Research Article

Miroslava Nedyalkova*, Sergio Madurga, Davide Ballabio, Ralitsa Robeva, Julia Romanova, Ilia Kichev, Atanaska Elenkova, Vasil Simeonov

# Diabetes mellitus type 2: Exploratory data analysis based on clinical reading

https://doi.org/10.1515/chem-2020-0086
received October 26, 2019; accepted March 19, 2020

**Abstract:** Diabetes mellitus type 2 (DMT2) is a severe and complex health problem. It is the most common type of diabetes. DMT2 is a chronic metabolic disorder that affects the way your body metabolizes sugar. With DMT2, your body either resists the effects of insulin or does not produce sufficient insulin to continue normal glucose levels. DMT2 is a disease that requires a multifactorial approach of controlling that includes lifestyle change and pharmacotherapy. Less than ideal management increases the risk of developing complications and comorbidities such as cardiovascular disease and numerous social and economic penalties. That is why the studies dedicated to the pathophysiological mechanisms and the treatment of DMT2 are extremely numerous and diverse. In this study, exploratory data analysis approaches are applied for the treatment of clinical and anthropometric readings of patients with DMT2. Since multivariate statistics is a well-known method for classification, modeling and interpretation of large collections of data, the major aim of the present study was to reveal latent relations between the objects of the investigation (group of patients and control group) and the variables describing the objects (clinical and anthropometric parameters). In the proposed method by the application of hierarchical cluster analysis and principal component analysis it is possible to identify reduced number of parameters which appear to be the most significant discriminant parameters to distinguish between four patterns of patients with DMT2. However, there is still lack of multivariate statistical studies using DMT2 data sets to assess different aspects of the problem like optimal rapid monitoring of the patients or specific separation of patients into patterns of similarity related to their health status which could be of help in preparation of data bases for DMT2 patients. The outcome from the study could be of custom for the selection of significant tests for rapid monitoring of patients and more detailed approach to the health status of DMT2 patients.

**Keywords:** diabetes mellitus type 2, exploratory data analysis, classification, PCA, CA, PLS-DA

## 1 Introduction

Diabetes mellitus represents a heterogeneous group of metabolic diseases characterized by hyperglycemia resulting from an inadequate insulin secretion or impaired insulin effects [1]. More than 90% of all diabetic patients are diagnosed with diabetes mellitus type 2 (DMT2) [2]. Insulin resistance in muscle, liver and fat cells and relative insulin deficiency due to beta-cell dysfunction are paramount for the development of DMT2; however, other pathophysiological mechanisms including impaired incretin effects, increased glucagon secretion, increased kidney glucose reabsorption and brain insulin resistance might also be of great importance [3]. DMT2 is a chronic disease associated with multiple complications, reduced quality of life, premature mortality and large economic burden, most directly affecting patients in low- and middle-income countries [4,5]. According to a large multinational study,

* **Corresponding author: Miroslava Nedyalkova,** Department of Inorganic Chemistry, Faculty of Chemistry and Pharmacy, University of Sofia "St. Kl. Ohridski", 1164 Sofia, 1, Ave. J. Bourchier, Bulgaria, e-mail: mici345@yahoo.com; tel: +359-81-61-799
**Sergio Madurga:** Department of Physical Chemistry and the Research Institute of Theoretical and Computational Chemistry (IQTCUB) of the University of Barcelona (UB), 08028 Barcelona, C/Martí i Franquès, 1, Spain
**Davide Ballabio:** Department of Earth and Environmental Sciences, Chemometrics and QSAR Research Group, University of Milano-Bicocca, Piazza della Scienza, 1, 20126 Milano, Italy
**Ralitsa Robeva, Atanaska Elenkova:** Faculty of Medicine, Medical University – Sofia, Department of Endocrinology, 1431 Sofia, USHATE Acad. Iv. Penchev, Bulgaria
**Julia Romanova, Ilia Kichev:** Department of Inorganic Chemistry, Faculty of Chemistry and Pharmacy, University of Sofia "St. Kl. Ohridski", 1164 Sofia, 1, Ave. J. Bourchier, Bulgaria
**Vasil Simeonov:** Department of Analytical Chemistry, Faculty of Chemistry and Pharmacy, University of Sofia "St. Kl. Ohridski", 1164 Sofia, 1, Ave. J. Bourchier, Bulgaria

the prevalence of micro- and macrovascular complications in patients with DMT2 was 53.5% and 27.2%, respectively, while 38.4% of the diabetic individuals suffered from diabetic neuropathy [6]. DMT2 and its complications were negatively associated with patients' quality of life in regard to occupation, family and sexual life as well as future perspectives [7]. The economic costs of diabetes have increased by 26% from 2012 to 2017 in parallel with the increased prevalence of diabetes and the increased cost per person with diabetes [8,9].

Recently, the chemometrics and machine learning methods were successfully applied to analyze metabolites and to reveal the potential biomarkers in the clinical diagnosis of DMT2. Free fatty acids were identified as potential biomarkers of DMT2 by using heuristic evolving latent projections, selective ion analysis and competitive adaptive reweighted sampling [10]. Fisher linear discriminate analysis, support vector machine and decision tree algorithms were employed to analyze elements in blood samples, and it has been demonstrated that the level of chromium and iron can serve as a valuable tool of diagnosing DMT2 [11]. The support vector machine approach was also applied to element concentration in urine and hair samples to distinguish between DMT1 and DMT2 [12]. By using orthogonal partial least-squares discriminant analysis for metabolomic analysis of human serum samples, it was revealed that metabolomics fingerprints can identify potential biomarkers of red meat consumption and can be related to the risk of development of DMT2 [13]. The elemental analysis of diabetic toenails and a large variety of machine learning algorithms were combined for the non-invasive diagnosis of DMT2, and it was found that the levels of aluminum, cesium, nickel, vanadium and zinc in toenails can serve as an indicator for the presence or absence of DMT2 [14].

The high incidence and costs as well as the predicted increase of the diabetes prevalence in the future [15] justifies the efforts to apply new multidisciplinary approaches in the diabetic research. Therefore, the present study aims to reveal hidden patterns and subgroups of diabetic patients through hierarchical cluster analysis and principal component analysis (PCA). The major goals of the present study are as follows:

- proper classification of DMT2 patients and members of control group;
- reduction of the number of variables for optimization of the monitoring process of DMT2 patients;
- detection of patterns of similarity within the class of DMT2 patients; and
- determination of discriminant parameters for each identified pattern of patients.

# 2 Materials and methods

## 2.1 Input data and methods of exploratory data analysis

### 2.1.1 Input data

Data subject to intelligent data analysis are collected from patients' records. Data were provided by the for active treatment in endocrinology "Acad. Iv. Penchev", Sofia, Bulgaria and other endocrinological practices. Totally, 100 patients (57 females and 43 males) of age between 36 and 86 with duration of the disease between 1 and 30 years were involved. Additionally, data for 20 healthy volunteers (11 females and 9 males) are added for comparison with the same parameters tested. In Supplementary File 1: Table S1, the clinical parameters measured and their code names are indicated. The following parameters were taken into account:

- anthropometric data – age, duration of disease, weight, height and body mass index (BMI) – calculated according to the well-known formula: weight (kg)/height$^2$ (m$^2$), waist and hip circumference;
- blood tests for thrombocytes, thrombocyte volume to total volume and erythrocyte sedimentation rate;
- liver and muscle functional tests – alanine aminotransferase (ALAT), gamma glutamyl transferase (GGT), albumin and creatinine phosphokinase (CPK);
- kidney function data – creatinine and uric acid;
- protein profile – total protein, lipid profile: high-density lipoproteins (HDLs), low-density lipoproteins (LDLs), very low-density lipoproteins (VLDLs), cholesterol and triglycerides;
- electrolyte content – potassium ions and sodium ions;
- glucose levels (for patients checked in and checked out during a session of medical treatment in a hospital) – hemoglobin $A_{1c}$, fasting glucose, postprandial glucose, before sleep glucose and mean value of glucose.

## 2.2 Exploratory data analysis methods

In order to search for specific relationships between the clinical parameters or between the DMT2 patients. The input matrix has dimensions (120 × 35).

The following methods were used for data interpretation.

### 2.2.1 Hierarchical cluster analysis

Hierarchical cluster analysis is used to detect groups of similarity (clusters) between the objects of interest

(DMT2 patients and control group) or between variables (clinical parameters) by agglomerative hierarchical clustering. The major steps in the analysis include data normalization (z-standardization) to eliminate differences in the variable's dimensions, squared Euclidean distance as similarity measure, Ward's method of linkage, Sneath's test of cluster significance and dendrogram plot as a graphical output [16].

#### 2.2.2 PCA

PCA is able to reduce the number of the variables describing the system of objects in the direction of its highest variance. New variables are introduced, and the coordinates of the existing variable space are replaced by new ones. These new coordinates are the so-called latent factors or principal components (PCs). Their correct interpretation is the main task since they carry specific information about new types of relationships within the original data set. Two sets of output results are considered factor scores giving the new coordinates of the factor space with the location of the objects and factor loadings informing on the relationship between the variables. Only statistically significant loadings (>0.70) are taken into account for the interpretation.

The new PCs selected for interpretation should explain a significant portion of the total variance of the system. Usually, the first principal component (PC1) explains the maximal part of the system variation and each additional PC has a respective contribution to the variance explanation but with less significance.

A reliable interpretation pattern requires normally such a number of PCs, so that over 75% of the total variation can be explained. Often, the Varimax rotated PCA solution is applied that allows a better explanation of the system since it strengthens the role of the latent factors with higher impact on the variation explanation and diminishes the role of PCs with lower impact [17].

#### 2.2.3 Partial least square – discriminant analysis (PLS-DA)

PLS-DA is a supervised linear classification method that combines the properties of partial least squares (PLS) regression with the discrimination power of a discriminant method. The PLS regression algorithm identifies latent variables with a maximum covariance with the classes [18], which are coded into a dummy matrix $Y$, which represents the membership of each sample in a binary form. The PLS2 model is then calibrated on the $Y$ matrix [19] and the probability that a sample belongs to a specific class can be calculated on the basis of the predicted responses [20]. Thus, each modeled class can be described by a classification function reporting the coefficients that determine the linear combination of the original variables to define the classification score. Before the PLS-DA calculation, data were auto scaled.

All calculation work was done by the use of the software package STATISTICA 8.0 [21], except for the classification PLS-DA models, which were calculated with the MATLAB toolbox [22].

**Ethical approval:** All procedures were carried out in accordance with the principles of ethics of the Declaration of Helsinki. The Institutional Ethics Committee approved the use of anonymous patients' data for the study goals.

# 3 Results and discussion

## 3.1 Basic statistics

In Table 1, the basic statistical data for all 120 objects are given.

Detailed statistical data about patients and control group indicate that there are no significant differences between the averages for both groups between indicators such as age, thrombocytes, ALAT, total protein, albumin, K, Na and the different cholesterol determinations ($p < 0.05$). Statistically significant differences ($p > 0.05$) are observed for BMI and the parameters related to it (weight, waist and hip), GGT, CPK, creatinine, uric acid, triglycerides, erythrocyte sedimentation rate (ESR) and all glucose tests. All this could be attributed to factors related to the disease not to demographic reasons.

The correlation analysis carried out for the DMT2 group shows that statistically significant correlations ($p > 0.05$) are found as follows:

*Age*: all parameters related to BMI (anthropometric indicators), creatinine, uric acid, thrombocytes, Na and HbA1c;

*Duration*: weight, ESR, ALAT, creatinine and uric acid;
*Weight*: all BMI indicators and Na;
*Height*: waist, ESR, ALAT and Na;
*BMI*: waist, hip, HDL and HbAc1;
*Waist*: hip, HDL and fast glucose;
*Thrombocytes*: ESR and total protein;
*ESR*: creatinine, albumin and HDL;

**Table 1:** Mean values and standard deviations for 34 clinical parameters (variable "sex" is omitted)

| Variables | Control | SD control | DMT2 patients | SD DMT2 patients |
|---|---|---|---|---|
| Age | 50.35 | 9.82 | 60.29 | 10.58 |
| Duration | 0.00 | 0.00 | 7.35 | 6.47 |
| Weight | 71.05 | 8.82 | 84.22 | 18.30 |
| Height | 173.60 | 7.71 | 165.70 | 10.38 |
| BMI | 22.90 | 1.94 | 30.87 | 6.02 |
| Waist (W) | 86.85 | 4.02 | 103.18 | 12.44 |
| Hip (H) | 88.60 | 6.46 | 109.65 | 14.20 |
| W/H | 0.98 | 0.05 | 0.95 | 0.07 |
| Thromboc | 276.95 | 48.16 | 256.10 | 72.41 |
| Thrvol/Vol | 0.26 | 0.04 | 0.22 | 0.07 |
| ESR | 16.35 | 6.53 | 23.84 | 21.51 |
| ALAT | 28.85 | 9.22 | 29.39 | 23.01 |
| GGT | 38.45 | 12.01 | 49.56 | 48.27 |
| CPK | 67.40 | 36.65 | 101.95 | 121.84 |
| Creatinine | 82.85 | 14.63 | 96.29 | 35.00 |
| Uric acid | 264.45 | 73.94 | 297.98 | 83.68 |
| Total protein | 70.01 | 6.20 | 70.68 | 5.26 |
| Albumin | 39.90 | 5.86 | 38.63 | 3.92 |
| HDL | 1.15 | 0.19 | 1.10 | 0.24 |
| LDL | 2.16 | 0.35 | 2.95 | 1.02 |
| VLDL | 0.76 | 0.40 | 0.86 | 0.41 |
| Cholesterol | 3.70 | 0.67 | 5.05 | 1.51 |
| Triglycerides | 1.17 | 0.40 | 2.54 | 3.29 |
| K | 4.08 | 0.61 | 4.61 | 0.52 |
| Na | 138.60 | 4.42 | 140.23 | 3.47 |
| HbA1c | 5.40 | 0.38 | 8.65 | 1.90 |
| Fast glucose 1 | 4.73 | 0.57 | 9.78 | 3.78 |
| Postprandial | 4.91 | 0.58 | 10.70 | 3.90 |
| Before sleep | 4.80 | 0.40 | 10.30 | 4.64 |
| Mean 1 | 4.83 | 0.45 | 10.09 | 3.37 |
| Fast glucose2 | 4.73 | 0.57 | 7.65 | 2.05 |
| Postprandial | 4.91 | 0.58 | 8.87 | 2.09 |
| Before sleep | 4.80 | 0.40 | 7.45 | 2.66 |
| Mean 2 | 4.83 | 0.45 | 7.87 | 1.64 |

*ALAT*: GGT, total protein, albumin and LDL;
*GGT*: uric acid, HDL, triglycerides, K, Na, postprandial glucose and mean glucose 1;
*Creatinine:* uric acid and total protein;
*Uric acid*: VLDL and triglycerides;
*Total protein*: albumin, LDL and cholesterol;
*Albumin*: HDL, VLDL, HbA1c, before sleep glucose and mean glucose 2;
*HDL, LDL and VLDL*: cholesterol and triglycerides;
*Cholesterol:* triglycerides, Na, fast glucose 1, mean glucose 1, fast glucose 2 and mean glucose 2;
*Triglycerides*: Na, HbAc1, post prandial glucose and mean glucose 1;
*K*: before sleep glucose 2 and mean glucose 2;
*Na*: HbAc1; and

For the control group very few significant correlations are observed.

It could be concluded that several groups of mutual correlated clinical parameters for DMT2 patients are registered as follows:

- glucose test parameters;
- BMI and related anthropometric indicators;
- cholesterol indicators and triglycerides; and
- creatinine, uric acid and proteins.

This is important preliminary information about links between the clinical and anthropometric indicators, which could be of help for the multivariate statistical data analysis.

### 3.1.1 Classification approach to separate controls from DMT2 patients

The goal is to find specific descriptors able to distinguish between the control group and the group of DMT2 patients.

The first approach used was hierarchical cluster analysis. In Figure 1, the hierarchical dendrogram for clustering of all 120 objects (20 of the control group and 100 DMT2 patients) is presented. The separation even in this simple way of grouping is satisfactory – all members of the control group are included in one single cluster. Only one DMT2 patient is wrongly classified as member of the control group. The most statistically significant descriptors for the separation are all glucose test values, all anthropometric indicators, ESR, GGT, CPK and cholesterol.

The same data set was treated by the PLS-DA approach. In Figure 2, the separation of the objects into two very different classes (N – control group and P – patients with DMT2) is indicated. The discriminant indicators are found to be as follows: postprandial glucose test, almost all anthropometric parameters, age and, surprisingly, potassium.

It is shown that reliable classification and separation of the control group from the DMT2 patients' group are possible by the use of all clinical and anthropometric parameters used. In the next steps of the multivariate statistical analysis, an effort will be made to reduce the number of parameters used to optimize the monitoring of the patients.

### 3.1.2 Hierarchical cluster analysis of clinical parameters

In Figures 3 and 4, the hierarchical dendrograms for clustering of the clinical parameters (normalized inputs, squared Euclidean distances as similarity measure,

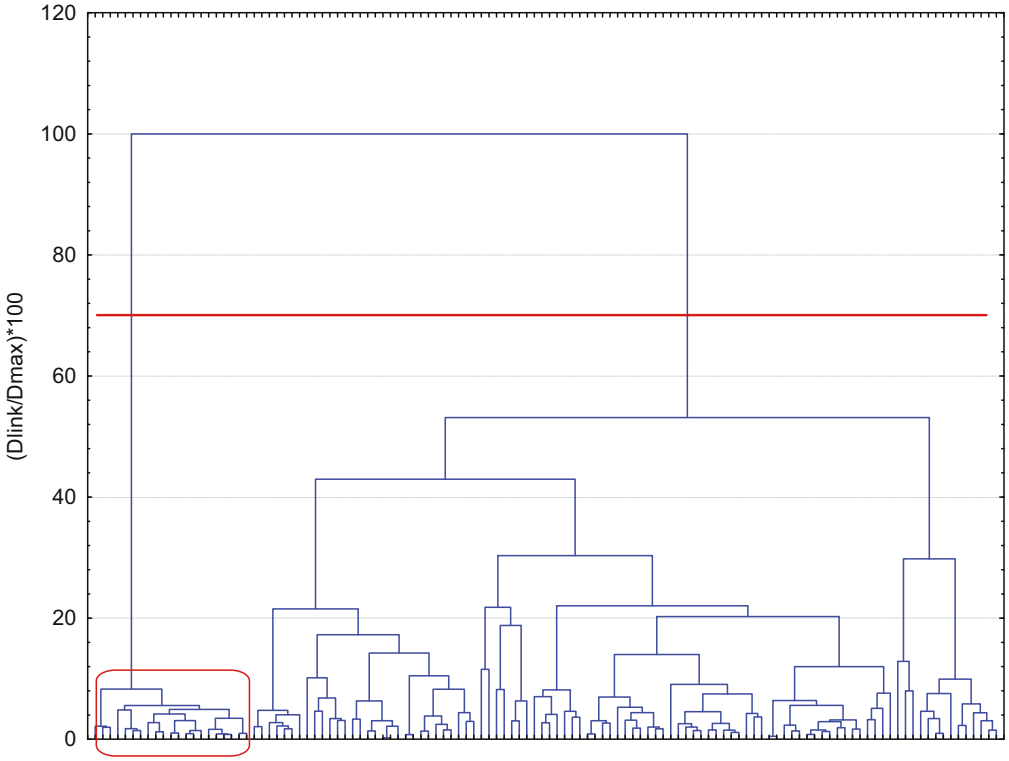


**Figure 1:** Separation of the control group (the cluster on the left side) from the DTM2 patients by hierarchical cluster analysis using all input parameters.

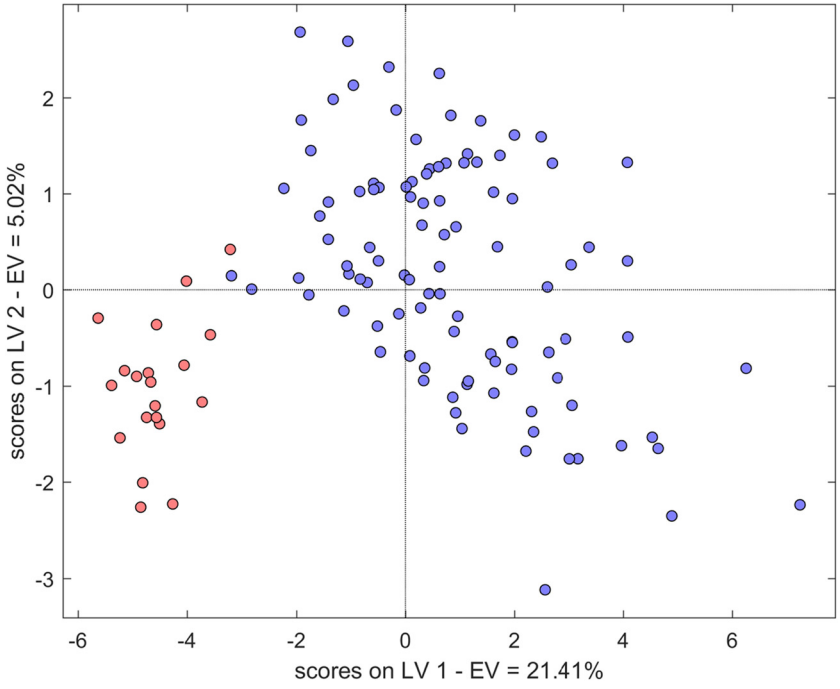


**Figure 2:** Separation of the control group (the cluster on the left side) from the DTM2 patients by PLS-DA approach.

method of Ward of linkage and Sneath's test for cluster significance) are presented (for the control group of 20 participants, Figure 1 and for the DMT2 group of 100 patients, Figure 2).

Four major clusters are formed: the first one includes all glucose tests (except for HbAc1), the second one – dominantly the different cholesterol indicators and enzyme function indicators, the third one – protein and albumin levels along with HbAc1 and the last one – blood parameters, anthropometric indicators and BMI, electrolytes, cholesterol and triglycerides. It could be assumed that in the control group the clustering of the tested parameters is distributed mainly with respect to the body systems and functions involved – enzyme control and cholesterol deposition, protein exchange, blood status and metabolic syndrome assessment.

In Figure 4, the same type of dendrogram for the group of 100 patients with DMT2 is given.

The separation of the 34 clinical parameters (parameter "sex" is eliminated from the analysis since preliminary studies have proven that there is no specific division between male and female patients) leads to the formation of the following five clusters:

K1: all glucose tests including the HbAc1 test, which is considered as one of the most important indicators for long-term glycemic control – *glucose indicator cluster*;

K2: (weight, waist, BMI and hip) – *anthropometric indicator cluster*;

K3: (thrombocytes, ESR, HDL, LDL, cholesterol, VLDL and triglycerides) – this cluster indicates the link between blood quality indicators and cholesterol deposition indicators – *cholesterol cluster*;

K4: (ALAT, gamma-glutamyl transferase (GGK), albumin, total protein, height and K) linkage between enzyme indicators and protein indicators; the link to K seems to be interesting – *enzyme cluster*;

K5: (age, Na, duration, creatinine, uric acid and CPK) – logical link between age and duration of the disease as well as between the indicators for the renal function – *renal function cluster*.

The clustering of the clinical parameters for the DMT2 patient group indicates the specificity of the assessment of the patients with respect to the impact of the disease on the different organs and systems of the human body. It is possible to separate a reduced set of indicators (one or two from each identified cluster) to perform a rapid assessment of the health status of the DMT2 patients.

### 3.1.3 PCA

In order to complete and confirm the results of the hierarchical clustering of the clinical parameters, PCA was additionally carried out. It could help for interpretation of the data structure both for the control group

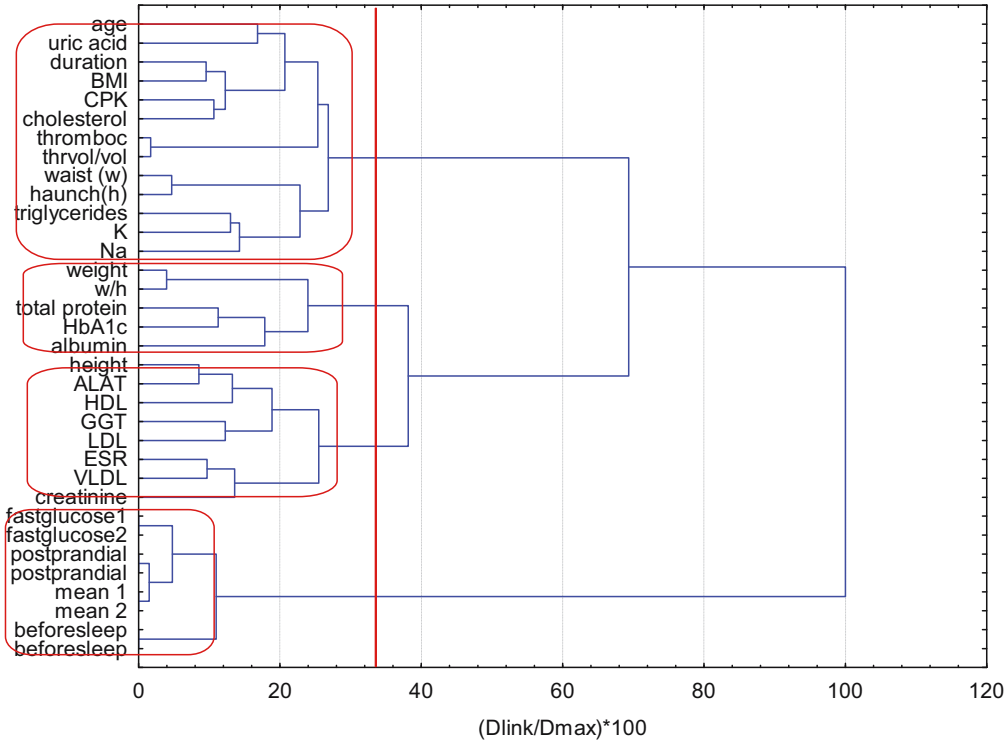


**Figure 3:** Hierarchical dendrogram for linkage of clinical parameters for control group.

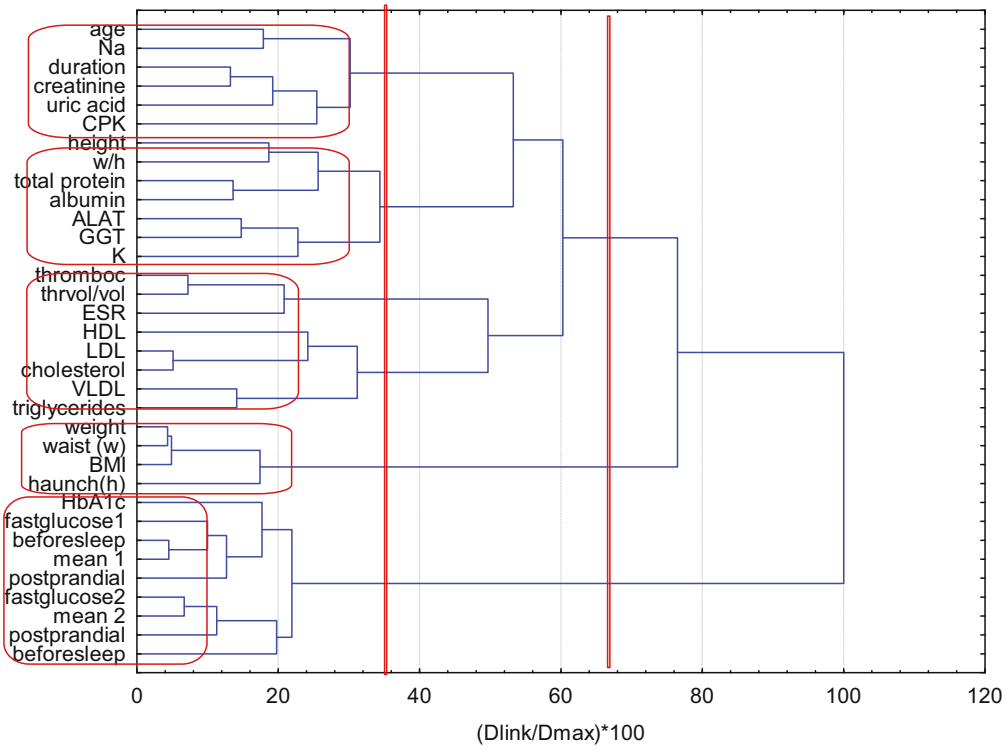


**Figure 4:** Hierarchical dendrogram for linkage of clinical parameters for the group of DMT2 patients.

and for patients with DMT2. Varimax rotation mode was used for both data sets (in Table S2 – SI).

Eight latent factors explain over 80% of the total variance of the system. In general, the significant grouping of clinical parameters resembles that of hierarchical clustering – all glucose tests have high factor loadings in PC1 but HbAc1 does not belong to this first principal component PC4; the anthropometric parameters weight, height and hip are correlated (high factor loadings in PC2). More detailed comparison is not very correct since in PCA one deals with eight latent factors since the identified number of clusters in the hierarchical clustering is four. But PCA gives the opportunity to reduce the number of variables by selecting representative variables from each latent factor (if it is needed to further interpretation).

PC1 (16.6% of explained variance) could be conditionally named "hyperglycemic factor" incorporating all glucose parameters with high factor loadings.

This most significant "hyperglycemic factor" (16.6%) described the increased glucose concentrations. It might reflect the inability of beta-cell to produce sufficient insulin to maintain normoglycemia, which is the cornerstone for the DMT2 development [3]. PC2 (11.9% of the total variance) indicates the close links between the anthropometric parameters and could be conditionally named "anthropometric or obesity factor". The "obesity" factor (11.9%) emphasized the important interrelations between visceral fat mass and carbohydrate dysregulation. Visceral obesity could influence negatively the glycemic control by increasing insulin resistance and gluconeogenesis [23]. Therefore, the acquisition of healthy eating patterns and maintaining of normal body weight are among the fundamental aspects of the DMT2 treatment plan [24].

High factor loadings for almost all cholesterol indicators are found in PC3 (7.7% explanation of the total variance). It is conditionally named "lipid factor".

The "lipid factor" might explain additional 7.7% of the group variation. The atherogenic lipid profile of DMT2 patients is characterized by specific alterations including hypertriglyceridemia, decreased HDL-cholesterol levels and a preponderance of smaller denser LDL cholesterol particles despite the normal LDL cholesterol blood levels [25]. Insulin resistance might determine not only the development of hyperglycemia but also the progress of lipid abnormalities. The increased secretion of free fatty acid from the adipose tissue as well as their decreased utilization in the skeletal muscles due to insulin resistance might enhance their efflux to the liver leading to impaired triglyceride metabolism [26]. The correction of lipid abnormalities in diabetic patients might decrease the risk for macrovascular complications

and reduce the cardiovascular morbidity and mortality [27].

The PC4 indicates good correlation between thrombocytes and ESR (7.4% explained variance) and the conditional name given is "inflammatory factor".

The bidirectional interrelations between the DMT2 and inflammation might explain 7.4% of the variations among diabetic patients and controls. Obesity and insulin resistance are often associated with an increased expression of various pro-inflammatory adipocytokines that might contribute to the maintenance of chronic low-grade systemic inflammation. The inflammatory response could facilitate the development of DMT2 by aggravating the insulin resistance and hyperglycemia, thus creating a vicious circle [28,29]. Since metabolic dysregulation itself maintains inflammation, the adequate treatment of the DMT2, obesity and dyslipidemia might reduce inflammation by improving the metabolic parameters [30]. The use of specific anti-inflammatory agents for reduction of insulin resistance is a matter of further research.

PC5 shows high loadings for age, duration of the disease (very logical link), uric acid and creatinine (explained variance of 7.3%). It indicates the impact of the disease on the kidneys and could be conditionally named "renal function factor".

The age and renal function are important determinants of the intragroup variation (7.3%). Ageing is related to specific difficulties in the diabetes care because of the pronounced heterogeneity in the health status of older adults, different patient's life expectancy, presence of comorbidities, increased risk for hypoglycemia and inability to transfer automatically the results from anti-diabetic studies conducted on younger patients to older ones [42]. The care for patients with renal impairment faces similar problems apart from the specifically limited therapeutic options [31].

The sixth latent factor explains additional 7.1% of the total variance. Its conditional name could be "liver function factor" as it demonstrates correlation between ALAT and GGT. Additionally, it offers a specific link between the enzyme indicators with potassium that could not be explained outside the context of unreported dietary habits or concomitant treatment.

The liver function is another crucial factor that might explain additional 7.1% of the total variance. Hepatocytes are main regulators of the glucose homeostasis through the processes of glycogen storage, glycogenolysis and gluconeogenesis. Thus, hyperglycemic states are often found in patients with hepatic diseases [32]. However, the treatment of diabetes mellitus in patients with liver disorders might be a challenge, because of the increased prevalence of concomitant malnutrition, alcohol abuse, increased risk of hypoglycemia as well as possible side effect of oral antidiabetic drugs metabolized in the liver [33].

PC7 (6.5% explanation of the total variance) is a conditional "protein factor" since it reveals high factor loadings for total protein and albumin.

The last involved latent component PC8 (5.1% explanation of the total variance) indicates the specific role of CPK in the assessment of the health status of the patients.

The importance of the other two latent factors, such as protein and CPK levels, is probably associated with an influence of concomitant conditions and/or medications. The described traits emphasize on the need of personalized complex care for the individuals at increased risk for hyperglycemia including a treatment of concomitant obesity, dyslipidemia and subclinical inflammation considering their age, renal and liver functions.

Since the PCA is a traditional method for space reduction, the further goals of the study were important to select respective variables from the strongly correlated (high factor loadings) parameters of each identified latent factor. Considering the important influence of the postprandial glucose load, obesity, inflammation, renal and liver functions for the health status of the patients with DM2 a restricted set of main indicators was chosen: postprandial glucose 1, BMI, cholesterol, thrombocytes, creatinine, uric acid, GGT and K. These indicators represent each latent factor and count for the role of the glucose tests, anthropometric indicators, liver function, renal function, inflammation markers, lipid profile and electrolytes for effective assessment of the health status of DMT2 patients.

In Figure 5, the hierarchical dendrogram for all 120 objects of the study (controls and DMT2 patients) is presented.

It is seen that the separation between both classes of objects is achieved. The only minor exception is that two patients with DMT2 are wrongly attributed to the control group. This is statistically completely acceptable.

One of the general objectives of the present study is to divide the DMT2 patients into groups of similarity (clusters) using a discrete number of important parameters. This classification could be of use to specific observation of the health status of the different patients and, additionally, to support in identifying symptoms of accompanying DMT2 diseases and complications.

In Figure 6, the hierarchical dendrogram for clustering of 100 DTM2 patients using 8 significant clinical and anthropometric indicators is shown.

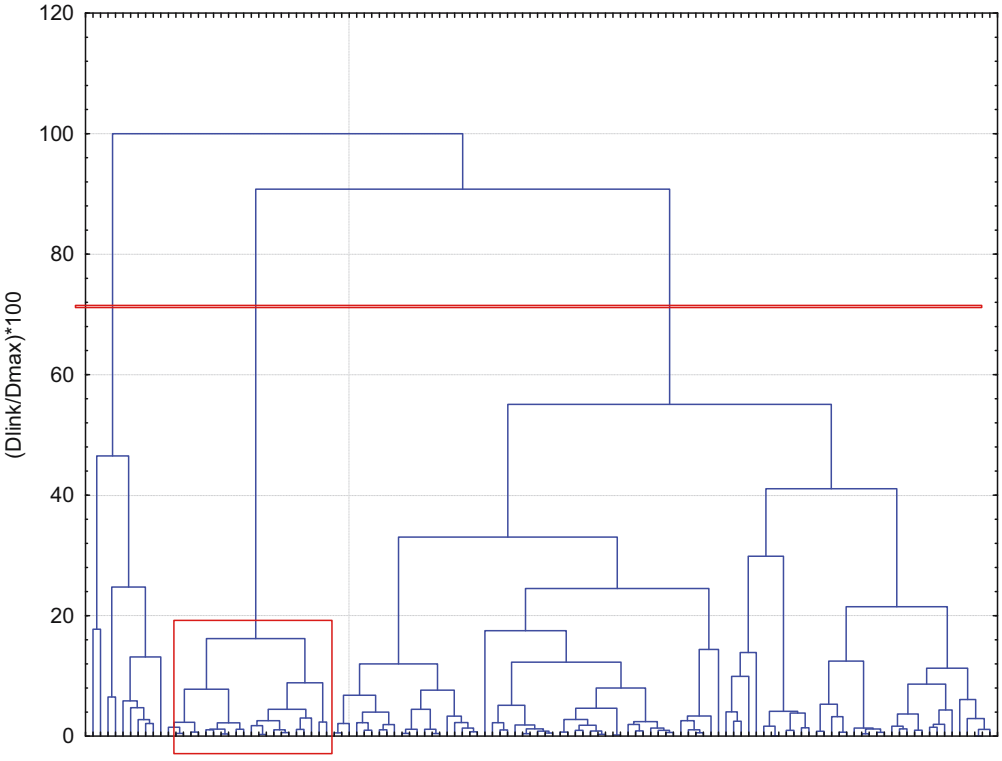


**Figure 5:** Hierarchical dendrogram for separation between objects of the control group (marked) and DMT2 patients using eight selected significant parameters.

Four significant clusters are formed. The members of each cluster are as follows:

Cluster 1 (25 members),

Cluster 2 (31 members),

Cluster 3 (38 members) and

Cluster 4 (6 members).

It could be assumed that each cluster represents patients with specific health status pattern. In order to determine the major discriminants for each pattern of patients, the average values (as standardized values) for each parameter for each cluster were calculated. Figure 7 represents the results.

Cluster 1 is characterized by highest levels for GGT, creatinine, uric acid and K. At the same time, it indicates low BMI, cholesterol and glucose levels. It could be assumed that the 25 members of this cluster might suffer from microvascular complications, such as diabetic nephropathy or they might have concomitant liver or renal diseases despite the relatively good glycemic control.

Cluster 2 includes DMT2 patients (quite significant number) with *worsened DMT2 status* indicated by the highest BMI and glucose level (the most significant discriminants for DMT2). There are no further indications for accompanying health problems.

Cluster 3 is the pattern of patients (largest number of members) with *improved DMT2 status* with no extreme values of the clinical and anthropometric indicators.

Cluster 4 involves a limited number of patients (only six). Although having low mean values for BMI, this pattern of patients shows still high glucose levels as well as and, additionally, highest cholesterol and thrombocyte levels, relatively high creatinine and uric acid levels. The DMT2 status requires significant improvement because of increased risk of micro- and macrovascular complications and cardiovascular morbidity and mortality.

The present study succeeded to distribute the DMT2 patients into groups of similarity (patterns or phenotypes) using a reduced number of commonly tested parameters. Using the described statistical approach, four phenotypes of diabetic patients could be additionally interpreted.

Phenotype 1 was characterized by parameters suggesting impaired renal and liver functions. The BMI, cholesterol and glucose levels were relatively low reflecting the complicated balance between optimal

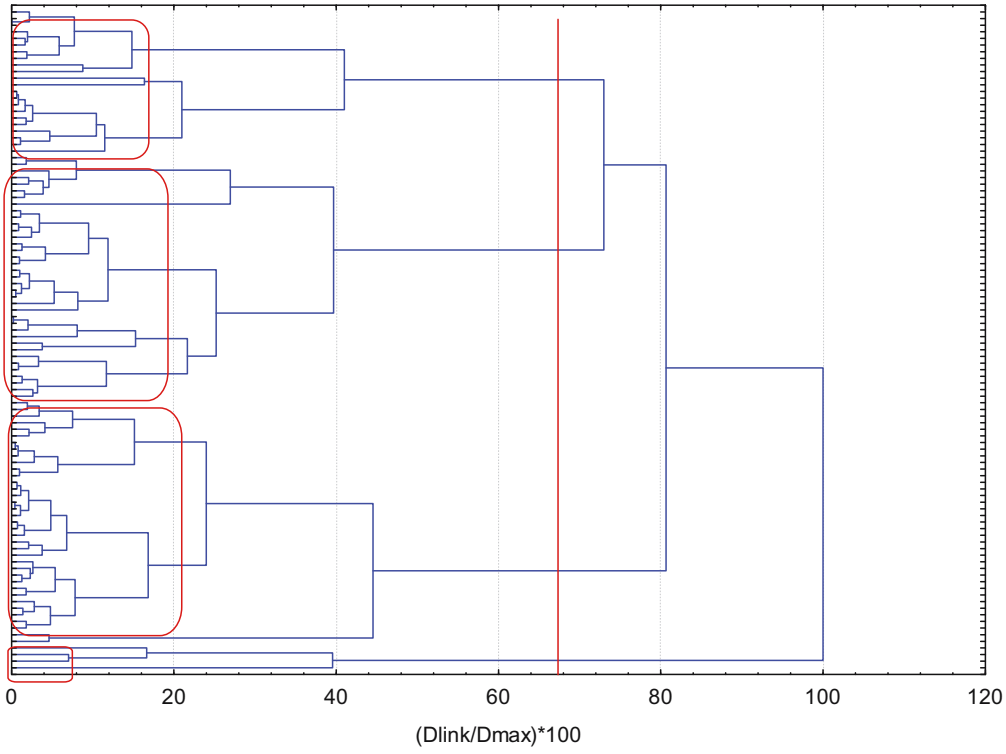


**Figure 6:** Hierarchical dendrogram for clustering of 100 DMT2 patients using 8 variables.

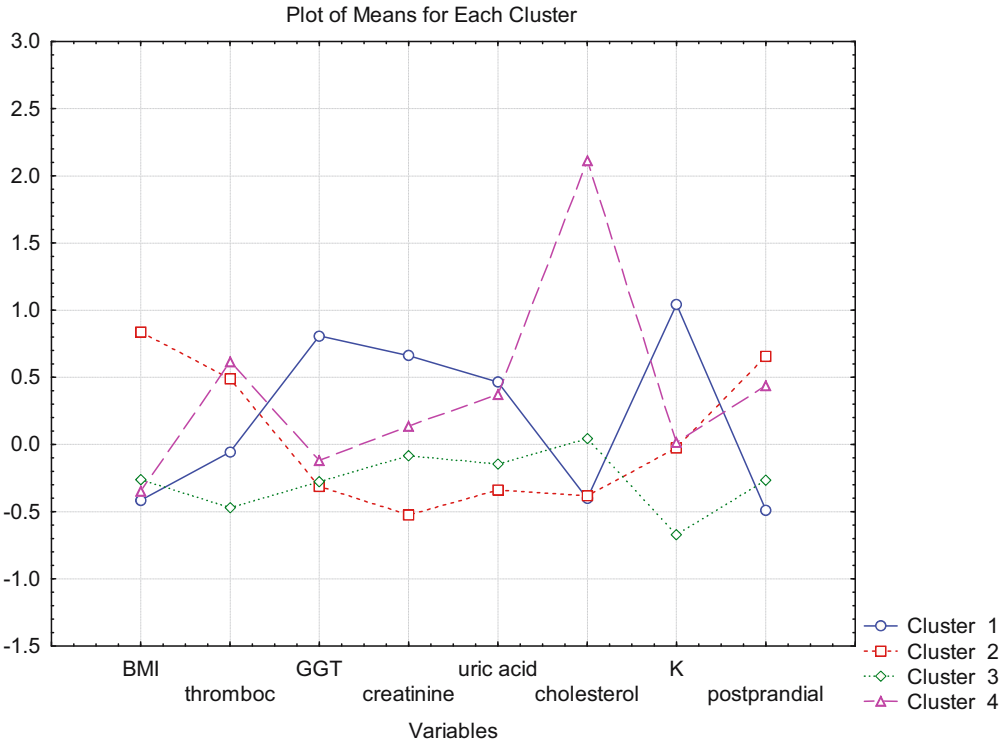


**Figure 7:** Plot of means (standardized values) for each parameter for each identified cluster.

glycemic control, malnutrition and the avoidance of hypoglycemia [31,33]. Therefore, the therapeutic intensity and goals might be less stringent in this category.

Phenotype 2 includes obese diabetic patients with poor glycemic control but no signs of concomitant health problems. The therapeutic approach in these patients should be focused on more intense therapy plan including lifestyle changes, healthy diet as well as antidiabetic drugs considering optimal medication adherence. The maintenance of healthy weight as well as the intensive glycemic control with the goal of achieving near-normoglycemia is crucial for the prevention of diabetic complications [34].

Phenotype 3 includes diabetic patients with optimal laboratory parameters and good control of the hyperglycemia. Thus, no therapy changes are needed in this group of individuals.

Phenotype 4 consists of a limited number of patients with unfavorable lipid and glucose values as well as increased uric acid levels despite the lack of obesity. Probably, these individuals belong to the group of so-called metabolically obese but normal weight patients, who are at increased cardiovascular risk due to increased insulin resistance, hyperglycemia and visceral fat deposition [35]. Moreover, normal-weight elderly people with metabolic disturbances have shown a higher risk of cardiovascular and all-cause mortality in comparison to obese individuals without metabolic disturbances [36]. Therefore, only the aggressive treatment of lipid abnormalities as well as the persistent efforts to optimize glycemic control might preclude the development of macrovascular complications in that group of patients.

# 4 Conclusion

The application of exploratory data analysis to classify, model and interpret clinical data of DMT2 patients has many aspects – to predict DMT2 by classification methods among large group of patients, to model the trajectories of the disease by interpretation of specific indicators, to identify metabolic and genetic biomarkers in patients with DMT2 and concomitant cardio-vascular factors by chemometric approaches, to study diabetic complications etc. [37–40].

In the present study, an effort is made to determine significant indicators out of all typical clinical and anthropometric data for DMT2 patients. The variable reduction offered (8 out of 35 variables) makes it possible to achieve the major goals of the study:

- To classify correctly into different class members of the control group of healthy volunteers from patients with DMT2.
- To determine by rapid tests, the specific health status of statistically significant patterns (clusters) of patients, which allows specific treatment and health care.
- To offer discriminant parameters for each identified specific pattern of DMT2 patients.
- To create a statistical basis for the personalized approach in the treatment of patients with DMT2 and concomitant diseases.

The present study has used intelligent data analysis to explain the variable traits of diabetic patients compared to the control group, which could reflect the differences in the pathophysiological mechanisms related to the disease. Moreover, different phenotypes of patients have been identified as in other studies, which might require distinct therapeutic approach and goals [41].

In conclusion, further efforts to differentiate distinct pathophysiological mechanisms and clinical subgroups through the PCA might contribute to the development of personalized approach in the management of diabetic patients.

**Acknowledgments:** Funding: This project was supported by the BSF grant number KP-06-OPR-03/14-2018 and by the Sofia University.

**Author contributions:** Conceptualization: M. N. and V. S.; methodology: M. N. and V. S.; software: D. B., M. N. and V. S; data writing – original draft preparation: M. N., R. R., J. R. and V. S.; and writing – review and editing: M. N., S. M., R. R., J. R., I. K., A. E. and V. S.

**Conflict of interest:** The authors declare no conflict of interest.

# References

[1] Blair M. Diabetes mellitus review. Urol Nurs. 2016;36:27–36.
[2] Bullard KM, Cowie CC, Lessem SE, Saydah SH, Menke A, Geiss LS, et al. Prevalence of diagnosed diabetes in adults by diabetes type – United States, 2016. Morb Mortal Wkly Rep. 2018;67:359–61.
[3] Defronzo RA. From the triumvirate to the “ominous octet”: a new paradigm for the treatment of type 2 diabetes mellitus. Clin Diabetol. 2009;10:101–28.

[4] Seuring T, Archangelidi O, Suhrcke M. The economic costs of type 2 diabetes: a global systematic review. Pharmacoeconomics. 2015;33:811–31.
[5] Sortsoe C, Green A, Jensen PB, Emneus M. Societal costs of diabetes mellitus in Denmark. Diabet Med. 2016;33:877–85.
[6] Litwak L, Goh S-Y, Hussein Z, Malek R, Prusty V, Khamseh ME. Prevalence of diabetes complications in people with type 2 diabetes mellitus and its association with baseline characteristics in the multinational A1chieve study. Diabetol Metab Syndr. 2013;5:57.
[7] Papazafiropoulou AK, Bakomitrou F, Trikallinou A, Ganotopoulou A, Verras C, Christofilidis G, et al. Diabetes-dependent quality of life (ADDQOL) and affecting factors in patients with diabetes mellitus type 2 in Greece. BMC Res Notes. 2015;8:786.
[8] American Diabetes Association. Economic costs of diabetes in the US in 2017. Diabetes Care. 2018;41:917–28.
[9] American Diabetes Association. Improving care and promoting health in populations: standards of medical care in diabetes-2018. Diabetes Care. 2018;41:7–12.
[10] Tan B, Liang Y, Yi L, Li H, Zhou Z, Ji X, et al. Identification of free fatty acids profiling of type 2 diabetes mellitus and exploring possible biomarkers by GC-MS coupled with chemometrics. Metabolomics. 2010;6:219–228.
[11] Chen H, Tan C. Prediction of type-2 diabetes based on several element levels in blood and chemometrics. Biol Trace Elem Res. 2012;147:67–74.
[12] Chen H, Tan C, Lin Z, Wu T. The diagnostics of diabetes mellitus based on ensemble modeling and hair/urine element level analysis. Comput Biol Med. 2014;50:70–5.
[13] Carrizo D, Chevallier OP, Woodside JV, Brennan SF, Cantwell MM, Cuskelly G, et al. Untargeted metabolomic analysis of human serum samples associated with exposure levels of persistent organic pollutants indicate important perturbations in Sphingolipids and Glycerophospholipids levels. Chemosphere. 2017;168:731–8.
[14] Carter JA, Long CS, Smith BP, Smith TL, Donati GL. Combining elemental analysis of toenails and machine learning techniques as a non-invasive diagnostic tool for the robust classification of type-2 diabetes. Expert Syst Appl. 2019;115:245–55.
[15] Ogurtsova K, da Rocha Fernandes JD, Huang Y, Linnenkamp U, Guariguata L, Cho NH, et al. IDF diabetes atlas: global estimates for the prevalence of diabetes for 2015 and 2040. Diabetes Res Clin Pract. 2017;128:40–50.
[16] Massart DL, Kaufman L. The interpretation of analytical chemical data by the use of cluster analysis. New York: Wiley; 1983.
[17] Vandeginste B, Massart D, De Jong S, Massaart D, Buydens L. Handbook of chemometrics and qualimetrics: art Bp. Elsevier: Amsterdam; 1998.
[18] Höskuldsson A. PLS regression methods. J Chemom. 1988;2:211–28.
[19] Barker M, Rayens W. Partial least squares for discrimination. J Chemom. 2003;17:166–73.
[20] Pérez NF, Ferré J, Boqué R. Calculation of the reliability of classification in discriminant partial least-squares binary classification. Chemom Intell Lab Syst. 2009;95:122–8.
[21] Hill T, Lewicki P, Lewicki P. Statistics: methods and applications: a comprehensive reference for science, industry, and data mining. StatSoft, Inc. United Kingdom; 2006.
[22] Ballabio D, Consonni V. Classification tools in chemistry. Part 1: linear models. PLS-DA. Anal Methods. 2013;5: 3790–8.
[23] Gastaldelli A, Miyazaki Y, Pettiti M, Matsuda M, Mahankali S, Santini E, et al. Metabolic effects of visceral fat accumulation in type 2 diabetes. J Clin Endocrinol Metab. 2002;87:5098–103.
[24] American Diabetes Association. Lifestyle management: standards of medical care in diabetes – 2019. Diabetes Care. 2019;42:46–60.
[25] Haffner SM. Management of dyslipidemia in adults with diabetes. Diabetes Care. 2003;26:83–6.
[26] Krauss RM. Lipids and lipoproteins in patients with type 2 diabetes. Diabetes Care. 2004;27:1496–504.
[27] Krentz AJ. Lipoprotein abnormalities and their consequences for patients with type 2 diabetes. Diabetes Obes Metab. 2003;5:19–27.
[28] Wieser V, Moschen AR, Tilg H. Inflammation, cytokines and insulin resistance: a clinical perspective. Arch Immunol Ther Exp (Warsz). 2013;61:119–25.
[29] Lontchi-Yimagou E, Sobngwi E, Matsha TE, Kengne AP. Diabetes mellitus and inflammation. Curr Diab Rep. 2013;13:435–44.
[30] Pollack RM, Donath MY, LeRoith D, Leibowitz G. Anti-inflammatory agents in the treatment of diabetes and its vascular complications. Diabetes Care. 2016;39:S244–52.
[31] Ioannidis I. Diabetes treatment in patients with renal disease: is the landscape clear enough? World J Diabetes. 2014;5:651–8.
[32] Picardi A, D'Avola D, Gentilucci UV, Galati G, Fiori E, Spataro S, et al. Diabetes in chronic liver disease: from old concepts to new evidence. Diabetes Metab Res Rev. 2006;22: 274–83.
[33] Gangopadhyay KK, Singh P. Consensus statement on dose modifications of antidiabetic agents in patients with hepatic impairment. Indian J Endocrinol Metab. 2017;21:341–54.
[34] American Diabetes Association. 11. Microvascular complications and foot care: standards of medical care in diabetes – 2019. Diabetes Care. 2019;42:S124–38.
[35] Conus F, Rabasa-Lhoret R, Peronnet F. Characteristics of metabolically obese normal-weight (MONW) subjects. Appl Physiol Nutr Metab. 2007;32:4–12.
[36] Choi KM, Cho HJ, Choi HY, Yang SJ, Yoo HJ, Seo JA, et al. Higher mortality in metabolically obese normal-weight people than in metabolically healthy obese subjects in elderly Koreans. Clin Endocrinol. 2013;79:364–70.
[37] Habibi S, Ahmadi M, Alizadeh S. Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining. Glob J Health Sci. 2015;7:304–10.
[38] Oh W, Kim E, Castro MR, Caraballo PJ, Kumar V, Steinbach MS, et al. Type 2 diabetes mellitus trajectories and associated risks. Big Data. 2016;4:25–30.
[39] Evangelista AF, Collares CV, Xavier DJ, Macedo C, Manoel-Caetano FS, Rassi DM, et al. Integrative analysis of the transcriptome profiles observed in type 1, type 2 and gestational diabetes mellitus reveals the role of inflammation. BMC Med Genom. 2014;7:28.
[40] Won JC, Im Y-J, Lee J-H, Kim CH, Kwon HS, Cha B-Y, et al. Clinical phenotype of diabetic peripheral neuropathy and

relation to symptom patterns: cluster and factor analysis in patients with type 2 diabetes in Korea. J Diabetes Res. 2017;2017:5751687.
[41] Montero RM, Herath A, Qureshi A, Esfandiari E, Pusey CD, Frankel AH, et al. Defining phenotypes in diabetic nephropathy: a novel approach using a cross-sectional analysis of a single centre cohort. Sci Rep. 2018;8:53,1–8.
[42] Bradley D, Hsueh W. Type 2 Diabetes in the Elderly: Challenges in a Unique Patient Population. J Geriatr Med Gerontol. 2016;2(2):14.