Bulgarian Journal of Chemistry Volume 3 Issue 1

Bulg J Chem

Article history:
Received: 5 April 2014
Revised: 9 May 2014
Accepted: 10 May 2014
Available online: 2 June 2014

Statistical assessment of medical data by use of cluster analysis part II. Classification of patients

PAPER

Pavlina Simeonova[1], Danail Simeonov[2], Lyubomir Spassov[3] and Vasil Simeonov[4]*

[1]*Laboratory of Environmental Physics, Georgi Nadjakov Institute of Solid State Physics, Bulgarian Academy of Sciences, Tzarigradsko Chaussee Blvd. 72, 1784 Sofia, Bulgaria*
[2]*Institute of General and Inorganic Chemistry, Bulgarian Academy of Sciences, Acad. G. Bontchev Str., Bl. 11, 1113 Sofia, Bulgaria*
[3]*Institute of Catalysis, Bulgarian Academy of Sciences, Acad. G. Bontchev Str., Bl. 11, 1113 Sofia, Bulgaria*
[4]*Laboratory of Chemometrics and Environmetrics, Chair of Analytical Chemistry, Faculty of Chemistry and Pharmacy, University of Sofia "St. Kl. Okhridski", 1164 Sofia, J. Bourchier Blvd. 1, Bulgaria*
**Автор за кореспонденция. E-mail: VSimeonov@chem.uni-sofia.bg*

The aim of the second part of the study is to interpret the sets of medical data collected for patient subjects to different treatments in hospital conditions (prolactinoma, diabetes mellitus type 2 and parneteral and mixed nutrition practice after surgery). In the present study, however, stress was put on the classification of patients. Again, patterns of similarity are sought in the data set with respect to the classification of the patients for each one of the cases in consideration. Each identified pattern is considered by its specific parameters and this way discriminating indicators for every pattern are found. This is achieved by multivariate statistical interpretation using cluster analysis.

Keywords: clinical data; cluster analysis; data treatment; diabetes mellitus type 2; nutrition; patients; prolactinoma.

Introduction

The detailed description of the diseases (prolactinoma and diabetis mellitus type 2) or the type of nutrition after surgery could be found Part I of the study. Thus, no differences with respect to the data collections could be found in this Part II. It has to be kept, however, in mind that the second part is dedicated to the multivariate statistical modeling of the cases of the input matrices – the patients. Very often it is of utmost interest to detect groups of similarity (patterns) between the patient subjects to medical care and treatment. Further, each pattern could be featured by specific indicators (clinical parameters). The goal of the present study is to identify patterns for each medical situation (prolactinoma, diabetis mellitus type 2, mode of nutrition) and to determine the parameters responsible for the pattern formation [1-6].

Experimental

Clinical data for prolactinoma patients

Forty six patients featured by 15 clinical indicators (**BMI ESR HGB RBC PLT Alb Glu Chol LDL Trigl ALAT CK Prol FSH Cortisol**) being thoroughly described in Part I are included. Thirty nine of them are women and seven are men.

Clinical data for diabetes mellitus type 2 patients

Altogether 100 patients (objects in the data set – 57 female and 43 male patients) of different age (between 36 and 86 years of age), and duration of disease (between one and 30 years of duration) were involved in this study. Their status is tested by 34 clinical indicators as described in Part I (**AGE DUR WEIGHT HEIGHT BMI WAIST HAUNCH W/H TROMB TROMB B/V SUE ALAT GGT CPK CREA URIC A TOT_PROT ALB HDL LDL VLDL HOLEST M2_GLU TRIGLI K NA HBA1C F_GLU PP_GLU BS1_GLU M1_GLU F2_GLU PP2_GLU BS2_GLU**)

Clinical data for patients subject to different nutrition

Sixty five patients with major gastrointestinal surgery and indications for postoperative nutritional support are included in the study. The patients are randomized into two groups: mixed nutrition (EEN and PN) - 33 and 32 with TPN after the respective operative interventions. Ten clinical indicators were measured (**Age BMI ALB (0, 2, 5, 10) PREAL (0, 2, 5, 10) CRP(0, 2, 5, 10) Ly (0, 2, 5, 10) ASA NRI SP2 LOS**)

Multivariate statistics

Cluster analysis (CA) is an exploratory data analysis tool for solving classification problems, based on unsupervised learning [7]. It is shortly described in Part I. CA enables objects stepwise aggregation according to the similarity of their features. As a result hierarchically or non-hierarchically ordered clusters are formed.

In this study the z- transformed input data were treated by the use of the squared Euclidean distance as similarity measure, Ward's method of linkage and Sneath's test of cluster significance.

The software package used for calculations is STATISTICA 8.0

Results and discussion

Classification of patients with prolactinoma

Before the classification procedure one of the patients was eliminated from the list of cases due to the very high level of prolactin hormone. As known, outliers deteriorate the classification and has to be avoided. Thus, totally 45 patients were subject to clustering. In *Fig. 1* the linkage of the 45 patients is shown.

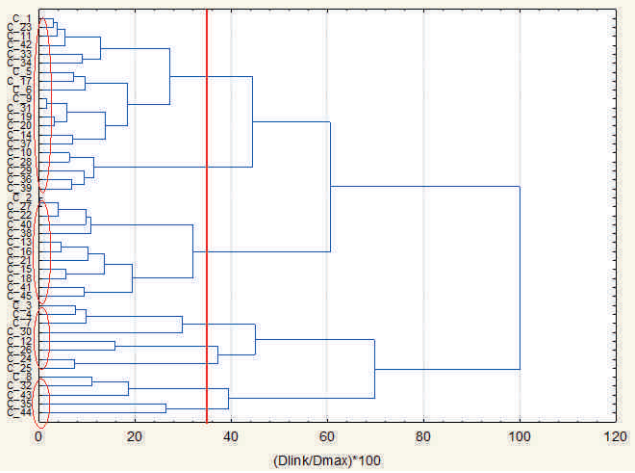

Figure 1. Hierarchical dendrogram for linkage of 45 prolactinoma patients.

Four major clusters are formed:
K1: 1, 23, 11, 42, 33, 34, 5, 12, 6, 9, 31, 19, 20, 14, 37, 10, 28, 29, 36, 39
K2: 2, 27, 22, 40, 38, 13, 16, 21, 15, 18, 41, 45
K3: 3, 4, 7, 30, 12, 26, 24, 25
K4: 8, 32, 43, 35, 44.

It is important to determine the structure of the patient data set with respect to the discriminating clinical parameters for each one of the clusters formed. It will help to understand better some specific features of the prolactinoma patients by their clinical indicators after process of treatment. The elimination of outliers is quite important in this situation as the estimation of the discriminators is based on calculation of the average value of each parameter for each cluster.

In *Table 1* the average values for the clinical parameters for each cluster are given.

Table 1. Average values for each clinical parameter in each cluster.

Variable	K- 1	K - 2	K 3	K - 4
BMI	23,7	**32,5**	32,3	28,5
ESR	9,9	5,8	***14,4***	14
HGB	134,7	**145,8**	129,5	117,8
RBC	4,4	**5**	4,3	4
PLT	**254,5**	227,5	204,1	252,8
ALB	39,3	**40,7**	31	34,9
GLU	4,6	5	**5,2**	4
CHOL	4,2	4,4	4,6	***6,5***
LDL	2,9	2,7	2,1	***4,9***
TRIGL	0,8	**1,5**	**1,5**	1,4
ALAT	13,6	***24,2***	12,4	21,2
CK	50,6	69,4	59,6	***70***
PROLAC	443	398,4	170	***571***
FSH	6,8	5,6	***36,9***	29,3
CORT	373,9	**382,4**	348	154,8

Patients with numbers 1 to 39
are female and 40 to 45 male patients.

For some of the clinical indicators the averages are statistically equal: BMI, HGB, RBC, PLT, ALB, GLU, TRIGLI, and CORT. It means that the whole group of patients does not show any sensitivity to the identified metabolic syndrome factor (BMI, GLU, TRIGLI), to the blood oxidation factor (HGB, RBC) and to blood damaging factor (PLT, ALB, CORT). Thus, they cannot serve as discriminators for the patterns of patients.

The rest of the clinical indicators could play a discriminating function. The patients, belonging to *K1*, do not show any specific response of discriminating indicators. They have relatively low values of BMI, GLU, CHOL, LDL, acceptable level of PROLAC and could be concluded that they form a pattern of "*patients of good health status*". This is the group with the most members (20).

The cluster *K2* (12 members) is characterized by the specific (high) level of ALAT which is probably an indication for some problems with the liver function. This group forms a pattern of patients with the highest levels of BMI, HGB, RBC, ALB, TRIGLI, CORT, which, although not discriminating parameters, indicate still a pathological status. Thus, this pattern of patients could be named "*patients still needing care and testing*". Probably, this group is formed by patients with higher level of initial problems – overweight, worse blood indicators, etc.

The next small cluster (8 patients) *K3* is characterized by highest levels of ESR and FSH (also GLU to some extent) but low levels of CHOL and LDL, ALB and PROLAC. This is a specific pattern of "*patients with good general status but needing improving of the blood parameters*".

The last cluster *K4* (5 patients) includes most of the male patients and could be described as pattern of "*male prolactinoma specificity*". It shows surprisingly high level of PROLAC (it has to be mentioned that the outlier patient 41 is a male patient).

Classification of patients with diabetes mellitus type 2

In *Fig. 2* the hierarchical dendrogram of the cases (patients) is shown.

It might be concluded that the total group of 100 patients is divided into 2 sub-groups – the one consisting of 68 cases (upper part of the dendrogram) and the second one – of 32 cases (lower part of the dendrogram). The major discriminating factor for the two sub-clusters is the average glucose level of the patients. The bigger group includes patients with lower levels, less duration of the disease and better anthropometric parameters. This cluster (68 cases) could be additionally separated into two smaller sub-clusters: the upper one (38 cases) is characterized by patients with higher levels of fat exchange parameters and worse renal function.

In the second cluster of 32 cases (higher glucose levels) another set of two bigger sub-clusters could be defined. In this situation the discrimination is based on differences in the liver function (enzyme content).

Therefore, the classification of the patients by the use of cluster analysis revealed information about separation due to different levels of clinical tests. It is indicated that the state of the various patients is depending on different

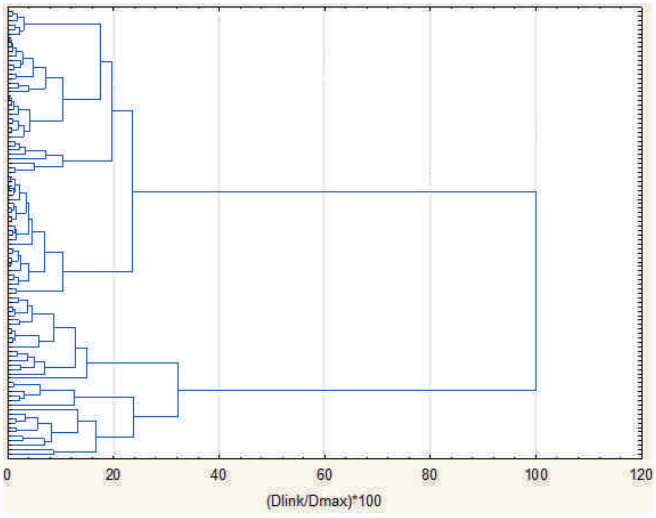

Figure 2. Hierarchical dendrogram for linkage of 100 diabetes mellitus type 2 patients.

links between the clinical parameters, e.g. lower glucose levels correlate to better fat exchange data as the most significant discriminator of the certain group and, vice versa, worse data for the glucose level should be attributed to worse liver function details.

It is important to note that in this situation just limited number of indicators is responsible for the recognition of the patterns formed – mostly the glucose levels during and after the treatment in hospital.

Classification of patients subject to different nutrition modes

It was interesting to additionally analyze the groups of patients with different modes of nutrition after major surgery by patients clustering. In *Figs. 3* and *4* the hierarchical dendrograms for both modes (mixed and parenteral) are shown.

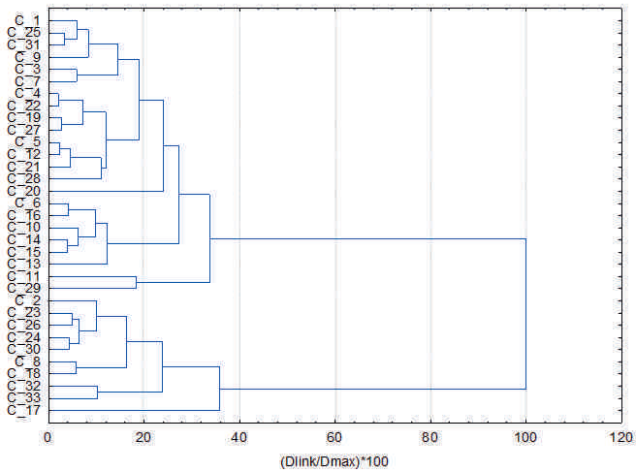

***Figure* 3.** Hierarchical dendrogram for linkage of 33 patients with mixed mode of nutrition.

For patients with mixed mode of nutrition (*Fig. 3*) three clusters are visible: *K1* with 10 patients (conditional numbers 2, 23, 26, 24, 30, 8, 18, 32, 33 and 17), *K2* with 8 patients (29, 11, 13, 15, 14, 10, 16, 6) and *K3* with 15 patients (20, 28, 21, 12, 5, 27, 19, 22, 4, 7, 3, 9, 31, 25, 1). In order to find discriminating parameters for each one of the identified clusters, the average values of all indicators for the cases included in each cluster were calculated.

For *K1* the highest values of BMI, NRI, all ALB and PREAL values are observed. Probably, these are patients responding to a specific "*albumin pattern*" having high levels of BMI I nutrition risk.

K2 is the group of younger patients with lower level of risky anthropometric indices and lower values of albumin indicators. They form a "*low risk*" pattern of patients subject to mixed mode of post-operational nutrition.

Finally, *K3* includes elderly patients (highest average age) with lowest levels of pre-albumin but increased levels of CRP. They could be attributed to the pattern of "*higher risk*" cases.

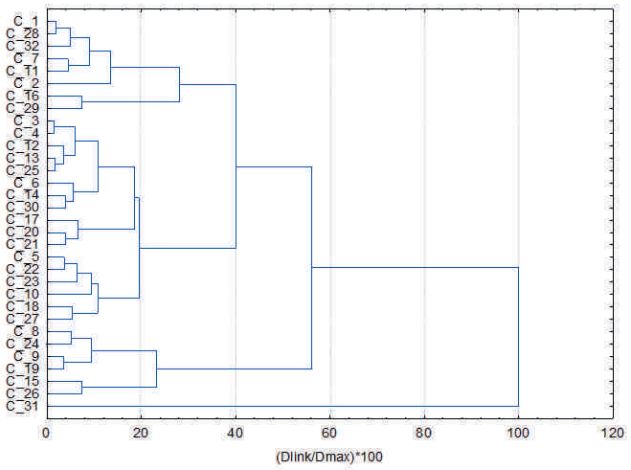

***Figure* 4.** Hierarchical dendrogram for linkage of 32 patients with only parenteral mode of nutrition.

The same discrimination was performed for the group of patients subject to TPN nutrition mode. As seen in *Fig. 4* three distinctive clusters are formed. One typical outlier is presented in the dendrogram (case 31). Cases 8, 24, 9, 19, 15, 26 belong to *K1*, *K2* includes cases 27, 18, 10, 23, 22, 5, 21, 20, 17, 30, 14, 6, 25, 13, 12, 4, 3. The rest of the cases (29, 16, 2, 11, 7, 32, 28, 1) are members of cluster *K3*.

Concerning the discriminating variables following could be mentioned. The outlier case is characterized by highest levels of prealbumins and CRP and by lowest levels of LY and LOS. This isolated case probably informs on a specific pattern of fast food transportation and malnutrition risks.

The first cluster K1 unifies elderly patients but with low BMI and NRI, lowest levels of albumin and CRP5 which is an indication for a reliable nutrition pattern. K2 is the cluster of the younger patients but with highest NRI values. The albumin values are the highest ones as well the LY values which is an indication for active and effective immune system response. Finally, K3 cases are characterized mainly by the highest levels of CRP (2, 5, and 10 day's period of monitoring). This group could be representative for a malnutrition pattern.

Conclusion

Five various patterns of similarity between the prolactinoma patients included in the group of observation were found: patients with good health status, patients needing additional care and observation, patients needing specific care to improve the blood parameters and male patients (as a separate pattern). The extraction of the indicator clinical parameters for each pattern allows a better decision making with respect to the health status of prolactinoma patients.

This study offers a reliable solution of the problem of finding groups (patterns) of similarity between clinical tests usually determined on patients with diabetes mellitus type 2 diagnosis. It can be concluded that all usual clinical tests are divided into 5 groups of similarity each one of them related to vital functions – glucose level, anthropometric indices, renal and liver function, fat exchange indicators. The approach makes it possible to get more information on the links between the functions of the body and to solve easier diagnostic and preventive checks.

The most important conclusion of the patient classification subject to different nutrition modes after surgery is that the mixed type of nutrition hardly differentiates the patients as it does the mixed mode of nutrition.

References

[1] S. Melmed, F. Casanueva, A. Hoffman, D. Kleinberg, V. Montori, J. Schlechte, J. Wass, *J. Clin. Endocrinol. Metab.* 96 (**2011**) 273.

[2] M. Orbetzova, Z. Kamenov, M. Andreeva, G. Genchev, S. Zacharieva, *J.Turk. Soc.Endocrinol.Metabol.* 5 (**2001**) Supplement 10.

[3] S. Wild, G. Roglic, A. Green, R. Sicree and H. King, *Diabetes Care* 27 (**2004**)1047.

[4] P. Tzaneva, K. Vassileva, *Anesthesiol. Intens. Treatment* 23(**1996**) 26.

[5] S.V. Shrikhande, G.S. Shetty, K. Singh, S. Ingle, *J. Cancer Res.Therap.* 5 (**2009**) 232.

[6] H. Baradi, R.M. Walsb, M. Henderson, D. Vogt, M. Popovich, *J. Gastrointest. Surg.* 8 (**2004**) 428.

[7] D. L. Massart, L. Kaufman: "The interpretation of analytical chemical data by the use of cluster analysis", J. Wiley & Sons, New York, **1983**.

Bulg. J. Chem. 3 (**2014**) 45-50

PAPER

Статистическа оценка на данни с помощта на кластерен анализ част II. Класификация на пациенти

Павлина Симеонова[1], Данаил Симеонов[2], Любомир Спасов[3] и Васил Симеонов[4*]

[1]*Лаборатория по Физика на Околната среда, Институт по Физика на Твърдото тяло „Акад. Г. Наджаков", БАН, бул. Цариградско шосе 72, 1784 София, България*
[2]*Институт по Обща и Неорганична Химия, БАН, ул. Акад. Г. Бончев, Бл. 11, 1113 София, България*
[3]*Институт по Катализа, БАН, ул. Акад. Г. Бончев, Бл. 11, 1113 София, България*
[4]*Лаборатория по Хемометрия и Екометрия, Катедра по Аналитична химия, Факултет по Химия и Фармация, Софийски Университет „Св. Кл. Охридски", бул. Дж. Баучер 1, 1164 София, България*
** Автор за кореспонденция. E-mail: VSimeonov@chem.uni-sofia.bg*

Получена: 5 април 2014
Редактирана: 9 май 2014
Приета: 10 май 2014
Излязла online: 2 юни 2014

Целта на втората част на изследването е да интерпретира медицински клинични данни за пациенти, подложени на различно третиране в болнични условия (пролактинома, захарен диабет тип 2, различен тип захранване след операции). Тук се обръща внимание на класификация на пациентите. Търсени са образци на подобие за всеки един от разглежданите клинични състояния. Всеки идентифициран образец се разглежда и по отношение на специфичните клинични индикатори (дискриминиращи индикатори). Целта се постига при прилагане на кластерен анализ.

Ключови думи: диабет тип 2; захранване; клинични данни; кластерен анализ; обработка на данни; пролактином.