 0.44
Group 3	Coronary artery disease	67	363	10245	16883	27558	5.47 ± 1.55	3.48 ± 1.29	1.26 ± 0.40
Group 4	Cerebrovascular disease	167	608	15779	33416	49970	6.27 ± 1.61	4.70 ± 1.28	1.42 ± 0.43
Group 5	Peripheral arterial disease	23	109	13373	48450	61955	6.39 ± 1.82	4.91 ± 1.26	1.45 ± 0.43
Group 6	Tendinous xanthoma	1	2	5	7	15	5,26 ± 0,90	3,15 ± 0,63	1,47 ± 0,52
Group 7	Corneal arcus	3	8	35	34	80	5,69 ± 1,75	3,81 ± 1,54	1,45 ± 0,25
Group 8	LDL-C > 3.4 mmol/L and triglyceride < 1.7 mmol/L	71	282	6563	4559	11475	7.67 ± 1.01	5.45 ± 0.77	1.60 ± 0.42
Group 9	Cholesterol levels >5.2 mmol/L and triglyceride < 1.7 mmol/L	107	492	9441	6918	16958	7.90 ± 1.02	5.47 ± 0.91	1.68 ± 0.46

Table 2. Lipid parameters of patients categorized according to the Dutch Lipid Clinic Network criteria

	DLCN score	**number of patients**	**Ratio (%)**	**Total cholesterol (mmol/L)**	**LDL-C (mmol/L)**	**HDL-C (mmol/L)**
Definite FH	>8	6	0.001	10.3 ± 0.7	8.1 ± 0.6	2.42 ± 0.4
Probable FH	6-8	225	0.017	9.38 ± 3.0	7.61 ± 2.4	1.54 ± 0.6
Possible FH	3-5	11706	0.87	7.34 ± 1.2	5.26 ± 0.8	1.58 ± 0.4
Total	**>3**	**11937**	**0.89**			

Table 3. Prevalence of cardovascular risk factors in patients categorized according to the Dutch Lipid Clinic Network criteria

	DLCN score	Number of patients	Ratio (%)	Current smoker (%)	Hypertension (%)	Diabetes mellitus (%)	Chronic kidney disease (%)	Obesity (%)	Low HDL-C level (%)	Hypothyroidism (%)
Definite FH	>8	6	0,001	16.9	65.5	23.2	21.4	8.9	25.7	8.9
Probable FH	6-8	225	0,017	18.5	71.3	21.8	30.6	5.6	19.0	10.5
Possible FH	3-5	11706	0.87	16.2	70.6	21.9	17.7	7.1	16.3	9.4

Figure legend

Figure 1. Demographics of hospital visitors *vs*. regional population in the Northern Great Plain region of Hungary.

References

[1] Blakely, T, Disney, G, Atkinson, J, et al., A Typology for Charting Socioeconomic Mortality Gradients: "Go Southwest", Epidemiology, 2017;28:594-603.

[2] Mackenbach, JP, Kulhánová, I, Menvielle, G, et al., Trends in inequalities in premature mortality: a study of 3.2 million deaths in 13 European countries, J Epidemiol Community Health, 2015;69:207-217; discussion 205-206.

[3] Fulcher, J, O'Connell, R, Voysey, M, et al., Efficacy and safety of LDL-lowering therapy among men and women: meta-analysis of individual data from 174,000 participants in 27 randomised trials, Lancet, 2015;385:1397-1405.

[4] Hartgers, ML, Ray, KK and Hovingh, GK, New Approaches in Detection and Treatment of Familial Hypercholesterolemia, Curr Cardiol Rep, 2015;17:109.

[5] Goldstein, JL, Schrott, HG, Hazzard, WR, et al., Hyperlipidemia in coronary heart disease. II. Genetic analysis of lipid levels in 176 families and delineation of a new inherited disorder, combined hyperlipidemia, J Clin Invest, 1973;52:1544-1568.

[6] Sjouke, B, Kusters, DM, Kindt, I, et al., Homozygous autosomal dominant hypercholesterolaemia in the Netherlands: prevalence, genotype-phenotype relationship, and clinical outcome, Eur Heart J, 2015;36:560-565.

[7] Akioyamen, LE, Genest, J, Shan, SD, et al., Estimating the prevalence of heterozygous familial hypercholesterolaemia: a systematic review and meta-analysis, BMJ Open, 2017;7:e016461.

[8] Mundal, L, Sarancic, M, Ose, L, et al., Mortality among patients with familial hypercholesterolemia: a registry-based study in Norway, 1992-2010, J Am Heart Assoc, 2014;3:e001236.

[9] McCrindle, BW and Gidding, SS, What Should Be the Screening Strategy for Familial Hypercholesterolemia?, N Engl J Med, 2016;375:1685-1686.

[10] Hovingh, GK and Kastelein, JJ, Diagnosis and Management of Individuals With Heterozygous Familial Hypercholesterolemia: Too Late and Too Little, Circulation, 2016;134:710-712.

[11] Pérez de Isla, L, Saltijeral Cerezo, A and Mata, P, Response by Pérez de Isla et al to Letter Regarding Article, "Predicting Cardiovascular Events in Familial Hypercholesterolemia: The SAFEHEART Registry (Spanish Familial Hypercholesterolemia Cohort Study)", Circulation, 2017;136:1984.

[12] Austin, MA, Hutter, CM, Zimmern, RL, et al., Genetic causes of monogenic heterozygous familial hypercholesterolemia: a HuGE prevalence review, Am J Epidemiol, 2004;160:407-420.

[13] Johns, BT and Jamieson, RK, A Large-Scale Analysis of Variance in Written Language, Cogn Sci, 2018.

[14] Larrañaga, P, Calvo, B, Santana, R, et al., Machine learning in bioinformatics, Brief Bioinform, 2006;7:86-112.

[15] Damgaard, D, Larsen, ML, Nissen, PH, et al., The relationship of molecular genetic to clinical diagnosis of familial hypercholesterolemia in a Danish population, Atherosclerosis, 2005;180:155-160.

[16] Bell, DA and Watts, GF, Progress in the care of familial hypercholesterolaemia: 2016, Med J Aust, 2016;205:232-236.

[17] Williams, RR, Hunt, SC, Schumacher, MC, et al., Diagnosing heterozygous familial hypercholesterolemia using new practical criteria validated by molecular genetics, Am J Cardiol, 1993;72:171-176.

[18] Benn, M, Watts, GF, Tybjaerg-Hansen, A, et al., Familial hypercholesterolemia in the danish general population: prevalence, coronary artery disease, and cholesterol-lowering medication, J Clin Endocrinol Metab, 2012;97:3956-3964.

[19] Benn, M, Watts, GF, Tybjærg-Hansen, A, et al., Mutations causative of familial hypercholesterolaemia: screening of 98 098 individuals from the Copenhagen General Population Study estimated a prevalence of 1 in 217, Eur Heart J, 2016;37:1384-1394.

[20] Arca, M, Old challenges and new opportunities in the clinical management of heterozygous familial hypercholesterolemia (HeFH): The promises of PCSK9 inhibitors, Atherosclerosis, 2017;256:134-145.

[21] Vallejo-Vaz, AJ, Akram, A, Kondapally Seshasai, SR, et al., Pooling and expanding registries of familial hypercholesterolaemia to assess gaps in care and improve disease management and outcomes: Rationale and design of the global EAS Familial Hypercholesterolaemia Studies Collaboration, Atheroscler Suppl, 2016;22:1-32.

[22] Ceska, R, Freiberger, T, Vaclova, M, et al., ScreenPro FH: from the Czech MedPed to international collaboration. ScreenPro FH is a participating project of the EAS-FHCS, Physiol Res, 2017;66:S85-S90.

[23] Bell, DA, Kirke, AB, Barbour, R, et al., Can patients be accurately assessed for familial hypercholesterolaemia in primary care?, Heart Lung Circ, 2014;23:1153-1157.

24] Kwok, S, Pang, J, Adam, S, et al., An online questionnaire survey of UK general practitioners' knowledge and management of familial hypercholesterolaemia, BMJ Open, 2016;6:e012691.

[25] Safarova, MS and Kullo, IJ, Lessening the Burden of Familial Hypercholesterolemia Using Health Information Technology, Circ Res, 2018;122:26-27.

[26] Liu, Y, Li, B, Tan, R, et al., A gradient-boosting approach for filtering de novo mutations in parent-offspring trios, Bioinformatics, 2014;30:1830-1836.

[27] Perfetti, R and Ricci, E, Analog neural network for support vector machine learning, IEEE Trans Neural Netw, 2006;17:1085-1091.

[28] LeCun, Y, Bengio, Y and Hinton, G, Deep learning, Nature, 2015;521:436-444.

[29] Catapano, AL, Graham, I, De Backer, G, et al., 2016 ESC/EAS Guidelines for the Management of Dyslipidaemias, Rev Esp Cardiol (Engl Ed), 2017;70:115.

[30] Kirke, AB, Barbour, RA, Burrows, S, et al., Systematic detection of familial hypercholesterolaemia in primary health care: a community based prospective study of three methods, Heart Lung Circ, 2015;24:250-256.

Dr. Arnold von Eckardstein
Editor-in-Chief, Atherosclerosis

Dr. Gerald F. Watts
Co-Editor, Atherosclerosis

Dear Sirs,

We have received the editorial response and the reviewers' comments and questions regarding our manuscript titled "Identifying patients with familial hypercholesterolemia using data mining methods in the Northern Great Plain region of Hungary" (ATH-D-18-00393). Thank you for the opportunity to submit our revised manuscript to your journal.

Below, please find our answers to the questions and comments. As it was requested, changes in the revised manuscript are highlighted in red. (We also submit the final corrected version of the manuscript.)

Reviewer #1:

We thank you for the thorough review and for the comments aimed to improving our manuscript. In the followings, we respond to your suggestions one by one.

Although there are limitations in data accumulation that have been adressed in the text, this paper gives nonconfirmatory information about FH prevalance in Hungary.
1-The following sentence needs to be moved to methods section: We initiated a cooperation with a clinical site partner company (Black Horse Group Ltd.) to utilize their medical system framework named "AescuLab" (www.aesculab.net).

As requested, we rephrased the above mentioned sentence and moved it into the methods section (page 5):
"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)."

2-The following sentence implies that the EAS/ESC guideline categorises FH as very high risk whereas FH without ASVD is only high risk in the guideline,this should be corrected: The latest,

7thHungarian Cardiovascular Consensus Conference kept FH as an option in the very-high risk category according to the 2016 EAS/ESC Guideline

Thank you for the comment, we corrected the following sentence (page 11) to:
"The latest, 7th Hungarian Cardiovascular Consensus Conference kept FH as an option in the very-high risk category since the 2016 EAS/ESC Guideline [29] considered FH without CVD as only high-risk."

Thanking for your comments, we do hope that our modifications will meet your expectations.

Reviewer #2:

This is an excellent example of combination of review and original article, with new, national/regional based very robust data (more than 1 300 000!). Authors clearly demonstrate the practice of examining large databases in order to generate new information on the example of FH. Paper is relatively short, condensed, the tables are easy to read and understand. Theoretically we could speculate about the difference between the region of Grate Plate and the capital of Hungary from FH prevalence point of view, but authors conclude sophistically conclude and comment their results.
In my mind this article can be published in the "Special FH issue", without revision.

Thanking you for your comments, we are honestly grateful for your recommendation.

Editors' comments:

This is the first attempt to define the frequency of FH in Hungary. The approach is based on interrogated hospital records and generates useful information. With these types of analysis the issue is false positives due to other conditions that cause high cholesterol. Can authors add information on diabetes, obesity and CKD; were secondary causes of high cholesterol such hypothyroidism also excluded ? What about other CV risk factors? Smoking, low HDL, hypertension ? Please try to include.

With our cordial thankfulness, we calculated the prevalence of the conditions mentioned above and added a sentence in the Results section (page 7) and a new table (Table 3) into the manuscript. We also addressed this issue in the Discussion (page 10).

Sentence in Results: ”We also assessed the prevalence of other CVD risk factors in the patients, including current smoking, hypertension, any type of diabetes mellitus, chronic kidney disease and obesity (defined as body mass index over 30 kg/m^2), low HDL-C levels and hypothyroidism ***(Table 3).”***

Table:

”Table 3. Prevalence of cardovascular risk factors in the patients categorized according to Dutch Lipid Clinic Network criteria”

	DLCN score	**number of patients**	**ratio (%)**	**current smoker (%)**	**hypertension (%)**	**diabetes mellitus (%)**	**chronic kidney disease (%)**	**obesity (%)**	**low HDL-C level (%)**	**hypothyroidism (%)**
Definite FH	>8	6	0,001	16.9	65.5	23.2	21.4	8.9	25.7	8.9
Probable FH	6-8	225	0,017	18.5	71.3	21.8	30.6	5.6	19.0	10.5
Possible FH	3-5	11706	0.87	16.2	70.6	21.9	17.7	7.1	16.3	9.4

Sentence in Discussion: ”Obesity, diabetes mellitus and low HDL-C levels tended to be more frequent among patients with DLCN scores above 8, while other CVD risk factors as smoking, hypertension chronic kidney disease were found to be less prevalent in these patients.”

The authors indicate that an adjustment of treated lipids was made to assess DLCNS; was this for LDL-C; if so, how was this estimated ? refs required.

Since we had no information about the general practitioners’ efforts (or the lack of them) about indicating lipid lowering treatment, we calculated the means of lipid parameters that were taken before the diagnosis and treatment that were proposed in the hospitals. Generally, Hungarian general practitioners tend to refer hyperlipidemic patients early to lipid centers. Considering that the

patients usually don't have wait more than 3-4 weeks, we believe that lipid lowering treatment initialized by the family doctors wouldn't have modified our data significantly. We addressed this issue as the followings (page 10):

"Although it is suggested to adjust LDL-C levels to calculate DLCN scores in patients receiving lipid lowering treatment [18], we were unaware of whether general practitioners (GPs) had initiated therapy or not before patient referral; therefore, we calculated the mean lipid levels that were taken prior to the diagnosis and treatment proposed in the medical centers."

2 others points:

1. The deficiency that primary care was not studied needs pointing out and references made to relevant studies from UK and Australia, both published on Heart. Most FH is in community; what is happening in Hungary?

Since this important issue closely relates to the subject that was addressed above, we continued our text in the manuscript as it follows (page 10). We also added new references to highlight the importance of the topic (refs 23, 24).

"Recent reports indicated that GPs were able to accurately identify FH patients using DLCN scores [23,24]; though, resulting from the regulations of the Hungarian national health system, financing issues and the general lack of LDL-C measurements in the community laboratories, general practitioners are disposed to refer hyperlipidemic patients early to the hospital, also for organizing risk stratification.

2. Safarova has published a good article in Circulation Research 2018 on use of health information technology and how this approach aligns with that used by authors in detecting FH is apposite and should be referenced and noted in Discussion.

This is also a very important comment, since databases of electronic health records are very hard to compare and making them comparable is extremely time consuming. Thanking for your suggestion, we referenced this paper in the manuscript (ref 25) and added the following section in the Discussion (page 11):

"Additionally, better interoperability of electronic health records would be recommended, since data comparability is of major interest to effectively utilize software applications in identifying potential FH patients [25]. Family history and genetic data as structured data elements in health

records as well as mapping family networks with the consent of the hyperlipidemic patients and FH probands would improve disease awareness and detection."

To briefly summarize these issues we added the followings into the penultimate sentence of the manuscript (page 11) with the help of a new reference (ref 30):
"Local activities including better cooperation between GPs, laboratories and medical centers [30], uniform and comparable electronic health records, as well as wider regional and international cooperation, such as EAS-FHSC [21] and ScreenPro FH [22] could contribute to those efforts."

The new references of the manuscript are Refs 23, 24, 25, and 30, respectively:

[23] Bell, DA, Kirke, AB, Barbour, R, et al., Can patients be accurately assessed for familial hypercholesterolaemia in primary care?, Heart Lung Circ, 2014;23:1153-1157.
24] Kwok, S, Pang, J, Adam, S, et al., An online questionnaire survey of UK general practitioners' knowledge and management of familial hypercholesterolaemia, BMJ Open, 2016;6:e012691.
[25] Safarova, MS and Kullo, IJ, Lessening the Burden of Familial Hypercholesterolemia Using Health Information Technology, Circ Res, 2018;122:26-27.
[30] Kirke, AB, Barbour, RA, Burrows, S, et al., Systematic detection of familial hypercholesterolaemia in primary health care: a community based prospective study of three methods, Heart Lung Circ, 2015;24:250-256.

We also corrected some typos.

In sum, we would like to express our gratitude to the reviewers and to the editor for their insightful comments to improve the quality of our manuscript. We hope our revised manuscript will fulfil the requirements of the journal and will be worth publishing.

Sincerely,
György Paragh

Statement of originality

All authors have seen and approved the final version of the manuscript being submitted. The article is the authors' original work, hasn't received prior publication and isn't under consideration for publication elsewhere.

Prof. György Paragh, MD, DSc
Department of Internal Medicine, University of Debrecen Faculty of Medicine
Address: Nagyerdei krt. 98, H-4032 Debrecen, Hungary.
Tel/Fax: + 36 52 442101
E-mail: paragh@belklinika.com

Conflict of Interest

There's no financial/personal interest or belief that could affect our objectivity.

All authors declare no conflict of interest.

Prof. György Paragh, MD, DSc

Department of Internal Medicine, University of Debrecen Faculty of Medicine

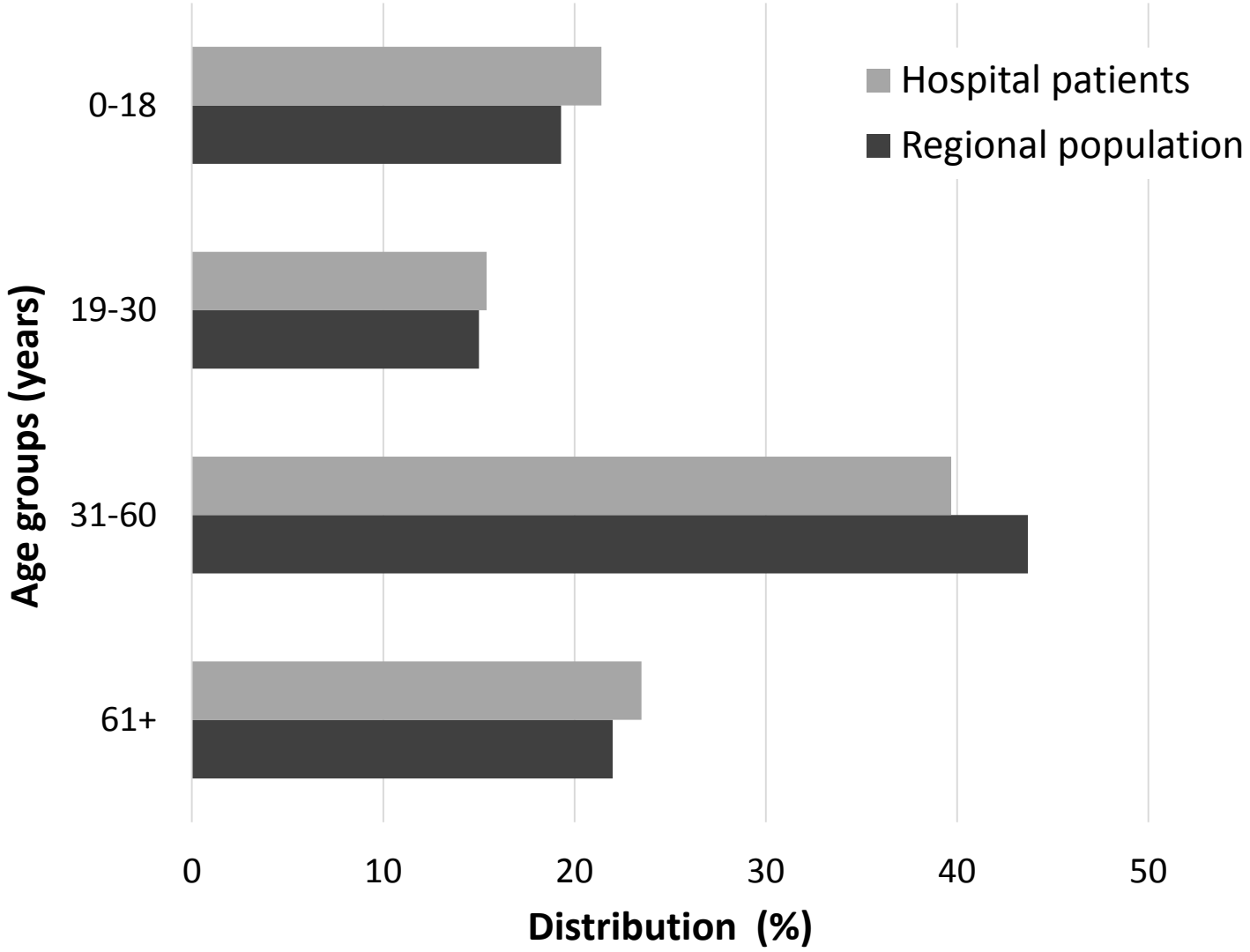
Address: Nagyerdei krt. 98, H-4032 Debrecen, Hungary.

Tel/Fax: + 36 52 442101

E-mail: paragh@belklinika.com

Figure(s)

Fig.1.



Atherosclerosis style guide checklist

Atherosclerosis applies format guidelines to all accepted papers, with the aim of improving their readability.

Manuscripts that do not conform to the format guidelines of the *Atherosclerosis* Journal will be returned to the authors for reformatting.

Please find below a questionnaire to guide authors to comply with the formatting requirements for revised submissions. For more detailed information, visit our website.

Please note that when you answer "No" to a question, editing of your manuscript is required before submission to *Atherosclerosis*.

Manuscript structure and style

Does your manuscript contain all the below essential elements, in this order? **Yes** **No**
(please stick to the headers as indicated below)

- Title
- Authors, Affiliations, Contact Information
- Abstract in the Atherosclerosis format (*Background and aims*, *Methods*, *Results*, *Conclusions*)
- Introduction
- Materials and methods (or Patients and methods)
- Results
- Discussion
- Conflict of interest (mandatory)
- Financial support (if applicable)
- Author contributions (mandatory)
- Acknowledgements (if applicable)
- References
- Figures and Tables (with legends in the suitable style)

Abstract style

Is the Abstract structured in the below sections? **Yes** **No**

- *Background and aims*
- *Methods*
- *Results*
- *Conclusions*

Figure and table legends

Are figure and table legends formatted as described below? **Yes** **No**

Each figure and table legend should have a brief overarching title that describes the entire figure without citing specific panels, followed by a description of each panel, and all symbols used.

If a figure or table contains multiple panels, the letter describing each panel should be capitalized and surrounded by parenthesis: i.e. (A)(B)(C)(D).

Please make sure to apply the formatting requirements to figures and tables where necessary (e.g. style of *p* values, gene and protein nomenclature).

Footnotes to tables

Are footnotes to tables formatted as described below? **Yes** **No**

Footnotes to tables should be listed with superscript lowercase letters, beginning with "[a]."
Footnotes must not be listed with numbers or symbols.

Abbreviations

Are abbreviations defined when first used in the text? **Yes** **No**

Use of abbreviations should be kept at a minimum.

Units

Are units expressed following the international system of units (SI)? **Yes** **No**

If other units are mentioned, please provide conversion factors into SI units.

DNA and protein sequences

Are gene names italicized? **Yes** **No**

Gene names should be italicized; protein products of the loci are not italicized.

For murine models, the gene and protein names are lowercase except for the first letter.
(e.g., gene: *Abcb4;* protein: Abcb4)

For humans, the whole gene name is capitalized.
(e.g., gene: *ABCB4;* protein ABCB4)

Mouse strains and cell lines

Are knock-out or transgenic mouse strains and cell lines italicized and the symbol superscripted? Yes No

(e.g. *ob/ob* , $p53^{+/+}$, $p53^{-/-}$)

p values

Are *p* values consistently formatted according to the below style throughout the manuscript (including figures and tables)? Yes **No**

p <X
p >X
p=X

Language

Is your manuscript written in good English? **Yes** **No**

Please make sure that you consistently use either American or British English, but not a mixture of them.

Please make sure that words are written consistently in the same way throughout the manuscript.
e.g. non-significant or nonsignificant
e.g. down-regulation or downregulation

Artwork

Have you submitted high-resolution versions of your original artwork? **Yes** **No**

Please make sure to use uniform lettering and sizing in your original artwork, including letters to indicate panels, consistently throughout all figures.