Cooperation Between Data Analysts and Medical Experts: A Case Study

Judita Rokošná, František Babič, Ljiljana Trtica Majnarić, Ľudmila Pusztová

Cooperation between Data Analysts and Medical Experts: A Case Study

Judita Rokošná[1], František Babič[1], Ljiljana Trtica Majnarić[2,3], Ľudmila Pusztová[1]

[1] Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Košice, 042 01 Košice, Slovakia
[2] Department of Internal medicine, Family Medicine and the History of Medicine, Faculty of Medicine, Josip Juraj Strossmayer University of Osijek, Josipa Huttlera 4, 31000 Osijek, Croatia
[3] Department of Public Health, Faculty of Dental Medicine and Health, Josip Juraj Strossmayer University of Osijek, Crkvena 21, 31000 Osijek, Croatia
judita.rokosna@student.tuke.sk, frantisek.babic@tuke.sk,ljiljana.majnaric@mefos.hr, ludmila.pusztova.2@tuke.sk

Abstract. The medical diagnosis and determine a correct medical procedure represent a comprehensive process that consists of many input information and potential associations. This information can lead to clinical reasoning to resolve a patient's health problem and set the treatment. Effective communication between the medical expert and data analyst can support this process more effectively, dependent on the available data. It is essential to create a shared vocabulary for this cooperation to reduce possible misunderstandings and unnecessary experiments. In our work, we performed exploratory data analysis, statistical tests, correlation analysis, and logistic regression in close cooperation with the participated expert thanks to whom we could verify the achieved results of our models. The collaboration between the medical expert and data analytic requires a lot of communication and explanation from the medical expert because of the correct interpretation of the medical data and its resulting associations. On the other hand, it requires a proper understanding of the task from the data analyst's point of view, a lot of iterations of graphs, and other models, which must be modified to be easy to read and interpret.

Keywords: medical expert, data analysts, collaboration, thyroid.

1 Introduction

The medical diagnosis and determine a correct medical procedure represent a comprehensive process, where it is necessary to consider some variables and relationships among them. The medical record consists of a quantity of information about the patient, such as necessary information (age, year of born, weight, height, etc.), its results of tests, disease data and its development, and so on. All the medical records' information is useful because of the determination of the right medical treatment of partic-

ular diseases. Based on them, the physician can find hidden dependencies between identifiable symptoms and potential definite diagnosis.

Among the critical skills of all physicians belong to constructing diagnosis, choosing diagnostic tests, and interpreting the results. These skills are essential in the diagnosis process. The diagnosis process [1] is a process like any other. The process is particularly complex patient-centered because it includes many handoffs of information or materials and clinical reasoning to determine a patient's health problem. The diagnostic process draws on an adaptation of a decision-making model that describes the cyclical process of information gathering, integration, and interpretation, and forming a diagnosis. This process and decision making can be particularly complicated when caring for older patients with aging diseases, such as hypertension. Aging diseases typically appear as multiple comorbidities, which imply complex, often non-linear relationships between disorders [2]. Little is known about these relationships. Patients' classification into smaller, more homogeneous groups would be necessary to help decision making.

Thyroid diseases can occur at different ages, and the type of disorder is somewhat age and sex-related. The most common thyroid disorder in older people, prevalently women, is subclinical hypothyroidism [3]. It is prevalent in persons with hypertension and with decreased renal function. Although without overt clinical symptoms, this disorder is associated with an increased risk of CVD and cognitive impairment. The relationships between the TSH hormone levels and BMI and other CV risk factors, depending on different stages of renal function decline, in persons with hypertension are poorly known. In this critical area of clinical decision making, there may be the benefit of data analytics methods and collaboration between a data analyst and the medical expert.

1.1 Thyroid disease and hypothyroidism

Thyroid gland diseases are the most prevalent endocrine disorders. They have a relatively high incidence in the general population, and these diseases occurred more often in women. Thyroid disease may occur in 5 – 10 % of the general population and in women in middle-aged to older age 10 – 15 %. The regions with iodine-deficient constitute a third of the world population, what it can be an essential risk factor for the onset of goiter and hypothyroidism. The prevalence of hypothyroidism and hyperthyroidism is different between women and men. Hyperthyroidism is prevalent in 0.5 – 2 % of women and 0.1 – 0.2 % of men, whereas hypothyroidism occurs in 0.06 – 1.2 % in women and 0.1 – 0.4 % in men [4].

Hypothyroidism [5] is a common endocrine disorder, and it affects hundreds of millions around the world. This disorder is resulting from a deficiency of thyroid hormones or by the complete loss of its function. About 95 % of hypothyroidism cases occur as a result of primary gland failure. One of the grades of hypothyroidism is subclinical hypothyroidism [6]. This disorder is a prevalent disorder among middle-aged and elderly patients. It represents a state with increased thyroid-stimulating hormone - TSH and typical values of thyroxine – T4 and triiodothyronine – T3. If the values of TSH are above 4.0 mU/l, it represents increased levels. By their increase,

the subclinical hypothyroidism can be divided into a mild form (values from 4.0 – 10.0 mU/l) and a more severe form (values above 10.0 mU/l). Most patients with subclinical hypothyroidism have no symptoms that would indicate this disorder. So, the diagnosis of this hypothyroidism is made based on laboratory findings. Among the general symptoms of hypothyroidism can include [7]: less energy, more fatigue, drier and itchier skin, drier and more brittle hair and more hair loss, loss of appetite, weight gain, slow thought and speech, muscle cramps and joint aches, slowing of heart rate, slightly higher blood pressure, higher cholesterol level, hoarse voice or depression. The only way to know whether you have hypothyroidism is with a blood test. The complete cure of hypothyroidism is cannot possible, but it can be treated by wholly controlled. Hypothyroidism is treated by replacing the amount of hormone that thyroid can no longer make to bring T4 and TSH levels back to normal levels. T4 replacement can restore the body's thyroid hormone levels and the body's function if the thyroid gland cannot work right.

1.2 Related works

Data analytics is becoming a strategically important tool for many organizations, including the healthcare sector. The goal is to search for new patterns and knowledge, which are not visible sometimes for the first look or to confirm an assumption of the physician on more large data samples. This process requires intensive collaboration between analysts and relevant domain experts. In this section, we present some works related to this crucial aspect of the analytical process and some existing studies related to thyroid disease.

The research group of professor Andreas Holzinger propose a concept called an interactive machine learning with a human-in-the-loop, i.e. “algorithms that can interact with agents and can optimize their learning behavior through these interactions, where the agents can also be human.” [8], [9]. Mao et al. investigated various aspects of the collaboration between data scientist and domain experts through the interviews with bio-medical experts collaborating with data scientists. They identified bottlenecks like collaboration and technology readiness or coupling of work dimensions [10]. Jeanquartier and Holzinger applied a visual analytics process in cell physiology [11]. Experts in this domain are using their domain knowledge in combination with both automatic and visual analysis together. They need to be guided by computer science experts to improve the choice of tools that are used. They point out that it is a gap between free available visualization tools and their usability for the experts.

Ionita I and Inoina L, in their study, used for experiments the dataset provided by UCI Machine Learning Repository containing data about clinical history patients with thyroid disorders [12]. The authors applied four types of classifications methods: Naïve Bayes, Decision Trees, Multilayer Perceptron, and RBF Network (Radial Basis Function Network). The best accuracy was obtained by the Decision tree (97.35%). Sidiq et al. worked with a clinical dataset from one of the leading diagnostic labs in Kashmir [13]. The authors generated several classification models based on the following methods: K-Nearest Neighbor, Support Vector Machine, Decision Tree, and Naïve Bayes to diagnose thyroid diseases. They used a programming language, Py-

thon, and 10-fold cross-validation. The best accuracy of 98.89% was achieved again by the Decision Tree approach. Logistic regression was used by the collective of authors to predict the patients without thyrotoxicosis and with thyrotoxicosis [14]. The outcomes demonstrate that the logistic regression obtains promising results in classifying regression on thyroid disease diagnosis 98.92%

2 Methods

Typically, the analytical process contains in the modeling phase (according to the CRISP-DM methodology) an application of suitable machine learning, artificial intelligence, or statistical methods. We have experience with all these methods, and sometimes it isn't easy to apply them based on the data quality or data range. On the other hand, the data analytics domain offers other unusual methods like EDA to meet the domain experts' expectations.

John W. Tukey defined exploratory data analysis (EDA) as “detective work – or more precisely numerical detective work” [15]. It is mainly the philosophy of data analysis. EDA's primary goal is to examine the data for distribution, outliers, and anomalies to direct specific testing of the hypothesis. The EDA gives analysts unparalleled power for gaining insight into the data and their visualizations [16].

The Chi-square test (also known as the Pearson Chi-square test) is the most useful a nonparametric statistical test that measures the association between two categorical variables [17]. This test utilizes a contingency table to analyse the data, where each row represents a category for one variable, and each column represents a category for the other variable. Each cell of the contingency table reflects the total count of cases for a specific pair of classes.

The most popular methods among normality tests are the Kolmogorov – Smirnov test, the Anderson – Darling test, and the Shapiro – Wilk normality test [18]. The Shapiro – Wilk test tests the null hypothesis on the significance level α that a sample $x_1, x_2, \ldots, x_n$ comes from a normally distributed population *norm* (μ, σ) with unspecified parameters μ and σ, where μ is median, and σ is the standard deviation [19].

Students’ t-test is one of the most commonly used methods of statistical hypothesis testing. This test is a parametric test, and it is used to compare samples of normally distributed data. There are several types of t-test: one-sample t-test, paired-samples t-test, unpaired two-sample t-test. Unpaired two samples t-test, also called independent t-test, is a parametric test, which compares the means of two independent groups [20]. This test aims to determine whether there is statistical evidence that the associated population means are significantly different. The main requirements for the independent t-test are that two groups of samples are normally distributed, and the variances of the two groups are equal.

Mann-Whitney U test is a nonparametric test, which is an alternative to the independent sample t-test [21]. This test is used to compare two population means that come from the same population. It can also apply to test whether two population means are equal or not. The test is used when the data fails the normality assumption or if the sample sizes in each group are too small to assess normality.

ROC curve or a Receiver operating characteristics curve is a useful technique for visualizing, organizing, and selecting classifier based on performance [22]. The receiver operating characteristics curve is computed by comparing a binary outcome with a continuous variable. Each observed level of a continuous variable is evaluated as a candidate cut point discriminating observed positive from negative [23]. Positive observations concerning the continuous measurement are these, exceeding the cut point, while those less than or equal to the cut point are classified negatively.

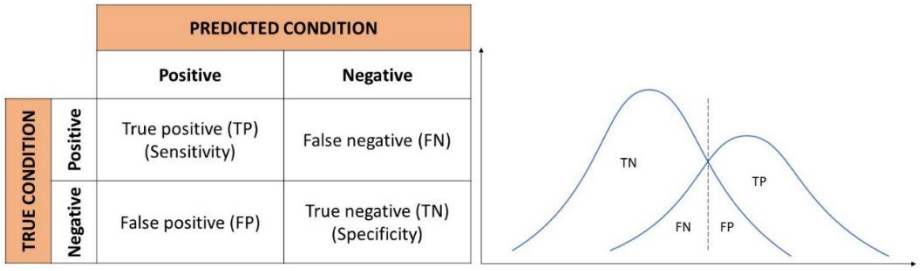

Fig. 1. Example of the contingency table.

As shown in the table, sensitivity is defined as a probability that observation with a positive outcome is correctly classified by a continuous measurement above a candidate cut point.

$$sensitivity = \frac{true\ positive}{true\ positive + false\ negative}$$

And specificity is defined as the probability that a continuous measurement correctly classifies an observed negative outcome at or below the candidate cut point.

$$specificity = \frac{true\ negative}{true\ negative + false\ positive}$$

The coordinates for the ROC curve are computed where the x-axis is 1 − specificity (= false positive rate; FPR), and the y-axis is sensitivity (= true positive rate; TPR). The best cut point given the data may be identified from the ROC curves coordinates with a criterion that maximizes TPR (True Positive Rate) and minimizes FPR (False Positive Rate).

To evaluate the ROC curve can be used the AUC (Area Under Curve), also known as the c-statistics [24]. The AUC metric varies between 0.5 and 1.00 (Ideal value), where values above 0.80 is an indication of a good classifier.

Logistic regression is a method of modelling the probability of an outcome that can only have two values [25]. The main is to find the best fitting model that described the relationship between the dichotomous variable (dependent variable – it contains data coded as 1/0, TRUE/FALSE, yes/no, etc.) and asset of independent variables. The final logistic regression generates the coefficients (standard errors, significance lev-

els) of a formula to predict the logit transformation of the probability of the presence of the characteristic of interest.

3 Analytical process

This study was motivated by previous successful cooperation between Slovak data analysts and medical experts from Croatia. This research team cooperated on several preliminary studies or experiments focused on effective diagnostics of Metabolic syndrome, Mild Cognitive Impairment, hypertension, or frailty within suitable analytical methods.

We used these experiences to support the medical diagnostics in the case of Hypothyroidism disease. The analytical process was interactive and iterative. The medical expert proposed the initial set of research questions (tasks). This set was continuously updated based on obtained results and their verification. The success of the proposed solution was evaluated not only by the novelty and correctness of new knowledge but also by the simple, understandable form of visualization for domain experts. All experiments were performed within the R programming language.

3.1 Data Description

The medical experts described all three datasets. Their deeper understanding required intensive communication, and the exchange of information resulted in a comprehensive overview. The first dataset contained 70 patients characterized by 32 variables like older than 50 years, with hypertension and diagnosed hypothyroidism (Table 1 and Table 2).

Table 1. Description of the numerical variables (dataset 1).

Variable	Minimum	Maximum	Median (interquartile range)	Mean (standard deviation)
TSH (mlU/L)	0.2	10.10	2.1 (1.25)	2.37 (1.55)
Age (years)	48	86	63.5 (13.75)	64.7 (9.69)
Menop (years)	0	35	0 (13.75)	7.45 (10.05)
Hy dur (years)	1	20	9 (8)	9.91 (5.35)
All drugs (count)	0	10	3 (3)	3.67 (2.23)
BMI (kg/m^2)	19	44	29 (6)	29.41 (4.92)
Wc (cm)	70.00	135.00	103.50 (13.75)	103.50 (12.60)
CRP (mg/L)	0.40	17.80	2.30 (1.47)	2.68 (2.41)

Erythro	3.68	5.83	4.72 (0.55)	4.76 (0.49)
Hct	0.297	1.74	0.43 (0.04)	0.47 (0.22)
Hgb (g/L)	90	173	144 (16.5)	144.10 (13.34)
MCV (fl)	72.9	101	90.00 (7.65)	90.09 (5.25)
F Glu (mmol/L)	4	16.9	5.6 (1.87)	6.46 (2.46)
Creatin (umol/L)	55	120	80 (16.75)	79.81 (14.18)
Gfr (ml/min/1.73m^2)	48	135	74.5 (22.25)	78.67 (15.27)
Chol (mmol/L)	2.1	8.1	5.25 (1.57)	5.317 (1.26)
LDL (mmol/L)	1	5.6	2.95 (1.6)	2.93 (1.10)
HDL (mmol/L)	0.7	10	1.4 (0.5)	1.523 (1.15)
Triglycer (mmol/L)	0.51	6.56	1.79 (1.10)	1.94 (1.09)
Dm dur (years)	0	15	0 (0)	1.84 (4.09)
All dg (count)	0	6	2 (1)	2.28 (1.34)
Cogn	0	16	8 (4)	6.81 (3.35)

Table 2. Description of the categorical variables (dataset 1).

Variable	Value	Number of values (%)
Gender	1 - Man	30 (42.89 %)
	2 - Woman	40 (57.14 %)
Hipolip drugs	0 – No	45 (64.29 %)
	1 - Yes	25 (35.71 %)
Phys act	0 – No	40 (57.14 %)
	1 - Yes	30 (42.86 %)
dm 2	0 – No	53 (75.71 %)
	1 - Yes	17 (24.29 %)

coronar	0 – No	61 (87.14 %)
	1 - Yes	9 (12.86 %)
fibril altri	0 – No	66 (94.29 %)
	1 - Yes	4 (5.71 %)
myocard	0 – No	62 (88.57 %)
	1 - Yes	8 (11.43 %)
walking	0 – No	19 (27.14 %)
	1 - Yes	51 (72.86 %)
Anxy depr	0 – 2 (Normal)	41 (58.58%)
	3 – 5 (Mild)	13 (18.57%)
	6 – 8 (Moderate)	12 (17.14%)
	9 – 12 (Severe)	4 (5.72%)

The second dataset contained 197 records about women between 50 – 59, characterized by 22 variables (Table 3 and Table 4).

Table 3. Description of the numerical variables (dataset 2).

Variable	Minimum	Maximum	Median (inter-quartile range)	Mean (standard deviation)
age (years)	50	59	52 (1.96)	52.6 (3)
FGlu (mmol/L)	4	13	5.50 (0.9)	5.723 (0.91)
TG (mmol/L)	0.42	4.49	1.68 (0.52)	1.77 (0.55)
Cho (mmol/L)	3.5	8.2	6.1 (1.7)	5.94 (0.98)
LDL (mmol/L)	1.02	5.79	3.00(1.72)	3.07 (1.03)
HDL (mmol/L)	0.80	7.20	1.20 (0.49)	1.45 (0.85)
Cre (mmol/L)	43	110	78 (19)	78.4(12.07)
GFR (ml/min/1.73m^2)	40.91	143.29	69.99 (19.04)	72.56 (14.12)
CRP (mg/L)	0.20	11.50	2.80 (2)	3.07 (1.49)

Hb (g/L)	106	155	137 (8)	135.4 (8.67)
Wc (cm)	78	120	90 (11)	90.57 (8.69)
BMI (kg/m^2)	21.11	38.28	26.89 (5.32)	27.56 (3.89)
chdi	1	10	5 (2)	4.91 (1.51)

Table 4. Description of the categorical variables (dataset 2).

Variable	Value	Number of values (%)
Hypdu	<5	28 (14.21 %)
	>10	56 (28.43 %)
	>5	113 (57.36 %)
DGDM	No	147 (74.62 %)
	Yes	50 (25.38 %)
CHD	No	184 (93.4 %)
	Yes	13 (6.6 %)
CoHD	No	181 (91.88 %)
	Yes	16 (8.12 %)
cogn	No	172 (87.31 %)
	Yes	25 (12.69 %)
depr	No	27 (13.71 %)
	Yes	170 (86.29 %)
psy	No	195 (98.98 %)
	Yes	2 (1.02 %)
sta	No	91 (46.19 %)
	Yes	106 (53.81 %)
MeSy	No	67 (34.01 %)
	Yes	130 (65.99 %)

The last dataset also contained women (135 records) but between 60 – 89, characterized by 20 variables (Table 5 and Table 6).

Table 5. Description of the numerical variables (dataset 3).

Variable	Minimum	Maximum	Median (inter-quartile range)	Mean (standard deviation)
age (years)	60	89	72 (10)	71.21 (6.46)
bmi (kg/m^2)	14.33	47.05	30.49 (6.67)	30.89 (4.97)
wc (cm)	50	148	98 (16)	98.39 (12.76)
glu_f (mmol/L)	3.9	16.2	5.7 (1.6)	6.24 (1.85)
chol (mmol/L)	2.9	9.3	5.96 (1.65)	5.96 (1.22)
ldl (mmol/L)	1.2	6.8	3.68 (1.4)	3.68 (1.10)
hdl (mmol/L)	0.8	2.3	1.5 (0.4)	1.47 (0.30)
Trig (mmol/L)	0.6	7.0	1.7 (0.85)	1.76 (0.91)
cre (mmol/L)	42	205	66 (17)	69.78 (22.96)
gfr (ml/min/1.73m^2)	24	191	86 (38.5)	86.32 (27.51)
hb (g/L)	68	158	133.4 (12.5)	133.4 (11.81)
erit	2.7	5.87	4.57 (0.42)	4.54 (0.42)
hct	0.22	0.48	0.41 (0.03)	0.41 (0.03)
som_com	1	8	3 (2)	3.48 (1.61)

Table 6. Description of the categorical variables (dataset 3).

Variable	Value	Number of values (%)
gender	f - female	135 (100 %)
hyp	no	1 (0.74 % %)
	yes	134 (99.26 %)
hyp_dur	less	53 (39.26 %)
	more	79 (58.52 %)

dm	no	3 (2.22 %)
	no	102 (75.56 %)
	yes	33 (24.44 %)
chd-nyha	higher	1 (0.74 %)
	lower	6 (4.44 %)
coron	no	128 (94.81 %)
	no	123 (91.11 %)
	yes	12 (8.89 %)

The first step provided by the data analysts was to explore data characteristics through suitable visualization techniques like histograms or within correlation analysis or statistical tests. All outputs were consulted within the expert to find possible adequate information for the research task solving. Also, we used them to identify potential outliers or incorrect or incomplete patient records. Based on a relatively large total number of variables, we present only one figure as an illustration. Also, we can state that sometimes it is necessary to generate them iteratively based on the expert's recommendations (description of axes, accurate scales).

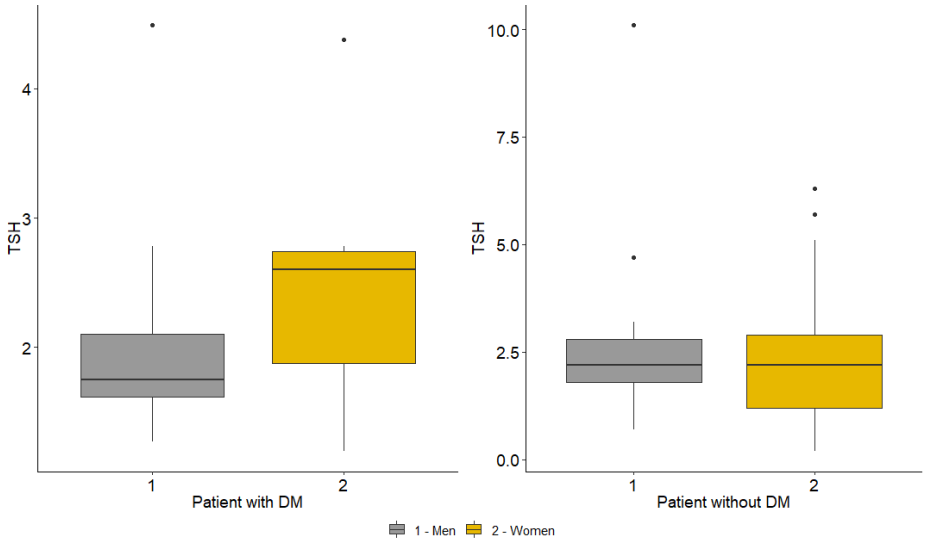

Fig. 2. Boxplots of TSH range in patients with DM and without DM by gender.

Figure 2 shows the range of TSH in patients with Diabetes mellitus (DM) and patients without Diabetes mellitus separately for men and women. The analysis shows

that the average values of TSH in patients with DM and without DM are about the same. The average value of TSH in men with Diabetes mellitus is 2.09 while in men without Diabetes mellitus is a little higher at 2.59. The analysis shows that the average value of TSH in women with Diabetes mellitus is higher than in men in the same group.

In the case of numerical variables, we performed a correlation analysis:

- strong correlation: *Creatinine* (a waste product in a blood, which is removed from blood by kidneys) – *Glomerular filtration rate* in the second dataset and *Haemoglobin – Hematocrit, Cholesterol and LDL* in the third dataset.
- moderate correlation: *Cholesterol and LDL* and *BMI – Waist circumference* in the first and second datasets; *All drugs – All diagnoses, Age – Menopause* in the first dataset; *Cholesterol – waist circumference* in the second *and Erythrocytes – Hematocrit, Haemoglobin – Erythrocytes, Creatinine – Glomerular filtration rate, BMI – Waist circumference, Age – Glomerular filtration rate.*

In the case of nominal variables, we applied the Pearson's Chi-squared test (a dependency exists):

- 1st dataset: *Gender – Physical activity, Hipolip drugs – diabetes mellitus 2, Physical activity – Walking, Cardiomyopathy – Fibrillation atriorum.*
- 2nd dataset: *almost half of the variables.*
- 3rd dataset: *Gender with Hypertension, Hypertension duration, Diabetes mellitus, Stages of chronic heart disease* and *coronary artery disease, Hypertension duration – Diabetes mellitus, Hypertension.*

Finally, we performed selected statistical tests within numerical variables. We started with the Shapiro-Wilk normality test: *Waist circumference, Erythrocytes, Mean corpuscular volume, Creatin, Cholesterol,* and *LDL*. Based on these results, we choose a statistical test – if variables were normally distributed, we used the Two-Sample t-test (Welch test), but if variables were not normally distributed, we used a Mann-Whitney test. The results show that the dependency exists between almost all the variables besides combinations: *Duration of menopause* with *Hematocrit, HDL, Triglycerides, C-reactive protein, All drugs, All diagnoses.*

Some of these results have a logical basis like total cholesterol and low-density lipoproteins or hypertension and its duration. The expert confirmed some of them based on the existing knowledge base and her experience. In general, we used them for features reduction, improving the data quality.

3.2 Task 1

The first experiment aimed on investigation of the possible differences in variable TSH (*test for thyroid stimulating hormone*) according different cut-off values for *Glomerular filtration rate*, such as > = 60 or <60 and > = 80 (Figure 3, right) or < 80 (Figure 3, left).

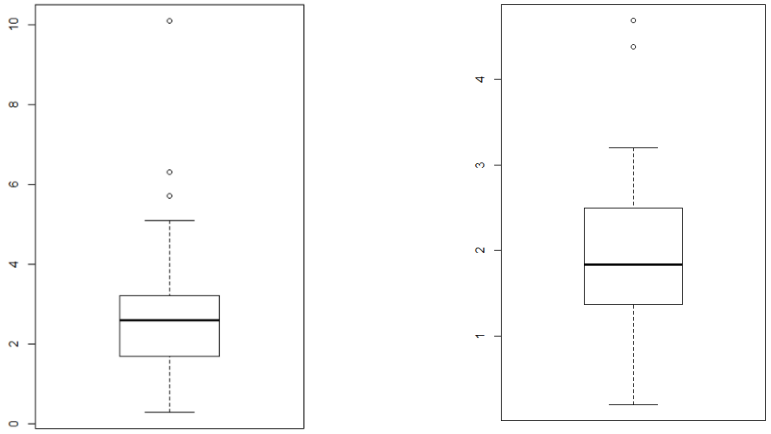

Fig. 3. The comparison of TSH distribution according to the Gfr cut-off value 80.

All combinations were evaluated by the expert to investigate a sign of mild kidney impairment. It means that the kidney is not able to excrete enough waste substances and redundant water from the blood and it is very often in older people with hypertension or diabetes.

3.3 Task 2

The second task focused on the investigation of the possible differences in other input variables according to the following condition: below the reference range (left), reference range (middle), and above the reference range of the *TSH* variable (right). The reference range of *thyroid stimulating hormone* is 0.46 - 4.68 lU/ml). As an illustration, we present a comparison of gender balance (Figure 4).

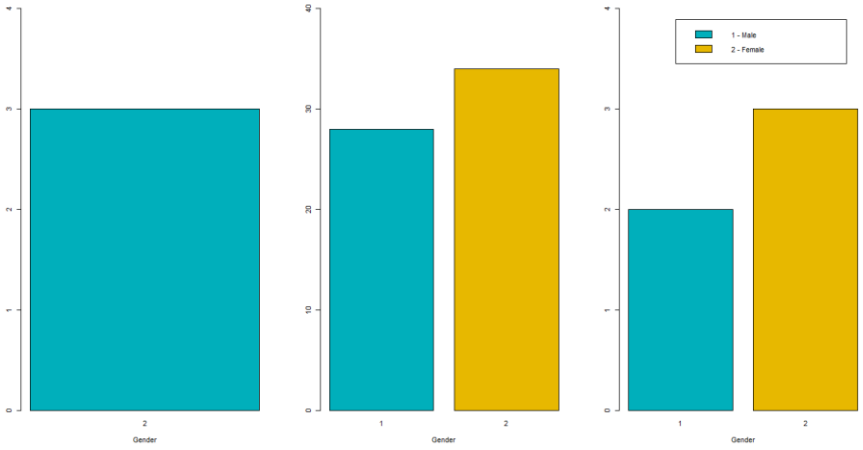

Fig. 4. The comparison of gender distribution according to three TSH levels.

The set of graphs showed several interesting findings evaluated by the expert. If *TSH* is below the reference range, it can indicate hyperthyroidism associated with more cardiac arrhythmias. But if *TSH* is within the reference range and above the reference range, it indicates hypothyroidism, and it may be related to the decreased renal function (low *Gfr*) and is known to be associated with higher *cholesterol, LDL* and *triglycerides*. Paradoxically, due to decreased renal function and associated frailty, is tends to be lower *BMI*.

3.4 Task 3

In 3rd task, we investigated collaborative possible trends between TSH and other variables like *age, glomerular filtration rate, waist circumference, BMI* (Figure 5), *total cholesterol, LDL, HDL, triglycerides, fasting blood glucose* and *c-reactive protein.* The results showed that *TSH* values increase as the metabolism increases until the point when the level of insulin resistance becomes high. It is the case at high *BMI* values and very high waist circumference values indicating obesity. That is also following with increased *triglycerides* and *LDL* values. Or when renal function (*Gfr*) becomes decreased - at the cut-off of 60, then because of malnutrition *BMI* may fall. But a high level of insulin resistance (the barrier for the action of insulin on cells – that is slowing down metabolism) causes that TSH trends become decreasing.

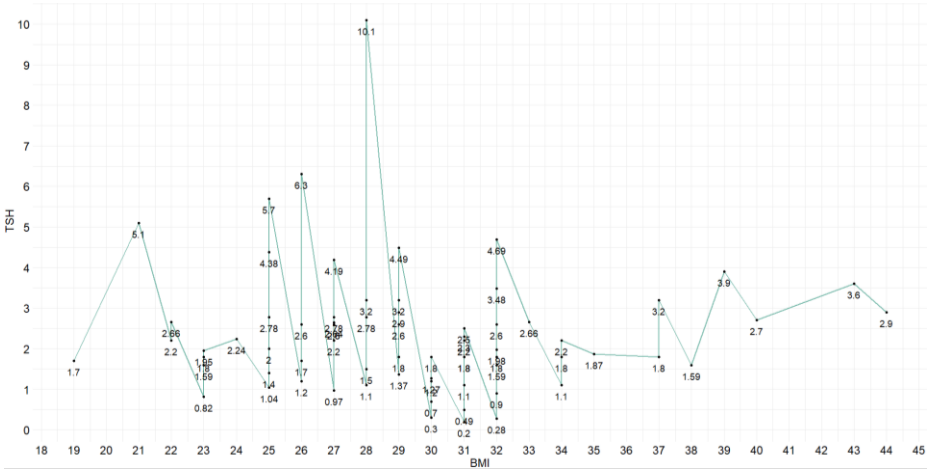

Fig. 5. Line chart visualizing a relation between variables TSH and BMI.

3.5 Task 4

This task was specified based on the outputs from previous ones. We aimed to estimate the significance of differences trough suitable statistical tests like Shapiro-Wilk normality test, two-sample t-test (Welch test), or Mann-Whiney test. The results confirmed the expert's expectations like *TSH* increases in parallel with increasing *age* and decreasing renal function (*Gfr*). Paradoxically when *TSH* increases, the variables

BMI, *waist circumference*, and *fasting blood glucose* decrease. If *TSH* decreases in parallel with reducing *haemoglobin*, *LDL*, and *triglycerides*, these factors are in accordance with hypermetabolism.

3.6 Task 5

The fifth task dealt with an experimental calculation of new cut-off values for variables *TSH* according to several target binary conditions like *Gfr* (< 60 , >= 60), *BMI* (< 30 , >= 30) or (< 25 , >= 25), *waist circumference* (< 80 , >= 80) or (< 88 , >= 88), *fasting blood glucose* (< 6.1 , >= 6.1), etc. For this purpose, we used Receiver operating characteristics (ROC) as simple understandable visualisation technique for the domain expert. The ROC analysis showed that hypothyroid patients had significantly lower *Gfr* (median 66.5, interquartile range 64.2-72.5) than *euthyroid* ones (median 78, interquartile range 68-91), as well as *BMI* (median 25.5, interquartile range 23-27 vs median 29, interquartile range 26-32).

Fig.6 shows the cut-off value 2.55 for *Gfr* indicating patients with mild renal impairment (AUC 0.62, sensitivity 0.53, specificity 0.77). In hypertensive patients, an increase in *TSH* values, already within the reference range, contributes to variations in cardiovascular risk factors if there is a mild renal impairment.

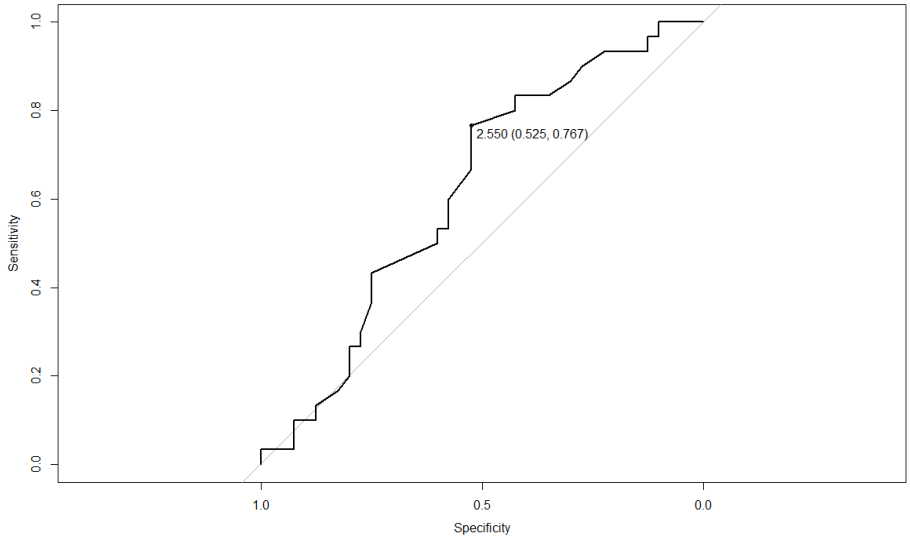

Fig. 6. ROC curve of variables TSH and Gfr.

The second ROC curves were generated from the new sample combining the second and third datasets. Anthropometric measures of obesity moderately depend on good renal function (*Gfr* >90) and obviously on other factors such as age, the time after menopause. Many patients with *Gfr* >90 (good renal function) have high cholesterol (Chol >6). Patients with older age of around 76 have a low renal function (*Gfr* <45). *BMI* indicating overweight (26-29) well correlates with low renal function but is

not exclusive for low renal function, because it may also be a characteristic of older patients with preserved renal function (*Gfr* >90). It means that patients have not been obese and did not get diabetes (15.7%) at higher rates. The relationship between low *Gfr* and *waist circumference* is a complex one because *waist circumference* is increased moderately.

3.7 Task 6

Finally, we decided to apply multiple logistic regressions on the joined dataset with women older than 50 years; one example is visualized in Fig.7.

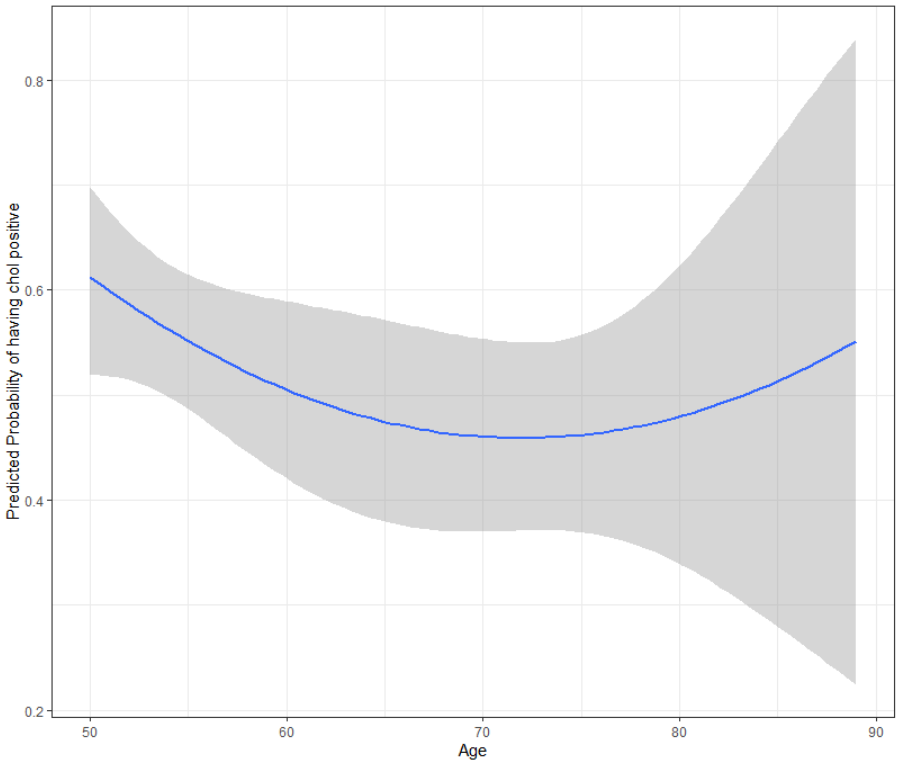

Fig. 7. Logistic regression of the variable age and total cholesterol.

The results show that there are two groups of women with the diagnosis of *hypertension*. The first group is younger (50 - 59 years old) and a higher prevalence of *diabetes*. The presence of *diabetes*, in combination with h*ypertension*, in the *age* of 54-55 years, is associated with a mildly reduced renal function (*Gfr* = 89-60 or 59-45). These women are usually overweight – *BMI* 26-30 (*BMI*>26. and *waist circumference* >100). The second group represents older women, and their age is higher than 65 and may be divided into two subgroups, according to the *Gfr* range values. The first subgroup is women *age* around 65 with good renal function (*Gfr* >90), but women in this

subgroup may be obese, their *BMI* may be higher than 30 or with the normal weight (*BMI*<26). The second subgroup is women age around 76 with a low renal function (*Gfr* <45). It can associate with a low weight (*BMI* <26), which indicates the frailty status, because of *low renal function* and high *age*. *Low renal function* is an insulin-resistant state and is associated with high *waist circumference* (>90), which is, typically, a characteristic of obesity for women with *age* from 50 to 60 years - years after post-menopause. But if women are in *age* higher than 76 years, their high *waist circumference* falls more intensively depending on advanced age that it increases with a drop of *Gfr*.

4 Conclusion

The practicing doctors mention two significant limitations of their decision/diagnostic process: existing literature and case studies contain a large volume of information, and doctors often lack time to find, consider, and use it all. The second constraint is sometimes too generic method, procedure, or recommendation not focused on the specific patient or situation. This situation is changing step by step with the deployment of electronic health records and suitable analytical approaches. Correctly set collaboration between medical experts and data analysts can answer the age-old question: what is truly best for each patient?

Mining medical data requires intensive and effective collaboration between medical experts and data analysts. We experimentally solved six research tasks through the analytical process dealt with thyroid disease and hypothyroidism. Our results confirmed an expectation that the incidence of thyroid dysfunction in diabetic patients is higher than in the general population [26]. In general, we can say that we generated a relatively large volume of results that were evaluated by the expert. This process was improved step by step by the common understanding and previous experience. The next step should be creating the knowledge base related to the thyroid disease as a basis for future development. A useful clinical decision support system will represent the final stage. But this development is not a simple task. It requires more resources and a broader collaboration of the research teams around the world, including data analysts, medical experts, and primary care doctors.

We have several recommendations for effective collaboration within domain experts: at first, have clearly defined tasks that reduce the risk of mutual misunderstanding; regular communication and joint decision; the easy-to-understand form of results and detailed explanations; state-of-the-art analysis to identify suitable methods; verification of the obtained results with existing literature or expert knowledge.

Acknowledgements. The work was partially supported by The Slovak Research and Development Agency under grants no. APVV-16-0213 and no. APVV-17-0550.

References

1. Balogh, E.P., Miller, B.T., Ball, J.R.: Improving Diagnosis in Health Care. The National Academies Press, Washington, DC (2015).
2. Onder, G., Palmer, K., Navickas, R., et al.: Time to face the challenge of multimorbidity. A European perspective from the joint action on chronic diseases and promoting healthy ageing across the life cycle (JA-CHRODIS). Eur. J. Intern. Med. 26, 157 -159 (2015).
3. Paul, L., Jeemon, P., Hewwit, J., McCallum, L., Higgins, P., Walters, M., et al.: Hematocrit Predicts Long-Term Mortality in a Nonlinear and Sex-Specific Manner in Hypertensive Adults. Hypertension 60(3), 631 - 638 (2012).
4. Silva, N.O., Ronsoni, M.F., Colombo, Bda.S., Correa, C.G., Hatanaka, S.A., et al.: Clinical and laboratory characteristics of patients with thyroid diseases with and without alanine aminotransferase levels above the upper tertile – Cross-sectional analytical study. Arch Endocrinol Metab 60(2), 101-107 (2016).
5. Ahmed, O. M., Ahmed, R.G.: Hypothyroidism. A new Look at Hypothyroidism. InTech, 1-20 (2012).
6. Cojić, M., Cvejanov-Kezunović, L.: Subclinical Hypothyroidism – Whether and When to Start Treatment?Open Access Maced J Med Sci 5(7), 1042 - 1046 (2017).
7. Carlé, A., Pedersen, I.B., Knudsen, N., Perrild, H., Ovesen, L., Laurberg, P.: Hypothyroid symptoms and the likelihood of overt thyroid failure: a population-based case-control study. Eur J Endocrinol 171(5), 593 - 602 (2014).
8. Girardi D., Kueng J., Holzinger A.: A Domain-Expert Centered Process Model for Knowledge Discovery in Medical Research: Putting the Expert-in-the-Loop. Brain Informatics and Health. LNAI 9250, 389 - 398 (2015).
9. Holzinger, A.: Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inf. 3, 119 - 131 (2016).
10. Mao, Y., Wang, D., Muller, M.J., Varshney, K.R., Baldini, I., et al.: How Data Scientists Work Together with Domain Experts in Scientific Collaborations: To Find the Right Answer or to Ask the Right Question? Proceedings of the ACM on Human-Computer Interaction 237 (2019).
11. Jeanquartier F., Holzinger A. On Visual Analytics and Evaluation in Cell Physiology: A Case Study. In: Cuzzocrea A., Kittl C., Simos D.E., Weippl E., Xu L. (eds) Availability, Reliability, and Security in Information Systems and HCI. CD-ARES 2013. Lecture Notes in Computer Science, Vol 8127. Springer, Berlin, Heidelberg (2013).
12. Ioniță, I., Ioniță, L.: Prediction of Thyroid Disease Using Data Mining Techniques. In: BRAIN. Artificial Intelligence and Neuroscience 7(3), 115 - 124 (2016).
13. Sidiq, U., Aaqib, S.M., Khan, R.A.: Diagnostic of Various Thyroid Ailments using Data Mining Classification Techinques. International Journal of Scientific Research in Computer Science, Engineering and Information Technology 5 (1), 131-135 (2019).
14. Gurram, D., Narasinga Rao, M.R.: A comparative study of support vector machine and logistic regression for the diagnosis of thyroid dysfunction. International Journal of Engineering & Technology 7(1.1), 326 - 328 (2018).
15. Cox, V.: Exploratory Data Analysis. In: Translating Statistics to Make Decisions. Apress, Berkeley, CA (2017).
16. Komorowski, M., Marshall, D.C., Salciccioli, J.D., Crutain, Y.: Exploratory Data Analysis. Secondary Analysis of Electronic Health Records. 185 - 203 (2016).
17. Franke, T.M., Ho, T., Christie, CH.A., The Chi-Square Test: Often Used and More Often Misinterpreted. American Journal of Evaluation 33(3), 448 - 458 (2012).

18. Ghasemi, A., Zahedias, S.: Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. Int J Endocrinol Metab. 10(2), 486 - 489 (2012).
19. Oztuna, D., Elhan, A.H., Tuccar, E.: Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. Turkish Journal of Medical Sciences 36(3), 171 - 176 (2006).
20. Rice, J.A.: Mathematical Statistics and Data Analysis (3rd ed.). Duxbury Advanced (2006).
21. Mann, H.B.; Whitney, D.R.: On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other". Annals of Mathematical Statistics 18(1), 50 - 60 (1947).
22. Hajian-Tilaki, K.: Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian J Intern Med 4(2), 627 - 635 (2013).
23. Rey deCASTRO, B.: Cumulative ROC curves for discriminating three or more ordinal outcomes with cutpoints on a shared continuous measurement scale. PLoS ONE 14 (8), (2019).
24. Carter, J.V., Pan, J., Rai, S.N, Galandiuk, S.: ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. Surgery 159(6),1638 - 1645 (2016).
25. Tolles, J., Meurer, W.J.: Logistic Regression Relating Patient Characteristics to Outcomes. JAMA 316 (5), 533 - 534 (2016).
26. Mohamed, G.A., Elsayed, A.M.: Subclinical hypothyroidism ups the risk of vascular complications in type 2 diabetes. Alexandria Journal of Medicine 53(3), 285 - 288 (2017).