

SHORT THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PhD)

**Diagnosing rare metabolic diseases using data mining and
machine learning**

by Ákos Németh

Supervisor: Péter Fülöp, PhD



UNIVERSITY OF DEBRECEN
DOCTORAL SCHOOL OF HEALTH SCIENCES

DEBRECEN, 2024

Diagnosing rare metabolic diseases using data mining and machine learning

By Ákos Németh, economist, chemical engineer

Supervisor: Péter Fülöp, MD, PhD

Doctoral School of Health Sciences., University of Debrecen

Head of the **Defense Committee**: Margit Balázs, PhD, DSc

Reviewers: András Benczúr, PhD
Rudolf Kolozsvári, PhD

Members of the Defense Committee: László Márk, PhD
Szilvia Fiatal, PhD
András Benczúr, PhD
Rudolf Kolozsvári, PhD

The PhD Defense takes place at the Lecture Hall of Bldg. “A”, Department of Internal Medicine, Faculty of Medicine, University of Debrecen, at 13:00 on 5th June, 2024

Introduction

According to health statistics, cardiovascular diseases (CVD) are the leading cause of death in Europe, while premature death from cardiovascular causes is three times higher in Central Europe than in Western Europe. Hyperlipidemia is a major risk factor for CVD, with high cholesterol and triglyceride levels contributing to the development of atherosclerosis even in young age. Some congenital diseases, including familial hypercholesterolemia (FH) and familial chylomicronemia syndrome (FCS), affects lipid metabolism leads to markedly elevated lipid levels, and carry a significant cardiovascular risk.

In this dissertation, I focus on congenital diseases of lipid metabolism, in particular FH and FCS. Both diseases are also of major importance as, in the absence of a proper diagnosis, the effectiveness of lipid-lowering therapy is markedly reduced in these already difficult-to-treat conditions.

FH is one of the most common monogenic diseases, caused mostly by loss-of-function mutations in the genes of low-density lipoprotein (LDL) receptor (LDLR) or apolipoprotein (Apo) B100 (Apo-B100), or by gain-of-function mutations in the gene of the proprotein convertase subtilisin/kexin type 9 (PCSK9). Previous literature data suggest that the heterozygous form of the disease develops at a frequency of approximately 1:500, whereas the homozygous form develops at a frequency of 1:1,000,000. Data from recent years imply that the real prevalence of the disease is somewhat higher than this, with the heterozygous form having a prevalence of about 1 in 300 and the homozygous form about 1 in 300,000. This means there could be 20-30 million FH patients worldwide, but at least 80% of them are undiagnosed. It is also worth noting that only 10% of known patients receive effective lipid-lowering treatment and reach cardiovascular risk targets. Thus, it is not surprising that FH patients die from cardiovascular complications some 15 years earlier compared to those without the gene defect.

Characterized by fasting chylomicronemia, FCS is also a rare disease with monogenic inheritance, in which the elimination of triglyceride-rich lipoproteins from the circulation is impaired due to reduced lipoprotein lipase (LPL) activity. This condition leads to severely high and persistent serum triglyceride (TG) concentrations, which can lead to a severe, potentially life-threatening condition with recurrent pancreatitis. According to previous estimates, FCS affects one in a million people, and in all cases, it is associated with a genetic variation that significantly impairs the function of the LPL protein, either in the form of homozygous or compound heterozygous mutations. Although loss-of-function homozygous mutations in the LPL gene can be detected in the vast majority of FCS patients, complex

mutations in Apo-C2, Apo-A5, lipoprotein lipase maturation factor 1 (LMF1), glycosyl-phosphatidylinositol-linked HDL-binding protein (GPIHBP1) and also in the glyceraldehyde-3-phosphate dehydrogenase (G3PDH1) genes are also common.

Rare diseases such as FH and FCS are generally underestimated in terms of their importance and prevalence. Their diagnosis is difficult and time-consuming, due to the complexity and rarity of the pathologies, and requires specialized medical knowledge. In respect to the above mentioned, our aim was to use computer-assisted diagnosis and therapy to estimate the exact prevalence of these diseases in our region and to provide physicians an effective diagnostic toolkit that can help them to accurately identify the diseases, also the causes and risk factors underlying the dynamics of these conditions. We also aimed to mark factors that might influence therapy without the need of specific expertise.

Exogenous and endogenous lipid metabolism

Cholesterol and triglycerides are physiologically very important plasma lipids. With a wide range of functions, cholesterol is a component of the cell membrane, and also serves as a regulator of co-localisation of signaling proteins and a precursor of the synthesis of certain hormones and bile acids. Triglycerides exist as the main source of energy for the body and, similarly to cholesterol, are the other main lipid component of lipoproteins circulating in the blood. The main lipoprotein fractions, such as chylomicron (CM), very-low density lipoprotein (VLDL), low density lipoprotein (LDL) and high density lipoprotein (HDL), differ in both their physical and chemical properties and thus in their function.

Lipid metabolism can be divided into two main processes, named exogenous and endogenous lipid metabolism. During the exogenous cycle, dietary fat is converted into chylomicrons in the epithelial cells of the small intestine following digestion. In the circulation, chylomicrons lose glycerol and fatty acids due to the hydrolysis mediated by the vascular endothelium-bound LPL. The remaining chylomicron-remnant particles are taken up by the liver, which uses the lipids for energy storage or for the synthesis of bile acids and other lipid derivatives. At the beginning of the endogenous cycle, lipids synthesized in the liver leave the hepatocytes in the form of VLDL. In the circulation, VLDL also loses fatty acids due to the activity of LPL, thereby reducing its TG content and transforming into IDL and then LDL. Cholesterol transported by LDL is a source of various cellular processes for the peripheral tissues, however, is also a key molecule in atherosclerosis, as it can initiate and maintain the process of atherosclerosis undergoing oxidization and subsequently

accumulating in the vascular wall. In contrast, HDL is the key protective molecule of atherosclerosis, participating in the so-called reverse cholesterol transport and removing cholesterol from the vessel wall and from the peripheral tissues and returning it to the liver. HDL also carries important antioxidant enzymes on its surface, contributing to its anti-atherosclerotic capacity.

Dyslipidemias

Etiologically, dyslipidemias can be grouped into primary (determined by some genetic variations, often showing familial accumulation) and secondary (i.e. associated with other diseases or medications) causes. Based on the serum lipid parameters, hypercholesterolemia, hypertriglyceridemia and combined hyperlipidemia may exist.

Due to the early development of atherosclerosis and its serious complications, childhood FH screening is of paramount importance to provide treatment against atherosclerotic cardiovascular diseases (ASCVD). In such cases, either general screening and blood sampling or family history can be used to identify the potential patients. Screening in childhood is also important in FCS, as acute pancreatitis may be a very serious and life-threatening complication, too. Data mining-assisted screening, which can be used to collect independent tests and calculate risk scores, has recently gained importance and may play an important role in supporting general medical screening and increasing its cost-effectiveness. The most common diagnostic laboratory tests for screening for the mentioned diseases are the measurement of circulating levels of serum cholesterol, TG, HDL-C and LDL-C.

Resulting in Fredrickson type IIA dyslipidemia, FH is characterized by a markedly increased cholesterol level with a significantly elevated LDL-C value. As cholesterol deposits in the tissues, 20-40% of the patients may also show some xanthomas of the Achilles tendon, xanthelasmas of the eyelids, or annular discoloration of the cornea termed arcus corneae. FH is diagnosed by the Dutch Lipid Clinic Network (DLCN) criteria and scoring system, which takes into account the patients' untreated LDL-C values, physical and genetic examination results, as well as their own clinical history and family history.

Resulting in Fredrickson type I dyslipidemia, FCS is characterized by an extreme hypertriglyceridemia with consequently elevated CM and VLDL levels. Eruptive xanthoma, lipemic plasma and lipemia retinalis are common, although not specific symptoms. High serum CM levels can lead to liver and spleen enlargement, however, the most serious

potential complication of FCS is the acute, sometimes recurrent pancreatitis with a mortality rate up to 5-6%. An FCS scoring system is also available to support diagnosis of the disease, which was proposed by Moulin *et al.* (FCS score).

In addition to the routine lipid parameters recommended above, newer recommendations also suggest more frequent measurements of apoB100, as its serum levels are proportional to VLDL, LDL and IDL and are a good indicator of cardiovascular risk. Serum lipoprotein(a) [Lp(a)] is also an independent risk factor for ASCVD but difficult to control with standard lipid-lowering therapy. Recent guidelines advise its measurement at least once to provide better estimation of CVD risk. Moreover, the relationship between Lp(a) and morbidity and mortality in FH patients is still not fully understood according to the literature.

Machine processing of text data

Text mining deals with textual data processing and includes the scientific results and techniques of computing and science, statistics, and linguistics. Its purpose is to process unstructured data and efficiently extract information using various techniques. During information retrieval (IR), we assign structured information to the raw, unorganized text. Important steps of the process are "tokenization", filtering and "stemming", which contribute to the pre-processing of unstructured texts. An important element of text mining is the use of the so-called "regexes" (i.e. regular expressions), which provide an opportunity to easily describe (and discover) rules and patterns. They can also be used on already pre-processed unstructured texts to search for simpler and more complex expressions.

Ranking documents and terms is one of the most important text mining tasks. TF-IDF (Term Frequency - Inverse Document Frequency) infers the importance and relevance of words and phrases by examining their frequency of occurrence in documents. The TF-IDF value of a term increases proportionally with the frequency of occurrence of the term in the categorized documents, but the value is compensated by the number of documents in which the term occurs (e.g. a frequently used conjunction in a text will receive a low TF-IDF value).

During machine learning, the most advanced and complex way of processing textual data is the machine learning-based methods of natural language processing (NLP), which aim at reproducing the written or spoken language structure and are therefore suitable for complex morphological, syntactic and semantic analysis of texts. Recently, neural networks are mostly used, but there are also statistical and rule-based approaches.

The discrepancy between estimated and actual probability of occurrence is the perplexity, which may serve as a performance metric for language models. That is, the lower the perplexity the more successful a language model is in understanding the structure of a given text. Modern neural network-based language models describe the text in terms of multidimensional vector spaces. In word2vec models, each word is interpreted as a vector, and the words are described in the vector space using a shallow, two-layer neural network (a process called word embedding). Word2vec models were surpassed by transformer-based models in the mid-2010's, which are currently the most widely used of the neural network-based language models. Transformer-based models take the position of the words within a sentence into account with respect to their relationship to other words ("position encoding" and "self-attention"); and their multi-layer structure (multi-layered but equally weighted encoder-decoder structures) allows parallel computation for each word, making them particularly efficient to learn on multi-threaded distributed cloud systems.

Binary classification

In binary classification, the object, i.e. the input sample, is classified into one of two categories (e.g. 0: not sick; 1: sick) using mathematical operations. In the process, a set of attributes (x) is assigned to a label (y) using a classification model. The input of the classification learning task is a finite list of n (x, y) "tuple", and the classification task is mostly descriptive or predictive. In mathematical-statistical terms, classification can in fact be regarded as a special case of regression calculation (where the "dependent" variable can only take discrete values). During classification, a model is built from input data using learning algorithms that can describe the relationship between (x) and (y) as accurately as possible (training set), and can also correctly categorize data that were not part of the model building (test set). The most important and best known composite performance metric of binary classification is the ROC-AUC curve. The ROC (Receiver Operating Characteristics) is a probability curve that shows the effectiveness of binary separation at different thresholds, while the AUC (Area Under The Curve) indicates the overall success of the separation. The two measures are often referred to collectively as AUROC (Area Under the Receiver Operating Characteristics).

A wide range of classification models ("classifiers"), i.e. classification algorithms, are available. Decision trees (classification or regression trees) are among the simpler classification models. Decision trees consist of nodes and branches. Among the nodes, the

end point where the classification takes place ("leaf nodes") and the starting point from which the tree starts ("root node" or "root") are of particular importance. The NN classification ("k-nearest neighbors" classifier) is the most significant "lazy learner". The "lazy learner" (in contrast to the previous "eager learner" models) means that the training data is not generalized before classification. This is especially useful for large and frequently changing data sets. During the NN classification, we examine the categories of the k cases located in the vicinity of an examined point x in a d dimensional feature space. The Bayes classification (Bayes classifier) is based on the Bayes theorem describing the relationship between conditional and linked probabilities. With the help of "*a priori*" independent probabilities determined on the learning data, the probability that an uncategorized case falls into a given category can be calculated based on the Bayes theorem. Bayesian classification is very efficient on any data where the initial probability distribution is well defined and multicollinearity is low.

Artificial neural networks (ANNs) are the most popular and fastest growing models of machine learning. An ANN consists of nodes ("neurons") and connections between the nodes (similar to neuron-axon-dendrite connections in the nervous system). Neurons are activated by functions that determine the type of ANN, which can be feed-forward, where neurons are connected only to neurons in the next layer, or recurrent, where neurons are connected to each other or to other neurons in several layers. During the training of an ANN, backpropagation is most often used and the function parameters are optimized according to the gradient descent. ANN is a general approximator, so it can estimate virtually any real function, but it is particularly sensitive to input noise and overfitting, and has a significant trainingtime. SVM (support vector machine, or support vector network) is perhaps the most popular of the shallow learning methods. A major advantage of SVM is that the learning problem is a convex optimization problem, unlike the heuristic approach of ANN or trees, so that the minimum found during the process is surely also the global minimum of the objective function. Finally, in ensemble or combined classification solutions, we exploit the fact that the probability of classification error is smaller for multiple, partially independent classification algorithms together than for a single algorithm.

Objectives

We set out to solve the following problems during our studies:

- to determine the regional prevalence of FH patients by data mining;
- to determine the regional prevalence of FCS patients by data mining;
- to assess cardiovascular risk factors in FH patients;
- to investigate serum Lp(a) levels in FH patients and compare it with the non-FH population using data mining methods;
- to improve the identification of FH patients, especially the "probable FH" group, using machine learning methods;
- to ensure rapid and automated screening of FCS patients, rapid estimation of the FCS score using data mining and machine learning methods;
- to investigate the possibility of efficient identification of FCS patients based on other routine laboratory tests using data mining and machine learning methods;
- to determine the potential of machine learning and data mining tools to support everyday medical work and clinical research.

Materials and methods

Study population

We used traditional data mining (DM) tools to understand the examined diseases, estimate their prevalence and support diagnostics. During data processing, we used structured and anonymized data that was extracted from clinical and hospital databases from two high-traffic healthcare institutions in the Northern Great Plain region. All the medical documents of nearly one and a half million ($n = 1,342,124$) patients who appeared in the University of Debrecen Clinical Center and the J6sa Andr6s University Teaching Hospital in Ny6iregyh6za between January 1, 2007 and December 31, 2014 were analyzed.

Data used

The initial step in screening patients was to consolidate, format and clean the clinical data, so that we could run the written scripts and algorithms efficiently. Basically, 5 types of clinical data were used in our research: textual data of the patient examination, numerical data of laboratory tests and other important parameters (e.g. reference ranges, units of measurement, data on the machines performing the tests), data on presumed and confirmed diagnoses, results of interventions and tests, and non healthcare related statistical data on patients (e.g. place and date of birth, sex, education, occupation).

Within clinical data sources, there are naturally many duplications and inconsistencies on patient level. Those can make data processing difficult and easy at the same time, but because of this, they must be handled properly during processing. For example, duplicate diagnoses are common and useful because it is easier to find a diagnosis and categorize a patient, but the existence of inconsistent diagnoses makes it difficult to see clearly and often requires more detailed analyzes and deeper explorations.

The cleaning and preprocessing processes were partly performed using open source tools (Pandas, Numpy) and partly using self-developed Python 2.8.X and Python 3.X scripts, which resulted large, coherent and indexed data files containing the diagnosis of each medical case, laboratory tests, descriptions of interventions and diagnostic procedures, and all textual data. The anonymous statistical data of individuals were stored in a separate table. During the pre-processing, various serialization and buffering methods were also used to avoid memory problems associated with the massive data set. These tasks were mainly solved using the python-based pandas, the awk language popular in the unix environment, and the dask framework. At the end of the preprocessing, the data was stored in PostgreSQL 13.x database tables.

Data cleaning with data mining

Additional processing steps are required to structure the textual data and extract the information contained in it. To find and group important terms, parsing, stemming, stop word filtering, BOW (bag-of-words) modeling and ranking of terms (TF-IDF, W2V) were performed. We also determined certain expressions typical of the diseases with the help of MDs, which were then searched for in the patients' medical histories using regexes. Using the above mentioned methods, we reduced the initial term space of several million terms to two hundred and fifty thousand, and also succeeded in defining the most important terms (such as smoking or harmful alcohol consumption) and discretizing them. These steps were essential to accurately determine the elements of each scoring system (DLNCS, FCS score) affecting anamnesis and family medical history. At the same time, the pre-processing did not end here, since in order to more efficiently find the elements hidden in the text and to be able to efficiently compare the cases based on the text, we needed further language processing via known language models as well as language models trained by ourselves. For language processing, we used the NLTK (Natural Language Toolkit 3.x) Python library. During basic text preprocessing, the most important word embeddings, unigrams, bigrams and n-grams became available. Since these embeddings were only trained in Hungarian using a classical set of texts (corpus), we had to train our own language models. For this, we developed a GRU (Gated Recurrent Unit)-based model and for general terms, we used BERT (Bidirectional Encoder Representations from Transformers). After data cleaning, the set of clinical data ("dataset") was organized into a data matrix. The data matrix contained the most important attributes at the patient (case) level, as well as continuous (e.g. laboratory values), discrete (e.g. hypertension: yes/no) and ordinal (e.g. order of diagnoses) values for all the needed attributes.

Screening of patients for FH and FCS diagnosis

Before the DLNCS score (DLNCS) was determined, the mean lipid values of the patients were calculated, excluding all cases where lipid-lowering therapy was already known based on the language processing process detailed above. Then, based on the set of matched medical cases, we calculated the DLNCS score for each patient on the already cleaned and pre-processed data, thereby estimating the probability of FH. The FCS score was calculated similarly on the cleaned and pre-processed data. The data thus calculated and grouped according to the probability of FCS was subjected to expert assessment and then further

processed and trained using machine learning (ReLU neural network, adaptive boosting, gradient boosting, SVM) after clinical diagnosis.

Machine learning

The identification of FCS cases can be considered as a binary classification problem, since a finite descriptive vector of real numbers (lab tests, history, diagnostics, etc.) and a binary label (0: not an FCS patient; 1: an FCS patient) can be obtained for each individual. In order to test all machine learning methods and select the best performing one, we trained subsamples of the entire population with data from known and diagnosed FCS patients, and then evaluated the mathematical models trained on these subsamples on independent datasets with also known and diagnosed FCS patients using known performance metrics (ROC-AUC, ACC, SENS), with a 50-50% split of the data into training and test sets. In line with the Vapnik-Chervonenkis generalization theory, we used three, widely applicable mathematical models, namely ensemble trees (tree ensembles), "shallow" kernel-based (SVM) and "deep" ReLU activation function neural networks. Since no specific conditions on structure are met for FCS diagnoses, we used fully connected neural networks with ReLU activation function and adaptive momentum (Adam) optimization.

FCS diagnoses are small datasets where the order of the data matrix properties is not deterministic, so it was likely from the outset that ensemble trees and kernel-based models would be more efficient for classification than others. For binary classification of the most important performance metrics, we chose sensitivity (i.e., what percentage of actual FCS patients the model categorizes as FCS patients), specificity (i.e., what percentage of non-FCS patients the model categorizes as non-FCS patients) and accuracy (i.e., the proportion of patients in the total population who received an accurate classification regardless of whether they were FCS patients or non-FCS individuals) in addition to the ROC-AUC curve. The AUC value (as an integral of the ROC curve) can be interpreted as the probability that a positive sample is ranked higher in confidence than a negative sample. It is worth noting that the importance of the data matrix properties can be determined in several ways based on the trained models, while for SVM and deep learning the estimation of the (single) importance of these input variables is not straightforward. In the case of gradient boosting, these properties can be determined exactly from their position on the trees. For AdaBoost, SVM implementation, we used scikit-learn 0.23, while for ReLU meshes we used PyTorch 1.6 and for trees we used XGBoost 1.2.1.

Laboratory tests

The tests were always carried out according to the current protocol of the health institution, not necessarily after overnight fasting. Lp(a) levels were determined from fasting blood samples in the Department of Laboratory Medicine, University of Debrecen Clinical Center, using a Cobas c501 Analyser (Roche Ltd., Mannheim, Germany) according to the manufacturer's instructions and protocol. The reference range for Lp(a) analysis was < 300 mg/L.

Summary of the statistical analyses

The statistical analyses were performed with Python-supported data mining and statistical software packages, mainly in Conda 4.X environment, using Python 3.8, IPython 7.2x, Cython 0.2x, Pandas 0.2x and Numpy 1.2x. Significance was calculated in all cases at the 95% significance level and using unpaired t-tests. Statistical plots were generated using Matplotlib 3.5 software package and associated libraries (Seaborn).

Results

Regional prevalence of FH and FCS

We used two scoring systems to estimate prevalences of FH and FCS, such as the DLCN and FCS scores. First, we created subgroups based on diagnoses, symptoms and laboratory values, taking into account the medical logic behind the systems. Then we calculated the corresponding score for each patient, based on all available data representing the probability of the disease in the individual. This was also necessary because the prevalence estimation on clinical data sets is already considered a statistical process, where only limited information is available on individuals. The differences resulting from the two approaches can also provide additional useful information, such as belonging to one of the two cohorts may provide a strong probability of belonging to the other cohort even if the data for that individual is not available.

To identify patients with FH and thus estimate the disease prevalence, nine groups of patients were formed, taking into account cardiovascular risk groups. The first group included those for whom a definite FH diagnosis could be made using text mining, while the second group included those with a history of high cholesterol. The third, fourth and fifth groups included individuals with a diagnosis of CAD (coronary artery disease), CVD (cerebrovascular disease) and PAD (peripheral arterial disease). The sixth group included patients with a diagnosis of tendon xanthoma and the seventh group included patients with a diagnosis of arcus corneae. The eighth group included patients with high LDL-C and low triglycerides, while the ninth group was a selection of patients with high total cholesterol and low triglycerides. For each patient, the DLNC score was calculated and used to classify them into definite, probable and possible FH categories.

Our results suggest that 0.89% of patients in the region may be affected by familial hypercholesterolaemia. Since nearly 76% of the Hungarian population visits the public health care system once a year and the hospital goers' statistics are essentially in line with the 2011 regional census data, clinical prevalence may accurately estimate the actual prevalence. Genetic testing offered to patients with a DLSN score of three or above confirmed a genetic diagnosis of FH in 33% of patients in the literature. Based on these results, we calculate that 1 in every 340 people in our country may have familial hypercholesterolemia (33% of patients with DLSN values above 3 compared to the total patient population).

Our evaluation did not take into account the genetic and family history elements of the DLNCS, as these data were very limited. This was partially compensated by the relatively high annual attendance of the population in public health in our region and country, and in

addition, we used machine learning methods to compensate for these shortcomings. Thus, data from individuals with different DLCN scores were used as input for machine learning to test which patient composition for each score provides the best learning, i.e. the highest ROC-AUC curve. A total of 459 patients were identified, leading to the conclusion that one in every 1,300 people in the region (459:590,500) is suffering from FH.

We also calculated the FCS score for all patients with familial chylomicronemia syndrome. All individuals with a score ≥ 10 were considered to have FCS as proposed by Moulin *et al.* Consequently, a total of 26 patients were identified, resulting in a regional estimated prevalence of FCS at 19.4:1,000,000 (19.4 ppm).

A different approach to the rapid estimation of FCS prevalence was to define key characteristics that quickly reduce (or "cut") the large number of cases to a lower number of cases by orders of magnitude. In rare metabolic diseases, extreme variation in a single laboratory value is common, providing an opportunity for rapid, large-scale, crude screening of individuals. As a starting point, we used the FCS score system and the medical rationale behind the disease and integrated this body of knowledge using data mining and machine learning methods.

Patients with high FCS scores were selected for the rapid estimation. In selecting key disease characteristics, we used high-scoring and well-measured items. Such key characteristics were (i) a serum TG level above 10 mmol/l on three consecutive blood tests (5 points), which (ii) was never documented to be below 2 mmol/l (which would have resulted in a 5-point deduction); (iii) no secondary factor present (2 points); (iv) TG level above 20 mmol/l on at least one occasion; (v) symptoms before the age of 40 years; and (vi) positive history of pancreatitis.

Based on all this, the prevalence of FCS among patients in Debrecen was 8.47 ppm, while it was 5.32 ppm in Nyíregyháza. When the FCS score was calculated for each patient available in the database, the frequency of occurrence of patients with an FCS score of 10 or higher (i.e., a very probable diagnosis of FCS) was 27.11 ppm in Debrecen and 13.3 ppm in Nyíregyháza.

At the same time, these frequencies of occurrence were an order of magnitude higher than the previously known literature data. Therefore, we decided to subject the data of the 275 patients in Debrecen who were classified as unlikely and very likely FCS to a detailed medical check. At the end, we found a total of 21 (7 clinically certain and 14 very likely) FCS patients estimating the prevalence of FCS at 11.8-35.6 ppm among patients of University of Debrecen Clinical Center, similar to the data mining results.

Recommendations for the more effective diagnosis of FCS

With the help of machine learning, entirely new associations can be discovered on the trained medical data, which can be used to identify new, previously undiagnosed patients in the population and reveal the deeper medical connections of the disease.

Using the data of the FCS patients in Debrecen as vector spaces, we searched for other, potentially affected individuals in both health care institutions by building mathematical models with several machine learning methods. The data of the above mentioned 21 patients were used as a training set, while the data of 275 patients and randomly selected individuals were used as a test set. Again, the kernel methods and boosting algorithms performed extremely well. In machine learning, we tried to understand which characteristics of the patients are crucial for the models and how they can be used to make the clinical diagnosis of FCS more accurate.

Our observations have confirmed that TG levels are the most important diagnostic criteria when considering the relative importance of individual laboratory parameters, since FCS is also best characterized by serum triglyceride levels. We also found that the maximum and mean TG levels in the medical history were of particular importance, and the fluctuation between TG levels (the distribution and variation between the highest and lowest values) was also important in the decision of the models. Indeed, individuals were more likely to receive the diagnosis of FCS if there was not a large difference between the minimum and maximum TG levels during the medical history. An important new finding was that the mean and the maximum cholesterol levels are essential to distinguish FCS patients from non-FCS patients. This observation is particularly interesting in the light of the fact that cholesterol testing is not currently part of clinical practice in the diagnosis of FCS.

Further correlations were found when we examined the thresholds at which the mathematical models change their decision, i.e. move from FCS to non-FCS, and from non-FCS to FCS diagnosis. These cut-off values can be used for population screening, but they do not represent absolute importance, but indicate whether the probability of an FCS diagnosis increases or decreases below or above a given value. Besides the previously mentioned triglyceride and cholesterol cut-offs, laboratory results exceeding certain values of hemoglobin, MCHC, basophil granulocyte, lymphocyte or amylase results increase the chance of the diagnosis of FCS. On the other hand, elevated white blood cell, neutrophil granulocyte, sodium, glucose, CRP, liver and kidney function tests decrease the likelihood of

the diagnosis of FCS. Interestingly, the presence of elevated inflammatory parameters, such as higher white blood cell count and CRP also reduces the likelihood of FCS diagnosis.

Complex examination of FH patients according to Lp(a) categories

We carried out our Lp(a) studies among the patients attending the University of Debrecen Clinical Center (n=590,500). The best training results for DLCN scores were above 7 and 8, which, in accordance with the concept of DLCNS, indicated that individuals with a score of 7 or above were considered to be FH patients. Based on this model we identified 459 patients, of whom 221 (48%) had an Lp(a) laboratory measurement available.

FH patients had significantly higher Lp(a) values than the non-FH patient population [236 (92.5; 698.5) mg/l vs. 167 (80.2; 431.5) mg/l, $p < 0.01$]. In both groups, we found that women had significantly higher Lp(a) values than men [266 (108-875) mg/l vs. 182 (73.1-648) mg/l, $p < 0.05$] among FH patients and also among non-FH individuals [179 (81.5-458) mg/l vs. 152 (80.7-386) mg/l, $p < 0.01$]. Although atherosclerotic complications were found to be significantly more frequent in FH patients than in the non-FH group (46.6% vs. 13.9%), we found no significant association between Lp(a) levels and the frequency of atherosclerotic complications.

Based on our results, high blood pressure, smoking, obesity, hyperuricemia, diabetes mellitus and low HDL cholesterol levels are more common among FH patients. In order to clarify the role of Lp(a) in the population affected by the disease and those without it, we investigated the relationship between atherosclerotic diseases and cardiovascular risk factors. The most important risks were high blood pressure, age over 60 and lack of statin treatment, as well as smoking and obesity. Male gender was also considered to be a significant risk. Examining the risk factors in FH patients with high and low Lp(a) values, we found no significant differences between the two groups. In the comparison between the sexes, however, it was observed that high Lp(a) values occurred more often in women in all groups; however, the difference was not significant in the case of FH patients. In addition, the frequency of smoking and hypertension was lower among men without FH.

We also used the raw data set to determine the relative hazard ratios (HRs) of cardiovascular diseases associated with each risk factor. We found that hypertension and high LDL-C levels had the highest relative hazards, followed closely by smoking. In FH patients, both hypertension and smoking increased significantly the already high cardiovascular risk. However, a significant proportion of patients also had more than one risk factor. Among

patients with cumulative risk factors present, the relative hazard was 27.42, indicating a nearly 30 times higher cardiovascular risk in patients with multiple risk factors.

Finally, when analyzing Lp(a) levels in the whole population independently of other risk factors, we concluded that serum Lp(a) levels exceeding 1000 mg/l increase significantly the cardiovascular risk.

Summary of the new findings

- With data mining, we found the regional FH prevalence to be 1:340. Using machine learning methods, the ratio of patients with a certain FH diagnosis was calculated to be 1:1300.
- The regional frequency of FCS was estimated to be 19.4:1,000,000 with data mining; which was calculated to 13.3 and 27.11 ppm when examining the two healthcare providers separately. Using machine learning methods, we estimated the prevalence of FCS at 11.8-35.6 ppm, which is approximately 10 times higher than the recent prevalence data.
- With the help of ensemble methods and shallow neural networks, we developed effective mathematical models that are suitable for separating FCS patients from other clinical patients using routine blood sampling results (average AUC: 0.98) and respectively distinguish FCS patients from other non-FCS hypertriglyceridemic patients (mean AUC: 0.87).
- With our newly developed machine learning models, we found some new laboratory parameters that can further refine the recent FCS scoring system. Serum total cholesterol level is a particularly good candidate, since its high values improve the chance of FCS diagnosis, while very high levels of serum cholesterol reduce the probability of having FCS. In this way, a range can be determined in the future where the probability of FCS diagnosis is the highest (based on our current tests, this cholesterol range is between 6.5 and 11.0 mmol/l).
- We also confirmed that the ASCVD risk was significantly higher in FH patients (46.6% vs. 13.9%), as well as the Lp(a) serum values were significantly increased compared to the population without FH. We could not prove a direct correlation between higher Lp(a) values and increased cardiovascular risk in FH patients. At the same time, the incidence rate of each cardiovascular risk factor was significantly higher in FH patients, while the individual relative risk values were the same.

- Regarding gender differences, serum Lp(a) levels were found to be significantly higher among women, regardless of the presence of FH.
- We also found that cardiovascular risk increases significantly, when Lp(a) level is above 1000 mg/l. The relative individual risk is, however, lower compared to the common risk factors (such as smoking or hypertension).
- Overall, our results prove that data mining and machine learning methods applied on large amounts of clinical data is effective in supporting everyday clinical work and diagnostics and may also improve scientific research providing new insights on previously unknown associations.

Discussion

Prevalence of rare diseases

Although rare diseases affect fewer individuals in the population, the potential health consequences and the costs of the resulting medical care may impose a significant health burden on both the patient and society. Determining the exact prevalence of these diseases is also a major challenge. Depending on the genetic predisposition, familial hypercholesterolemia has a prevalence of 1:250 to 1:1,000,000 in the population, while the prevalence of familial chylomicronemia syndrome is also estimated at 1:1,000,000. Although the prevalence and severity of these diseases pose a significant challenge to medical practitioners in our region, the exact prevalences of FH and FCS are unknown.

To the best of our knowledge, we were the first in our region to use data mining and machine learning methods in order to estimate the prevalence of rare diseases on large amounts of health data. We calculated that there were 340 FH patients for every 1,000,000 people in the Northern Great Plain region, which is in line with data from several European countries. Screening for FH has a major role in the prevention of cardiovascular diseases and the use of newly developed and effective lipid-lowering therapies (e.g. PCSK9 inhibitors) may also contribute to accelerate FH screening, as more and more patients may be improved by these new therapeutic options. Using machine learning methods on the patient data from Debrecen, we found 459 probable FH patients, reducing slightly the prevalence of FH to 1:1300. It is important to note that we aimed to ensure that the identified individual was definitely an FH patient, so we may have underestimated the true prevalence.

When FCS was studied using similar methods, the prevalence of the disease was found to be 19.4 ppm, which is higher than the estimated global prevalence of 1 ppm. On the other hand, it should be noted that recent studies have provided data comparable to our results. Indeed, the prevalence of FCS was found to be 13 ppm in a population with a similar size attending health care providers. In a smaller patient population, FCS prevalence was found to be 1 in 300,000 children, which also suggests a higher prevalence than the currently generally accepted estimates.

The variation found in the prevalence may also be due to the differences in sampling, depending on whether the sample is taken from a healthy or an unhealthy population. Also, the estimated prevalence may be higher in centers providing higher levels of patient care, as patients requiring more thorough investigations are more likely to present in the most specialized institutions. Our studies yielded similar results, either with a quick estimate based on key characteristics or with individual scores describing each patient. In any case, our

analyses suggest that the estimated prevalence of FCS may be one order of magnitude higher than currently accepted.

The diagnosis of FCS currently relies heavily on genetic analysis and the measurement of LPL enzyme activity. On the other hand, the recently developed FCS score system supports a quick and effective recognition of FCS. As we proved, serum TG level and its specific values (e.g. a TG level occurring above 10 mmol/l at least three times and being never below 2 mmol/l) quickly narrows the potentially affected population to 1-2%. This close and obvious relationship between serum TG levels and the disease allows the successful application of simple scoring systems, but, at the same time we found that the patients actually diagnosed with FCS are often not the ones with the highest FCS score. This can also be explained by the fact that we did not have all the history data to reliably calculate the score. Taking all of this into account, we believe that the previously proposed FCS scoring system works with outstanding efficiency in all cases where all data of the patient is available, but its use is more limited and can be further refined when patient data is scant. Machine learning and data mining may, however, help in these situations.

Refinement of clinical criteria and protocols

Machine learning also seems to be a useful tool to support the clinician with developing accurate, usable protocols and to better understand the pathomechanism of rare diseases. With the help of our lipidologist experts, we found that kernel and boosting-based models were very efficient at separating FCS and non-FCS patients with high ROC/AUC values in binary classification, while neural networks performed significantly worse. With further training, we tried to find the key serum levels ("cuts") where the "patient - non-patient" binary separation occurs with the best efficiency. We found that the examination of serum TG values clearly helps the accurate diagnosis of FCS.

In our opinion, an interesting and important novelty is that the assessment of serum total cholesterol level may provide additional help to identify potential FCS patients. Our studies suggest that patients with total cholesterol levels below 4 mmol/l and above 11 mmol/l have a low probability of FCS diagnosis reflecting on the close association between triglyceride and cholesterol metabolism and carrier lipoproteins.

In addition to lipid parameters, we also found several other metabolic components influencing the likelihood of FCS. These observations are likely to be explained by the high prevalence of insulin resistance and related conditions such as obesity, type 2 diabetes and

non-alcoholic fatty liver disease among patients presenting to healthcare providers, which is in line with the previous findings.

Pathogenesis of FH in the light of metabolic deviations and other factors

Mathematical-statistical operations on large datasets can also be used to investigate disease pathogenesis, explore complex dependencies and causal relationships. Although all FH patients have high serum LDL-C levels, they possess various cardiovascular disease frequencies with a wide phenotypic heterogeneity. It is well known that the co-occurrence of different risk factors significantly increases the risk of atherosclerosis and contributes to phenotypic heterogeneity. Although the relationship between Lp(a) levels and CVD risk in heterozygous FH patients has been identified, the potential role of elevated Lp(a) concentrations as a risk factor for this disease remains unclear.

To our knowledge, we were the first in Hungary to map the cardiovascular risk factors of FH patients with data mining methods. The proportion of hypertensive patients turned out to be extremely high (86.3%), which is similar to other data from Central and Eastern Europe. At the same time, the prevalence of hypertension was significantly lower in some Western European and South American populations.

We also found that a surprisingly large number of FH patients smoke (66.4%), which is comparable to Polish data. The frequency of diabetes was significantly larger among our FH patients than in the non-FH population, which is very similar to Polish, Romanian or Mexican FH patient data.,. Our results also underline the importance of examining risk factors and highlight the need for extensive, international surveys in this patient group. It is also worth noting that some studies found a much lower prevalence of diabetes and observed that the severity of the consequences of mutations in FH patients is inversely proportional to the prevalence of diabetes; however, we were unable to confirm these results. It is believed that the conflicting data may be due to the varying socioeconomic status, as well as genetic differences among individuals. In addition, patients and their relatives usually have incomplete knowledge of both FH and its cardiovascular risk, which may also contribute to the very high prevalence of modifiable risk factors, including smoking. As with the FCS, raising awareness among patients and relatives of FH is of major importance.

Relatively few studies investigated Lp(a) levels in FH patients in detail, however the data available to date are in substantial agreement with our results. It is also worth mentioning that, regardless of the presence of FH, we found significantly higher Lp(a) values among women, which is also in agreement with literature data. Although the exact reason for

the gender difference is not known, it is assumed that the effect of estrogen on Apo(a) protein expression may be responsible for this variation.

Although the low number of Lp(a) measurements prevented to identify significant differences in FH patients, the found tendencies suggest that extremely high Lp(a) values further increase the risk of cardiovascular events. However, subsequent analyses are needed to confirm these findings and to calculate the exact hazard.

Our data suggest that the very high prevalence of vascular disease in the FH population of the Northern Great Plain region is largely explained by the significant clustering of major cardiovascular risk factors (e.g. smoking, hypertension and obesity). Patients with multiple metabolic risk factors may particularly benefit from new treatment options and personalized therapeutic strategies. PCSK9 inhibitors antisense oligonucleotides and newer short interfering RNA preparations can achieve Lp(a) reductions of up to 80-90%, providing potential benefit to both FH and non-FH patients with high serum Lp(a) levels. On the other hand, the cost of these new therapeutic options is very high, so adequate control of other modifiable risk factors and adverse habits is of paramount importance. Non-smoking patients with optimal body weight and LDL-C values, who have both hypertension and diabetes may largely benefit from the new treatment options, which is closely supported by our HR calculations.

Limitations

The fundamental shortcoming of retrospective research based on large amounts of data is that we cannot influence the data quality and structural characteristics. Thus, we had limited or no access to family history and previous genetic analysis of the patients. Also, modification of LDL-C levels is recommended in patients on lipid lowering treatment when determining DLCN scores, however, we had no information on whether GPs initiated such therapy prior to patient referral. For this reason, we used the average of the lipid levels measured before treatment in all such cases. Finally, some tests including Lp(a) determination were not generally available, therefore were only checked in certain patients, mostly older subjects with suspected or already diagnosed CVD.

Summary

We found that methods of data mining and machine learning, when applied to a large amount of clinical and hospital data, was extremely useful in estimating the prevalence of rare diseases and identifying patients with these conditions.

We estimated higher FH and FCS prevalence compared to previous literature data. We also managed to prove that DCLN and FCS scoring systems are very effective in diagnosing FH and FCS patients, being also useful in identifying pathologies caused by severe genetic involvement.

Additionally, we concluded that the FCS scoring system itself can be further improved using machine learning tools. This also seems to be useful, because score calculation mainly relies on information that is often unavailable in health databases. Teaching on a large dataset may also reveal previously unknown correlations in the disease course. Indeed, we strongly believe that measuring serum total cholesterol levels would allow a more accurate diagnosis of FCS and a more confident identification of such patients.

The applicability of the method is partly hampered by the "black box" nature of some machine learning models, i.e. that exact decision paths are difficult or impossible to infer due to the complex weighting and the existence of various parameters in the model. Yet, these modern solutions might be important elements of the mass, although personalized, diagnostics of the future, mainly due to their fundamental property of rapid and mass applicability. Most patients visit healthcare providers several times, during which a standard medical examination and routine blood sampling usually take place. This data may provide sufficient information to a rapid estimate of the cardiovascular risk of FH or FCS and would allow earlier identification of rare diseases other than the above mentioned contributing to early treatment.

Computer-assisted testing may also help to select suitable candidates for further diagnostic and genetic testing, thereby reducing the resources spent on diagnostics. Our research demonstrated that data mining and machine learning methods were also effective in quantifying risk factors, and may therefore serve as an important element of scientific research beside diagnostics.

Acknowledgements

I would like to thank Professors Dr. József Balla and Dr. György Paragh, current and former directors of Institute of Internal Medicine, University of Debrecen Faculty of

Medicine and Department of Internal Medicine, University of Debrecen Clinical Center for encouraging and allowing me to participate in the scientific work in the department.

I would also like to thank Professor Dr. István Édes, former president of University of Debrecen Clinical Center for his help in creating and launching the machine learning and data mining research from which the results of this thesis has grown.

I would like to express my gratitude to my supervisor Dr. Péter Fülöp, who supported and guided my scientific work, helped me to understand the medical background, planned the research and solved the problems that arose.

I am also unspeakably grateful to Professor Dr. Harangi Mariann, for her help in my research and work, who was always available when I needed help.

I would like to thank Dr. Zsolt Karányi, bioinformatics expert of Institute of Internal Medicine, for being a prominent supporter of the research, as well as for his useful advice and for the continuous expansion of my knowledge.

Special thanks go to the daily mentoring and friendly help of Dr. Bálint Daróczy, without whom this dissertation could not have been completed. His vast knowledge of machine learning and his constant availability were very precious to me.

I would like to express my gratitude to Dr. Hajnalka Lőrincz, as well as to the complete staff of the Research Laboratory of Institute of Internal Medicine and to the co-authors of my scientific papers for their help and support.

Finally, I am extremely grateful to my family and friends for their patience when I was unable to spend more time with them.



Registry number: DEENK/26/2024.PL
Subject: PhD Publication List

Candidate: Ákos Németh
Doctoral School: Doctoral School of Health Sciences

List of publications related to the dissertation

1. **Németh, Á.**, Daróczy, B., Juhász, L., Fülöp, P., Harangi, M., Paragh, G.: Assessment of associations between serum lipoprotein (a) levels and atherosclerotic vascular diseases in Hungarian patients with familial hypercholesterolemia using data mining and machine learning.
Front. Genet. 13, 1-10, 2022.
DOI: <http://dx.doi.org/10.3389/fgene.2022.849197>
IF: 3.7
2. **Németh, Á.**, Harangi, M., Daróczy, B., Juhász, L., Paragh, G., Fülöp, P.: Identifying Patients with Familial Chylomicronemia Syndrome Using FCS Score-Based Data Mining Methods.
J Clin Med. 11, 1-14, 2022.
DOI: <http://dx.doi.org/10.3390/jcm11154311>
IF: 3.9
3. Paragh, G., Harangi, M., Karányi, Z., Daróczy, B., **Németh, Á.**, Fülöp, P.: Identifying patients with familial hypercholesterolemia using data mining methods in the Northern Great Plain region of Hungary.
Atherosclerosis. 277, 262-266, 2018.
DOI: <http://dx.doi.org/10.1016/j.atherosclerosis.2018.05.039>
IF: 4.255

List of other publications

4. Kovács, B., **Németh, Á.**, Daróczy, B., Karányi, Z., Maroda, L., Diószegi, Á., Harangi, M., **Pál, D.**: Assessment of Hypertensive Patients' Complex Metabolic Status Using Data Mining Methods.
J. Cardiovasc. Dev. Dis. 10 (8), 1-14, 2023.
DOI: <http://dx.doi.org/10.3390/jcdd10080345>
IF: 2.4 (2022)





5. Kovács, B., **Németh, Á.**, Daróczy, B., Karányi, Z., Maroda, L., Diószegi, Á., Nádró, B., Szabó, T., Harangi, M., Páll, D.: Determining the prevalence of childhood hypertension and its concomitant metabolic abnormalities using data mining methods in the Northeastern region of Hungary.
Front. Cardiovasc. Med. 9, 1-10, 2023.
DOI: <http://dx.doi.org/10.3389/fcvm.2022.1081986>
IF: 3.6 (2022)
6. Sturm, Á., Sharma, H., Bodnár, F., Aslam, M., Kovács, T., **Németh, Á.**, Hotzi, B., Billes, V., Sigmond, T., Tátrai, K., Egyed, B., Téglás-Huszár, B., Schlosser, G., Charnpilas, N., Ploumi, C., Perczel, A., Tavernarakis, N., Vellai, T.: N6-Methyladenine Progressively Accumulates in Mitochondrial DNA during Aging.
Int. J. Mol. Sci. 24 (19), 1-19, 2023.
DOI: <http://dx.doi.org/10.3390/ijms241914858>
IF: 5.6 (2022)
7. Alimohammadi, S., Péntes, Z., Horváth, D., Gyetvai, Á., Bácsi, A., Kis, G., **Németh, Á.**, Arany, J., Oláh, A., Lisztes, E., Tóth, I. B., Bíró, T., Szöllősi, A. G.: TRPV4 activation increases the expression of CD207 (Langerin) of monocyte-derived Langerhans cells, without affecting their maturation.
J. Invest. Dermatol. 143 (5), 801-811, 2023.
DOI: <http://dx.doi.org/10.1016/j.jid.2022.10.022>
IF: 6.5 (2022)
8. Paragh, G., **Németh, Á.**, Harangi, M., Banach, M., Fülöp, P.: Causes, clinical findings and therapeutic options in chylomicronemia syndrome, a special form of hypertriglyceridemia.
Lipids Health Dis. 21 (1), 1-14, 2022.
DOI: <http://dx.doi.org/10.1186/s12944-022-01631-z>
IF: 4.5

Total IF of journals (all publications): 34,455

Total IF of journals (publications related to the dissertation): 11,855

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.



26 January, 2024