On Patient's Characteristics Extraction for Metabolic Syndrome Diagnosis: Predictive Modelling Based on Machine Learning

František Babič[1], Ljiljana Majnarić[2], Alexandra Lukáčová[1],
Ján Paralič[1], and Andreas Holzinger[3]

[1] Technical University of Košice, Faculty of Electrical Engineering and Informatics,
Department of Cybernetics and Artificial Intelligence, Letná 9/B, 042 00 Košice, Slovakia
[2] Josip Juraj Strossmayer University, Osijek, Croatia
[3] Medical University Graz, Institute for Medical Informatics, Statistics and Documentation
Research Unit HCI, Auenbruggerplatz 2/V, A-8036 Graz, Austria
{frantisek.babic,alexandra.lukacova,jan.paralic}@tuke.sk,
ljiljana.majnaric@hi.t-com.hr, a.holzinger@hci4all.at

Abstract. The work presented in this paper demonstrates how different data mining approaches can be applied to extend conventional combinations of variables determining the Metabolic Syndrome with new influential variables, which are easily available in the everyday physician`s practice. The results have important consequences: patients with the Metabolic Syndrome can be recognized by using only some, one, or none of the conventional variables, when replaced with some other surrogate variables, available in patient health records, making diagnosis feasible in different work environments and at different time points of patient care. In addition, the results showed that there is a large diversity of patient groups, much larger than it was supposed earlier on when their identification was based on the conventional variables approach, indicating the underlying complexity of this syndrome. Finally, the discovered novel variables, indicating yet unknown pathogenetic pathways can be used to inspire future research.

Keywords: biomedical data mining, metabolic syndrome, machine learning.

1 Introduction

Metabolic Syndrome (MetSy) is a well-known cluster of cardiovascular risk factors, components of which include central obesity (abdominal fat accumulation), impaired glucose tolerance, hypertension and atherogenic dyslipidemia, defined as increased serum triglycerides (TG) and decreased HDL-cholesterol (HDL) [1]. It is based on continuous rather than dichotomous variables and diagnostic criteria or cut-off values vary between studies and recommendations. This combined disorder is common in modern society, encompassing almost a quarter of the world`s adult population.

Insulin resistance, a blockade of insulin action in peripheral tissues, and abdominal obesity, are considered as the key mechanisms. However, novel findings also indicate

M. Bursa et al. (Eds.): ITBAM 2014, LNCS 8649, pp. 118–132, 2014.

the role of microcirculation and endothelial cells dysfunction and of increased oxidative stress and inflammation, as well [2, 3]. Hyperhomocysteinemia, a marker of impaired remethylation reaction, and the neuroendocrine stress axis, have also been implicated in the pathogenesis of this syndrome [4, 5]. The fact that this syndrome can also appear in frail elderly persons, not only in obese ones, and that its manifestations can differ between men and women, indicate that there could be a variety of patient groups and heterogeneity of underlying mechanisms [6].

By applying different data mining approaches on the large dataset prepared in a way that collected parameters from many aspects describe the health status of patients, we wanted to find out whether there are some important extensions to the classical definition of the Metabolic Syndrome, to add value to clinical reasoning, or to map novel variables and pathways that can be used to direct future research. Experiments were performed by using R software with the installed package "OptimalCutpoints" [7] and data mining workbench SPSS Clementine 10.1.

2 Related Work

Information on cardio-vascular (CV) risk factors and their clustering is available mainly from large prospective population studies which are known to provide the highest level of evidence [8]. Therefore, these factors have rarely been an object for predictive modelling, based on machine learning methods, as these methods are used for solving medical uncertainties. However, evidence is growing on the existence of novel CV risk factors, not yet proved as biomarkers [9]. In addition, the awareness is increasing, that these factors may vary in the composition and intensity, depending on the population they were drawn from, or socio-demographic characteristics of the examined population group [10]. For all these reasons, the existing CV risk assessment scores and prediction support systems become increasingly insufficient to meet the needs for prediction in different real-world situations [11]. This requirement is a challenge for the application of machine intelligence [12]. In a recently published work, authors used Bayesian networks for predicting the Metabolic Syndrome from a dataset composed of a total of 18 attributes and 1193 subject records, collected in the Yonchon County, Korea [13].

A further related work was also done in the Far East: A group in Thailand [14] explored the relationship between hematological parameters and glycemic status in the establishment of a quantitative population-health relationship model for the identification of individuals with or without diabetes mellitus. For this purpose they ran a cross-sectional study of 190 participants which they classified into three groups based on their blood glucose levels. Hematological (white blood cell (WBC), red blood cell (RBC), hemoglobin (Hb) and hematocrite (Hct)) and glucose parameters were used as input variables while the glycemic status was used as output variable. They applied support vector machine (SVM) and artificial neural network (ANN) as machine learning approaches for identifying the glycemic status and applied association analysis (AA) for the knowledge discovery process of health parameters that frequently occur together. A major barrier for the realization of personalized medicine is in the identification of biomarkers [15], [16].

A further interesting recent work was done by the University Hospital Zurich [17]. The authors described a two-stage strategy for the discovery of serum biomarker signatures corresponding to specific cancer-causing mutations and its application to prostate cancer in the context of the commonly occurring phosphatase and tensin homolog (PTEN) tumor-suppressor gene inactivation. The authors identified 775 N-linked glycoproteins from sera and prostate tissue of wild-type and Pten-null micea and the resulting proteomic profiles were analyzed by machine learning methods, i.e. random forests, to build predictive regression models for tissue PTEN status and diagnosis and grading of a prostate carcinoma (PCa).

3 Data Understanding

Data were collected in a family practice located in an urban area of the town of Osijek, the north-eastern part of Croatia, the region known by high prevalence of CV and other chronic diseases, higher than average for Croatia. A total number of 93 subjects, 35 male and 58 female, 50-89 years old (median 69), gave their consent and were included in the study. Data, only low-cost, easily available parameters were collected systematically, that means in a way to determine the health status of examined patients by many aspects. A large proportion of these collected parameters are routinely collected data from patients' health records. Nominal parameters indicate age and sex, diagnoses of the main groups of chronic diseases, information on drugs use and anthropometric measures. A number of laboratory tests were also performed, indicating the main age-related pathophysiologic changes, including information on: inflammation, the nutritional status, the metabolic status, chronic renal impairment, latent infections, humoral (antibody-mediated) immunity and the neuroendocrine status (Table 1). As performed on the small sample, this presented method is not likely to allow for definite answers, but may be used as the first step approach for solving some medical uncertainties. In this sense, this method is likely to allow new variables and hidden relationships, not easily detectable in clinical studies, to be mapped in otherwise unknown input space. Results got by this method are to be further tested.

MetSy database contains 93 patients' records including 61 medical variables and one variable describing target diagnosis called Metabolic Syndrome. 60 patients in the analyzed dataset have diagnosed syndrome and 33 do not. One of the traditional ways to determine this diagnosis is to use the IDF (International Diabetes Federation) definition including following expert rules using combination of major input variables and their values [18]. Meaning of particular variables can be found in Table 1 below.

Criteria for female: (w/h > 0.85 *OR* BMI > 30) *AND* at least 2 out of the 4 following conditions must be fulfilled Hypertension (yes) *OR* TG > 1.7 *OR* HDL < 1.3 *OR* fasting glucose ≥ 5.6 *OR* Diabetes mellitus (yes).

Criteria for male: (w/h > 0.9 *OR* BM I> 30) *AND* at least 2 out of the 4 following conditions must be fulfilled: Hypertension (yes) *OR* TG > 1.7 *OR* HDL < 1.0 *OR* fasting glucose ≥ 5.6 *OR* Diabetes mellitus (yes)

Table 1. Description of all variables included in experiments

Variable code	Variable description
age	Age (years)
sex	M=Male, F=Female
Hyper	Hypertension (yes, no)
DM	Diabetes mellitus (yes, IGT=Impaired glucose tolerance, No)
F Glu	Fasting blood glucose (mmol/L)
HbA1c	Glycosilated Haemoglobin (%) - showing average blood glucose during last three months
Chol	Total Cholesterol (mmol/L)
TG	Triglycerides (mmol/L)
HDL	HDL-cholesterol (mmol/L)
Statins	Therapy with statins (yes,no)
Anticoag	Therapy with anticoagulant/antiaggregant drugs (yes,no)
CVD	Cardiovascular diseases as myocardial infarction, angina, history of revascularisation, stroke, transient ischaemic cerebral event, peripheral vascular disease (yes, no)
BMI	Body Mass Index (kg/m^2)
w/h	Waist/hip ratio
Arm cir	Mid arm circumference (mm)
skinf	Triceps skinfold thickness (mm)
gastro	Gastroduodenal disorders as gastritis, ulcer (yes,no)
uro	Chronic urinary tract disorders (yes,no) - recurrent cystitis in women, symptoms of prostatism in men
COPB	Chronic obstructive pulmonary disease (yes,no)
Aller d	Allergy (Rhinitis and/or Asthma) (yes,no)
dr aller	Drugs allergy (yes, no)
analg	Therapy with analgetics/NSAR (yes,no)
derm	Chronic skin disorders as chronic dermatitis, dermatomycosis (yes,no)
neo	Malignancy (yes,no)
OSP	Osteoporosis (yes, no)
Psy	Neuropsychiatric disorders as anxiety/depression, Parkinson`s disease, cognitive impairments (yes,no)
MMS	Mini Mental Score – test for screening on cognitive dysfunction, Max Score =30, Score <24 indicates cognitive impairment
CMV	Cytomegalovirus specific IgG antibodies (IU/ml)
EBV	Epstein-Barr virus specific IgG (IU/ml)
HBG	Helicobacter pylori specific IgG (IU/ml)
HPA	Helicobacter pylori specific IgA (IU/ml)
LE	Leukocytes Number x10^9/L
NEU	Neutrophils % in White Blood Cell differential
EO	Eosinophils % in White Blood Cell differential
MO	Monocytes % in White Blood Cell differential

Table 1. (*continued*)

LY	Lymphocytes % in White Blood Cell differential
CRP	C-reactive protein (mg/L)
E	Erythrocytes number x10^{12}/L
HB	Haemoglobin (g/L)
HTC	Haematocrite (erythrocyte volume blood fraction)
MCV	Mean cell Volume (fL)
FE	Iron (g/L)
PROT	Total serum proteins (g/L)
ALB	Serum albumin (g/L)
clear	Creatinine clearance (ml/s/1.73m^2)
HOMCIS	Homocistein (µmol/L)
ALFA1	Serum protein electrophoresis (g/L)
ALFA2	Serum protein electrophoresis (g/L)
BETA	Serum protein electrophoresis (g/L)
GAMA	Serum protein electrophoresis (g/L)
RF	Rheumatoid Factor level (IU/ml)
VITB12	Vitamin B12 (pmol/L)
FOLNA	Folic acid (mM/L)
INS	Insulin (µIU/L)
CORTIS	Cortisol in the morning (nmol/L)
PRL	Prolactin in the morning (mIU/L)
TSH	Thyroid-stimulating hormone (IU/ml)
FT3	Free triiodothyronine (pmol/L)
FT4	Free thyroxine (pmol/L)
ANA	Antinuclear antibodies (autoantibodies) (µIU/ml)
IGE	IgE (kIU/L)
MetSy	0 – without diagnosed Met Sy, 1 – diagnosed MetSy

4 Experiments

In our experiments, we focused on the possibility to use selected methods from machine learning theory to provide the answers on specified medical questions. In cases where not only classification or prediction accuracy is important, both patterns need also to be understood by human experts, methods extracting patterns in form of decision trees have proved to be very successful and effective. Decision trees provide a classification structure and can be easily transformed also into form of decision rules. This is in contrast to classifiers like neural network models, which may provide nice classification results, but as a kind of black box. Decision tree models can be extracted by different algorithms, such as e.g. CART, or C4.5 [19]. We used decision trees for different purposes, e.g. also for selecting of most important classification attributes from predefined groups of attributes. In our experiment we used two alternative instances of algorithm C4.5: J48 implemented in Weka data mining tool and C5.0 provided by SPSS data mining software. In both cases we have tested different parameters and their values to find the combination with good precision and optimal decision ability. Because of the small number of input records we have instead of the traditional division into training and test sample used 10-fold cross validation.

Afterwards we searched for their optimal cut-off values as follows. For finding the optimal cut-off points c, which best distinguish diseased and healthy patients, we used the measure called Youden index (J) [20], defined as

$$J = \mathrm{maxc}\{\mathrm{Sensitivity}(c) + \mathrm{Specificity}(c) - 1)\} \qquad (1)$$

Its advantage is in offering the best result with respect to the maximum overall correct classification by maximizing the sum of sensitivity and specificity. The parameter c is understood as the optimal cut point [21]. The range of J is <0,1>, where the value 1 stands that all diseased and healthy patients are correctly classified and the value 0, on the contrary, means that the selected cut off point is completely ineffective [22]. The confidence level was set to 0.95. We considered the cut off values only in the case, when importance of particular variable was statistically significant (i.e. $p<0.05$). We used student's unpaired t-test to for this purpose.

4.1 Decision Trees for MetSy Determination

We performed different experiments, starting with the whole database of patients, than with the data sample including only female patients and on the other hand over the sample including only men. This division provided opportunity to identify characteristics relevant for female and male patients, respectively.

First decision tree was generated through all the records in the dataset (93 records) and it largely confirmed the original rules specified by IDF, e.g.:

IF Fglu (Fasting blood glucose > 5.4 AND HDL (HDL-cholesterol) <= 1.72 THEN MetSy = 1 (100% strength of this rule covering 44/60 records)

Strength of all decision rules is affected by sample size from which these rules were extracted. In experiments where male records were used, they included 23 patients with MetSy (out of 35 male patient records in our database) and on the other side female records included 37 patients with MetSy (out of 58 female patient records in our database).

Decision Rules for Female Patients

Next experiments were performed on divided dataset; we created two samples, one for male (35 patients) and the second for female (58 patients). We have applied the same algorithms C5.0 as in the first case, but we obtained different rules for target variable MetSy. Decision tree for female contained three variables and can be transformed in the following rules:

IF HbA1C (average level of blood glucose over the previous 3 months) > 4.41 THEN MetSy = 1 (100% force of rule within this sample 21/21)

IF HbA1C (average level of blood glucose over the previous 3 months) =< 4.41 AND TG (triglycerides) > 1.7 THEN MetSy = 1 (92.3%, 12/13)

IF HbA1C (average level of blood glucose over the previous 3 months) =< 4.41 AND TG (triglycerides) =< 1.7 and EBV (Epstein-Barr virus specific IgG) =< 20.8 then MetSy = 1 (100%, 2/2)

IF HbA1C (average level of blood glucose over the previous 3 months) =< 4.41 AND TG (triglycerides) =< 1.7 and EBV EBV (Epstein-Barr virus specific IgG) > 20.8 then MetSy = 0 (90.9%, 20/22)

Decision Rules for Male Patients

Application of C5.0 algorithm on male patients' records brought interesting finding, because generated decision tree contained only one variable for classification – FOLNA (level of Folic acid), see Fig. 1.

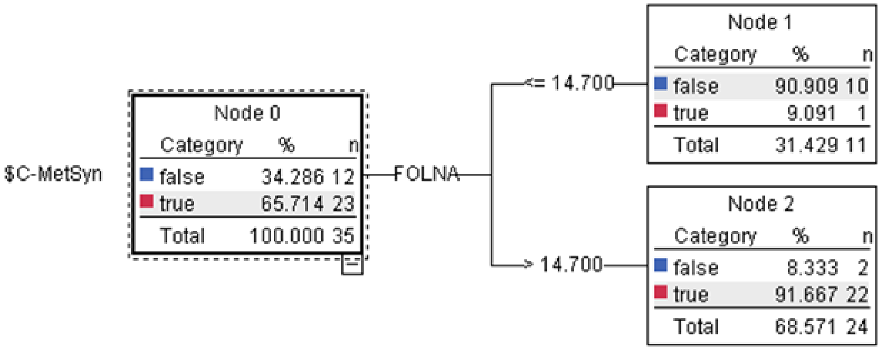

Fig. 1. Simple decision tree for men records

This simple decision rule has motivated us to try some other algorithms for generation of decision tree models. Algorithms CHAID (Chi-squared Automatic Interaction Detection) and Quest (Quick, Unbiased, Efficient Statistical Tree) brought following rules:

IF Fasting blood glucose =< 4.9 THEN MetSy = 0 (100%, 2/2)

IF Fasting blood glucose > 4.9 AND age < 70 AND age > 73 THEN Met Sy = 1 (100%, 21/21); 2 patients from age interval (70,73) hadn't diagnosed Metabolic Syndrome.

IF Fasting blood glucose > 5.9 THEN MetSy = 1 (92.3%, 12/13)

IF Fasting blood glucose =< 5.9 AND Serum protein electrophoresis ALPHA2 > 6.1 THEN MetSy = 0 (81.8%, 9/11)

IF Fasting blood glucose =< 5.9 AND Serum protein electrophoresis ALPHA2 =< 6.1 AND Mean cell Volume > 85.759 THEN MetSy = 1 (90%, 9/10)

Decision Rules without IDF Factors (Female Patients)

The second group of experiments was realized within reduced data sample, in which we have eliminated variables specified by IDF as factors causing the MetSy, e.g. TG, HDL, Fasting glucose, etc. New dataset included patient records described by 55 variables. The aim of this operation was to identify some other important variables that have strong impact on MetSy diagnosis. In the case of female, we identified following rules as interesting:

IF HbA1C (average level of blood glucose over the previous 3 months) > 4.41 THEN MetSy = 1 (the same rule as in previous experiment with all 62 variables)

IF HbA1C (average level of blood glucose over the previous 3 months) =< 4.41 AND Insulin > 27.1 THEN MetSy = 1 (100%, 6/6)

IF HbA1C (average level of blood glucose over the previous 3 months) =< 4.41 AND Insulin =< 27.1 AND Cardiovascular diseases = yes AND Cortisol in the morning > 457.6 THEN MetSy = 0 (100%, 2/2)

IF HbA1C (average level of blood glucose over the previous 3 months) =< 4.41 AND Insulin =< 27.1 AND Cardiovascular diseases = yes AND Cortisol in the morning =< 457.6 THEN MetSy = 1 (100%, 4/4)

IF HbA1C (average level of blood glucose over the previous 3 months) =< 4.41 AND Insulin =< 27.1 AND Cardiovascular diseases = no AND Drug allergy = yes THEN MetSy = 1 (75%, 3/4)

IF HbA1C (average level of blood glucose over the previous 3 months) =< 4.41 AND Insulin =< 27.1 AND Cardiovascular diseases = no AND Drug allergy = no AND Serum protein electrophoresis GAMA > 11.8 THEN MetSy = 0 (100%, 15/15)

Decision Rules without IDF Factors (Male Patients)

Reduction to the 55 variables did not produce any significant changes in the rules extractions from the records of male patients. Based on this fact, we decided to eliminate variable FOLNA from decision tree inputs in order to identify other possible important variables for MetSy determination. The resulting rules were:

IF Rheumatoid Factor level > 9.0 THEN MetSy = 0 (100%, 3/3)

IF Rheumatoid Factor level =< 9.0 AND Diabetes mellitus = IGT/yes THEN MetSy = 1 (100%, 9/9)

IF Rheumatoid Factor level =< 9.0 AND Diabetes mellitus = no AND w_h > 1.0 THEN MetSy = 0 (85.7%, 6/7)

IF Rheumatoid Factor level =< 9.0 AND Diabetes mellitus = no AND w_h =< 1.0 AND Insulin > 13.4 THEN MetSy = 1 (100%, 10/10)

4.2 Cut-Off Values for Better Characterization of Patients with MetSy

The second direction of our experiments was devoted to identification of new cut-off values for selected variables in order to evaluate their influence on MetSy determination. We focused especially on the following risk factors based on medical expert recommendations:

- Inflammation - variables: CRP, Le, Mo/Neu
- Age - years
- Renal dysfunction - variables: clear, HOMCIS
- Malnutrition - variables: alb, vitB12, folna
- The thyroid gland malfunction – variables: TSH, FT3, FT4
- hormones: PRL, CORTIS
- anemia/blood viscosity: E, HB, HTC
- anthropometric measures – malnutrition: skinf, Arm cir
- average 3-month glucose level : HbA1c

Based on the fact that MetSy has different characteristics for male and female, we performed this type of experiments over the two data samples (35M/58F) as in previous case.

The results of student's unpaired t-test indicated only variables FOLNA and HbA1c for men and MO and TSH for women as statistically significant. Optimal cut off points for these variables are presented in Table 2.

Table 2. Optimal cut-off values for identified variables (PPV - positive predicted value, NPV - negative predicted value)

	Cut-off value	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
FOLNA (M)	15.6	95.65	83.33	91.67	90.91
HbA1c (M)	4.5	39.13	100	100	46.15
MO (F)	5.5	86.5	14.3	64	37.5
TSH (F)	2.69	22.22	100	100	41.67

Next, we compared calculated cut-off values with those that were used for partition in decision trees models, but only for two attributes that appeared also in generated decision tree models; see Table 3.

Table 3. Comparison of identified cut-off values for two variables

Variable	Cut-off value in decision tree model	Cut-off value by statistical analysis
FOLNA	14.7	15.6
HbA1C	4.41	4.5

As can be seen from presented comparison, generated decision models for MetSy determination have used very similar cut-off values. On the other hand, performed statistical analysis resulted in some potentially interesting findings for medical expert to improve daily diagnostics in physician`s practice.

5 Discussion

In the first set of experiments, the target variable, i.e. patients with MetSy, diagnosed by using conventional variables, were contrasted with those ones not diagnosed with MetSy. Some interesting rules for determination of MetSy, based on using Decision trees method, were established. Results were obtained for the whole group of patients (93) and by using the whole dataset (61 variables), and separately for women (58) and men (35).

In general, the experiment performed on the whole dataset confirmed the original rules specified by IDF definition. However, when rules obtained for females were contrasted with those obtained for males, some important differences could be observed. This is likely to be in line to the existing knowledge indicating different metabolic pathways by which men vs. women attain CV diseases [2, 6]. In comparison to men, women seem to be more prone to Diabetes and use metabolic variables associated with diabetes (in our experiments indicated with the variable triglycerides), while men are prone to abdominal obesity and run CV risks through MetSy, rather than through Diabetes [2]. Our results are also in line to the experiences showing that men more frequently present with impaired fasting glucose (see experiments for men), whereas impaired glucose tolerance (indicated with nonfasting glucose levels) more frequently occurs in women [23]. The latter statement is likely to be confirmed with our experiments for women, where the variable *HbA1c* is known to cope with glucose variability, in a great part dependent on variation in nonfasting glucose levels [24]. Another variable which makes distinction between men and women is the variable *EBV*, selected in experiments for women. According to the current knowledge, this variable (indicating re-activation of the latent infection with Epstein-Barr virus), may be a marker of increased inflammation and inflammation-mediated aging of the immune system [25]. In this respect, inflammation has been recognized as a part of the MetSy [2]. It is not, however, quite clear from our experiments, whether increased or decreased specific anti-EBV IgG antibody levels represent the conditional criteria for MetSy determination in women, so further research is needed in this direction. Another disctinction between men and women was in respect to the variable *FOLNA*. Namely, rules obtained for males emphasized the role of Folic acid serum concentrations for classification of patients as to have MetSy or not (Fig. 1). In this regard, results of the recent research indicate the relevance of folate deficiency conditions for triggering oxidative-stress and apoptotic cell death, which may have implications for the development of both, impairment of insulin biosynthesis in pancreatic islet β-cells, as well as peripheral vasculopathy and insulin resistance [26, 27]. We may only speculate on the reasons why this variable is selected in men but not in women. One reason might be due to the fact that gastroduodenal disorders are more frequent in males, than in females, leading to malabsorption and folate deficiency. The second possible link between male gender, folate deficiency and MetSy, as according to the current knowledge, might be, on the contrary, due to the fact that men are much more dependent on genetics, than women, and less on environmental factors, in achieving aging and age-related diseases [28].

In this regard, it is known that folic acid is necessary, together with cobalamin (vitamin B12) and pyridoxine (vitamin B6), for maintaining some vital biological processes such as DNA methylation and that the activity of the enzymes included in these reactions are genetically controlled [26]. (The statement *Mean cell Volume>85.759*, indicating megaloblastic anemia, is complementary to folate deficiency, the main cause of this type of anemia).

When conventional variables were excluded from the dataset, some interesting rules, otherwise hidden, were obtained. In the female subject group, the variable *HbA1c*, routinely used measure of long term blood glucose control (other than measure fasting blood glucose, excluded from the experiment), was emphasized as the component of the MetSy. Significance of this variable as a measure of impaired glucose metabolism, suitable for a large scale studies, has recently been confirmed, in the study indicating this measure as a robust biomarker of mortality in diabetic patients [29]. In our experiments obtained rules confirmed the known fact on the associations between impaired glucose metabolism (indicating with *HbA1c*), hyperinsulinemia (a measure of insulin resistance) and CV diseases [30]. Moreover, our experiments provided these associations with new information. This information, for example, includes the rule stating that, in patients with CV diseases but normal blood glucose control, disturbed diurnal rhythm of the hypothalamus-pituitary-adrenal stress axis (indicated by decreased cortisol blood concentrations in the morning) can be used to select patients with MetSy. On the contrary, the preserved neuro-endocrine stress response may indicate persons who have not got MetSy. Another obtained interesting rule deals with information that the presence of *Drug Allergy*, in absence of other typical markers of MetSy, may be used to recognize, with high level probability, female patients with MetSy. This result suggests that there might be an association between genetic variations of drug-metabolizing enzymes and disturbed glucose metabolism, as it has already been established for the association between genetic variations of these enzymes and the individual`s susceptibility to cancer [31]. Our results further indicate that in female subjects not having *Drug allergy* and free from other typical markers of MetSy, normal (not decreased) serum gamma-globulins levels, indicating preserved immune system functions, may be used as a protective mechanism against MetSy development.

When conventional variables, together with the variable *FOLNA*, identified as an important one in our experiments, were excluded from the dataset, the decision rules performed in the male patient group, revealed the importance of the rheumatoid factor (variable *RF*) for determinantion of MetSy. Although *RF* positivity was found in a small number of patients in the sample, this found association confirms already known close relationships between rheumatoid arthritis and atherosclerotic CV diseases [32]. Moreover, increased CV risk, as according to the knowledge, occurs early during the course of rheumatoid arthritis (when only RF serum concentrations are increased, without visible clinical symptoms and signs of disease) and may be considered as a possible preclinical manifestation of this disease. From our results, however, it is not quite clear whether increased RF serum concentrations in men mean protection from, or predisposition for MetSy, and the nature of this found relationship requires further clarification. Other rules obtained for males further suggest that in the absence of *RF* and conventional MetSy variables, indications for MetSy, specifically in males, may be attained through the presence of Diabetes mellitus and high insulin levels (hyperinsulinemia), the latter disorder present in the absence of abdominal adiposity (indicated with *w/h =<1.0*). These rules confirm, once again, that Diabetes mellitus may have only minor role in the pathogenesis of MetSy in men and that hyperinsulinemia, the hallmark of MetSy, may exist independently of abdominal adiposity, which all together indicates that there might be different mechanisms underlying MetSy in men, in comparison to women.

In the second group of experiments, the aim was to evaluate the influence of selected variables, recommended by the medical expert, on conventionally defined MetSy and to determine their appropriate cut-off values. From relatively large set of variables, indicating important age-related pathogenetic disorders, only four of them reached statistically significant level (Table 2). They included variables *FOLNA* and *HbA1c* for men and *MO* and *TSH* for women. Only two of these four significant variables, *FOLNA* and *HbA1c*, appeared also in previously generated Decision tree models. When considered their statistical properties, only variable FOLNA showed excellent results of all statistical measures, including sensitivity, specificity, PPV (positive predicted value) and NPV (negative predicted value) (Table 2). Because of these properties, the variable *FOLNA* can be considered as a new biomarker of MetSy, particularly suitable for screening in general male population.

Other identified significant variables may also be used in decision-making. The variable *HbA1c* better performed for females in Decision tree models, while according to the statistics, it is better to use for males. For these contradictory results, its usability in relation to gender requires further confirmation. In general, if *HbA1c* is measured, then values above the identified cut-off value (Table 2) confirm the diagnosis of MetSy, but in this way only less than a half of affected persons can be identified (Table 3).

Variables selected as significant for female population, *MO* and *TSH*, might also be useful for practical purposes, although relations of their PPV and NPV measures are not satisfactory enough (Table 2). Since variable *MO* shows better results of sensitivity measure and variable *TSH* of specificity measure, their combination into the model would be of greater predictive utility. The question is only whether this combination is feasible, when compared with the classical scoring method, based on using conventional definition of MetSy. In any way, if a woman has *TSH* parameter measured its value below the identified cut-off value (2.69 mU/L) means that this woman is burdened with MetSy. Latent hypothyroidism, a frequent disorder in older population, characterized with isolated TSH elevation, even yet within the reference range, has just recently been recognized as the risk factor for the MetSy development [33]. Interestingly, the cut-off value for TSH, of 2.5 mU/L and below, found in the epidemiologic studies as significant for MetSy expression, is very similar to what we got in our results [34]. This example thus contributes in favor of our approach for testing ideas. Monocytes % (indicated by the variable *MO*), a part of the White Blood Cells Differential, is easy to perform in everyday medical practice and is frequently ordered for many purposes. According to our results, when we find a woman with Mo% of 5.5 and over, that means that she is susceptible to MetSy, but the diagnosis to be confirmed, this would require further testing, because of low specificity measure of the variable *MO*. Another purpose of this variable might be its use in the first step population screening, because of its good sensitivity measure and low cost performance.

6 Conclusion

We presented here an approach for testing ideas and hypotheses in the clinical domain. For this purpose, on an initial data set, consisting of systematically collected health data which describe the health status of a group of older patients from many aspects, we applied different data mining approaches, in order to test hypotheses given by the medical expert. The leading idea in this work was extraction of possible interesting rules for determination of MetSy from collected data and identification of suitable cut-off values for selected variables, in order to provide better inputs for proper diagnosis. Finally, we joined the best results from both methodological approaches to provide effective supporting mechanism for the diagnosis decision process. We obtained many interesting rules which can be used to test their practical usefulness on real-life data, or as an introduction for planning population-based research. In the case of latter, already tested hypotheses would provide guidelines for conducting research, allowing shortcuts and more efficient research designs. All obtained experiences and knowledge create a good starting point for experiments with larger data samples of a similar nature. But, this approach requires cooperation with a larger number of physicians or with the whole healthcare network and from technological point of view discussions about suitable methods for storing, preprocessing and further analyzing of these data sets. Our paper represents the first step for establishing such kind of collaboration between data mining research groups and application domain expert.

Acknowledgment. This publication is the result of the Project implementation: University Science Park TECHNICOM for Innovation Applications Supported by Knowledge Technology, ITMS: 26220220182, supported by the Research & Development Operational Programme funded by the ERDF (70%); partially supported by the Slovak Grant Agency of Ministry of Education and Academy of Science of the Slovak Republic under grant No. 1/1147/12 (30%).

References

1. Eckel, R.A., Grundy, S.M., Zimmet, P.Z.: The metabolic syndrome. Lancet 365, 1415–1428 (2005)
2. Festa, A., D'Agostino, R., Howard, G., et al.: Chronic subclinical inflammation as part of the insulin resistance syndrome. Circulation 102, 42–47 (2000)
3. Goodwill, H.G., Frisbee, J.C.: Oxidant stress and skeletal muscle microvasculopathy in the metabolic syndrome. Vascul. Pharmacol. 57(5-6), 150–159 (2012), doi:1016/j.vph.2012.07.002. Epub July 11, 2012
4. Oron-Herman, M., Rosenthal, T., Sela, B.A.: Hyperhomocysteinemia as a component of syndrome X. Metabolism 52, 1491–1495 (2003) [PubMed: 14624412]

5. Hjemdahl, P.: Stress and the Metabolic syndrome: an interesting but enigmatic association. Circulation 106, 2634–2636 (2002), doi:10.1161/01.CIR.0000041502.43564.79
6. Onat, A., Hergenc, G., Keles, T., et al.: Sex difference in development of diabetes and cardiovascular disease on the way from obesity and metabolic syndrome. Metabolism 54(6), 800–808 (2005)
7. Lopey-Raton, M., Rodriguez-Alvarez, M.X.: R Package, "OptimalCutpoints" (2013)
8. Lerner, D.J., Kannel, W.B.: Patterns of coronary heart disease morbidity and mortality in the sexes: a 26-year follow-up of the Framingham population. Am. Heart J., 383–390 (February 1986)
9. The MONICA, risk, genetics, archiving and monograph (MORGAM) biomarker project, Contribution of 30 biomarkers to 10-year cardiovascular risk estimation in 2 population cohorts. Circulation 121, 2388–2397 (2010)
10. Engstrom, G., Jerntrop, I., Pessah-Rasmussen, H., et al.: Geographic distribution of stroke incidence within an urban population: relations to socioeconomic circumstances and prevalence of cardiovascular risk factors. Stroke 32(5), 1098–1103 (2001)
11. Ajani, U.A., Ford, E.S.: Has the risk for coronary heart disease changed among U.S. adults? J. Am. Coll. Cardiol. 48(6), 1177–1182 (2006)
12. Holzinger, A., Jurisica, I.: Knowledge Discovery and Data Mining in Biomedical Informatics: The Future Is in Integrative, Interactive Machine Learning Solutions. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 1–18. Springer, Heidelberg (2014)
13. Han-Saem, P., Sung-Bae, C.: Evolutionary attribute ordering in Bayesian networks for predicting the metabolic syndrome. Expert Systems with Applications 39(4), 4240–4249 (2012)
14. Worachartcheewan, A., Nantasenamat, C., Prasertsrithong, P., Amranan, J., Monnor, T., Chaisatit, T., Nuchpramool, W., Prachayasittikul, V.: Machine Learning Approaches for discerning intercorrelation of Hematological Parameters and Glucose Level for identification of diabetes mellitus. EXCLI Journal 12, 885–893 (2013)
15. Holzinger, A., Dehmer, M., Jurisica, I.: Knowledge Discovery and Interactive Data Mining in Bioinformatics – State-of-the-Art, Future challenges and Research Directions. BMC Bioinformatics 15(suppl. 6), I1 (2014)
16. Huppertz, B., Holzinger, A.: Biobanks – A Source of Large Biological Data Sets: Open Problems and Future Challenges. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 317–330. Springer, Heidelberg (2014)
17. Cima, I., Schiess, R., Wild, P., Kaelin, M., Schuffler, P., Lange, V., Picotti, P., Ossola, R., Templeton, A., Schubert, O., Fuchs, T., Leippold, T., Wyler, S., Zehetner, J., Jochum, W., Buhmann, J., Cerny, T., Moch, H., Gillessen, S., Aebersold, R., Krek, W.: Cancer genetics-guided discovery of serum biomarker signatures for diagnosis and prognosis of prostate cancer. Proc. Natl. Acad. Sci. U. S. A. 108, 3342–3347 (2011)
18. International Diabetes Federation. The IDF consensus worldwide definition of the Metabolic Syndrome (2006), `http://www.idf.org/webdata/does/IDF_Meta_def_final.pdf`
19. Holzinger, A., Zupan, M.: KNODWAT: A scientific framework application for testing knowledge discovery methods for the biomedical domain. BMC Bioinformatics 14, 191 (2013)
20. Youden, W.J.: Index for rating diagnostic tests. Cancer 3, 32–35 (1950)
21. Yin, J., Tian, L.: Optimal linear combinations of multiple diagnostic biomarkers based on Youden index. Statistics in Medicine (2013)

22. Lai, C.-Y., Tian, L., Schisterman, E.F.: Exact confidence interval estimation for the Youden index and its corresponding optimal cut-point. Computational Statistics & Data Analysis 56, 1103–1114 (2012)
23. Regitz-Zagrosek, V., Lehmkuhl, E., Weickert, M.O.: Gender differences in the metabolic syndrome and their role for cardiovascular disease. Clin. Res. Cardiol. 95(3), 136–147 (2006)
24. Monnier, L., Colette, C.: Glycemic variability. Diabetes Care 31(suppl. 2), S150–S154 (2008)
25. Franceschi, C., Bonafe, M., Valensin, S., et al.: Human immunosenescence: the prevailing of innate immunity, the failing of clonotypic immunity and the filling of immunological space. Vaccine 18(16), 1717–1720 (2000)
26. Hung-Chih, H., Jeng-Fong, C., Yu-Huei, W., et al.: Folate deficiency triggers an oxidative-nitrosative stress-mediated apoptotic cell death and impedes insulin biosynthesis in RNm5F pancreatic islet β-cells: relevant to the pathogenesis of Diabetes. PLoS ONE 8(11), e77931 (2013), doi:10.1371/journal.pone.0077931
27. Schneider, M.P., Schlaich, M.P., Harazy, J.M., et al.: Folic acid treatment normalizes NOS-dependence of vascular tone in the metabolic syndrome Obesity (Silver Spring) 19(5), 960–967 (2011), doi:10.1038/oby.2010.210. Epub September 23, 2010
28. Franceschi, C., Motta, L., Valensin, S., et al.: Do men and women follow different trajectories to reach extreme longevity? Italian Multicenter Study on Centenarians (IMUSCE) Aging (Milano) 12(2), 77–84 (2000)
29. Sluik, D., Boeing, H., Montonen, J., et al.: HbA1c measured in stored erythrocytes is positively linearly associated with mortality in individuals with Diabetes mellitus. PLoS ONE 7(6), e38877 (2012), doi:10.1371/journal.pone.0038877
30. The Task Force on Diabetes and Cardiovascular Diseases of the European Society of Cardiology and the European Association for the Study of Diabetes: Guidelines on diabetes, pre-diabetes and cardiovascular diseases. Eur. Heart J. (2007), doi:10.1093/eurheartj/ehl261
31. Nebert, D.N., McKinnon, R.A., Puga, A.: Human drug-metabolizing enzyme polymorphisms: effects on risk of toxicity and cancer. DNA Cell Biol. 15(4), 273–280 (1996)
32. Cavagno, L., Boffini, N., Cagnotto, G., et al.: Atherosclerosis and rheumatoid arthritis: more than a simple association. Mediators of Inflammation, Article ID 147354 (2012), doi:10.1155/2012/147354
33. Waring, A.C., Rodondi, N., Harrison, S., et al.: Thyroid function and prevalent and incident metabolic syndrome in older adults: the health, aging and body composition study. Clin. Endocrinol (Oxf.) 76(6), 911–918 (2012), doi:10.1111/i.1365-226.2011.03428.x
34. Ruhla, S., Weickert, M.O., Arafat, A.M., et al.: A high normal TSH is associated with the metabolic syndrome. Clin. Endocrinol (Oxf.) 72(5), 696–701 (2010), doi:10.1111/j.1365-2265.20090369.x