Vojnotehnicki glasnik/Military Technical Courier

ISSN: 0042-8469

vojnotehnicki.glasnik@mod.gov.rs

University of Defence

Serbia

J. Paskota, Mira; S. Raškovi, Sanvila; Peri-Popadi, Aleksandra Ž.; D. uri, Vojislav; Jovii, Žikica M.; M. Perovi, Aleksandar

STATISTICAL APPROACH TO SELECTING THE OPTIMAL PARAMETERS FOR DIAGNOSIS OF SOME CONNECTIVE TISSUE DISEASES

Vojnotehnicki glasnik/Military Technical Courier, vol. 67, núm. 3, 2019, pp. 538-560

University of Defence

STATISTICAL APPROACH TO SELECTING THE OPTIMAL PARAMETERS FOR DIAGNOSIS OF SOME CONNECTIVE TISSUE DISEASES

Mira J. Paskota[a], *Sanvila* S. Rašković[b], *Aleksandra* Ž. Perić-Popadić[c], *Vojislav* D. Đurić[d], *Žikica* M. Jovičić[e], *Aleksandar* M. Perović[f]

[a] Universtiy of Belgrade, Faculty of Transport and Traffic Engineering, Belgrade, Republic of Serbia,
e-mail: m.paskota@sf.bg.ac.rs,
ORCID iD: http://orcid.org/0000-0002-7625-6155

[b] Universtiy of Belgrade, School of Medicine, Clinical Center of Serbia, Clinic of Allergolorgy and Immunology, Belgrade, Republic of Serbia,
e-mail: sanvila.raskovic@kcs.ac.rs,
ORCID iD: http://orcid.org/0000-0002-4625-5485

[c] Universtiy of Belgrade, School of Medicine, Clinical Center of Serbia, Clinic of Allergolorgy and Immunology, Belgrade, Republic of Serbia,
e-mail: aleksandra.popadic@kcs.ac.rs,
ORCID iD: http://orcid.org/0000-0001-9718-2688

[d] Universtiy of Belgrade, School of Medicine, Clinical Center of Serbia, Clinic of Allergolorgy and Immunology, Belgrade, Republic of Serbia,
e-mail: vojislav.djuric@kcs.ac.rs,
ORCID iD: http://orcid.org/0000-0002-7544-8307

[e] Universtiy of Belgrade, School of Medicine, Clinical Center of Serbia, Clinic of Allergolorgy and Immunology, Belgrade, Republic of Serbia,
e-mail: zikica.jovicic@kcs.ac.rs,
ORCID iD: http://orcid.org/0000-0002-8805-8391

[f] Universtiy of Belgrade, Faculty of Transport and Traffic Engineering, Belgrade, Republic of Serbia,
e-mail: pera@sf.bg.ac.rs,
ORCID iD: http://orcid.org/0000-0002-8326-8007

DOI: 10.5937/vojtehg67-21023; https://doi.org/10.5937/vojtehg67-21023

FIELD: Mathematics
ARTICLE TYPE: Original scientific paper
ARTICLE LANGUAGE: English

Abstract:

In order to choose the optimal parameters for easier diagnosis of systemic autoimmune diseases, the authors focused on data dimensionality reduction, using both feature selection and feature extraction.

ACKNOWLEDGMENT: This study is partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, through grants nos. III44006, III41013 and TR36001.

The Multiple Correspondence Analysis was used as a feature extraction method, with the aim of exploring the underlying data structure and detecting the crucial latent variables. The obtained latent variables were used as an input for the Discriminant Analysis which correctly classified 86.5% of all analyzed cases. The high rate of correctly classified objects indicates that it would be possible to automate diagnostic processes, which would lead towards the development of decision support systems in this area of medicine. In addition to their knowledge and experience, clinical experts would have further help in decision support systems. That can allow easier learning, faster checking of diagnostic steps, lower rates of misdiagnosed cases and easier communication with experts from other medical centers.

Key words: multiple correspondence analysis, dimensionality reduction, discriminant analysis, connective tussue diseases, autoimmunity, diagnosis.

Introduction

Autoimmune systemic diseases like systemic lupus erythematosus, progressive systemic sclerosis and Sjögren's syndrome can be very difficult to diagnose in practice. Doctors at the primary and secondary level of a typical health care system are usually not qualified enough to recognize connective tissue diseases. Even for specialists at clinics it can be a challenge. The additional problem lies in the fact that many autoimmune diseases patients suffer from more than one condition at the same time. This is why a great number of different parameters are typically needed for the correct diagnosis of these diseases.

In practice, various variables are used for the identification and classification of patients with systemic connective tissue diseases (Hoogen et al, 2013), (Shiboski et al, 2012). For the purpose of research, the American College of Rheumatology developed the Classification criteria for systemic lupus erythematosus in 1982. In 1997, these criteria were revised. The Systemic Lupus International Collaborating Clinics (SLICC) proposed new classification criteria in 2012 (Petri et al, 2012). The SLICC variables were selected after the statistical analysis of patients' medical records by experts, using logistic regression analyses. These variables were then used for the recursive partitioning analysis. The final selection of the variables was performed by the committee of medical experts, but it was strongly influenced by the statistical analysis. Thus, both expert opinion and statistical methods were used in attempts to classify systemic connective tissues diseases (Nadashkevich et al, 2004), (Vitali et al, 2002).

Even though we could not find any application of the Multiple Correspondence Analysis (MCA) in the research of the connective tissue or autoimmune diseases, the applications of the correspondence analysis in medicine are not new. Crichton and Hinde in (Crichton & Hinde, 1989) used a simple Correspondence Analysis (CA) to help diagnose patients with chest pain and acute abdominal pain. Greenacre in (Greenacre, 1992) gives several applications of the CA in different fields of medicine. The same author also gives an example of the application of the MCA in medicine. In (Almeida et al, 2009) the authors are using the MCA in building a logistic model for the predictor selection in living-donor kidney transplant data.

Concerning the other statistical methods applied in the study of autoimmune diseases, we refer the reader to (Armañanzas et al, 2009), where a combination a multivariate correlation and certain machine learning techniques are used for the application of the microarray analysis in study of SLE and PAPS (primary antiphospholipid syndrome).

The rest of the paper is organized as follows: Section 2 gives a short description of the analyzed data set and the available variables. In Section 3, we describe the statistical methods used for the analysis of the data set. Section 4 presents and discusses the results, while the concluding remarks are given in Section 5.

The data set and variables description

The data set consists of 37 patients treated at the Clinic of Allergology and Immunology in Belgrade in the period 2012/2013. Among them, eleven were diagnosed as systemic lupus erythematosus (SLE), fourteen as Sjögren's syndrome (Sy Sjögren), nine as progressive systemic sclerosis (PSS) and three had both SLE and Sy Sjögren. The patients were diagnosed according to the ARA criteria (Hochberg, 1997).

The connective tissue diseases are relatively difficult to diagnose, requiring a broad picture of the patient's medical history, usually assessed through a large number of variables. All the subjects from our study were evaluated using 87 different variables belonging to three different groups, classified according to their 'availability' and 'cost'. The first group consists of 33 variables relatively easy to obtain, and consequentially considered to be 'cheap' (variables 1 to 33, Table 1). These were the variables obtained during the anamnesis and clinical examination of the patients. The second group of 37 variables (variables 34 to 70, Table 1) were the laboratory results of different blood tests, while the 17 variables from the third group (71 to 87, Table 1) are the

results of more invasive diagnostic procedures such as salivary gland histopathology or kidney histopathology, and therefore the most 'expensive' to obtain. It is important to note that the final diagnosis was not included in the data set in any way.

The diagnostics process typically varies among individual patients depending on their condition, so not all of the mentioned diagnostics procedures were needed for all patients and there are some missing cases in the data set.

Methods

A short description of the multivariate correspondence analysis and the discriminant analysis is given in order to familiarize the reader with them and make the text and the results easier to follow and understand.

Multivariate Correspondence Analysis

The MCA is an exploratory statistical technique suitable for analyzing nominal variables, usually applied with the aim of learning something previously unknown about the analyzed data. By the results researchers can get from it and the field of application, the MCA is considered to be the equivalent of the principal component analysis (PCA) for nominal variables. The main features of the MCA are the possibilities of underlying structure exploration/detection and dimension reduction, usually resulting in a set of latent variables. The MCA is a generalization of the Simple Correspondence Analysis (CA), a very popular method for the analysis of contingency tables (Benzécri, 1973), (Greenacre, 1984). While the CA is suitable for the analysis of only two nominal variables, the MCA can be used for the simultaneous analysis of any number of nominal variables. Since the MCA is basically an optimal scaling method, it can also been used for the quantification of nominal variables. Good and detailed descriptions of the MCA, its characteristics and examples of application can be found in the literature, see for instance (Gifi, 1990), (Greenacre & Blasius, 2006), (Le Roux & Rouanet, 2004).

Discriminant analysis

The important results of the MCA are object scores, coefficients of all objects regarding virtual dimensions of the solution. Since these coefficients are numerical, as opposed to original variables being categorical, it is possible to think of the MCA as of a method of quantification. However, it is important to mention that a one to one

relationship between the original and quantified variables does not exist, because object scores are virtual variables, in many ways equivalent to principal components. Keeping that in mind, it is possible to apply any statistical method suitable for the analysis of numerical data on such virtual variables. In this study, the discriminant analysis was used to control the validity of classification.

As usually explained in the literature (Klecka, 1980), (McLachlan, 1992), the discriminant analysis in practice has two main purposes: to find a linear combination of the variables which separate the elements in the best possible way, and to allocate the sample elements into previously defined groups using these linear combinations, usually called discriminant functions. The first and necessary step, finding the discriminant functions, is also a form of data reduction. In some applications, the functions are used as a linear classifier for the allocation of the elements to the previously defined groups. In this research, the discriminant analysis was used with that purpose.

Results and the discussion

The analysis of frequencies, as a necessary first step in every statistical analysis of nominal variables, showed that out of 87 total variables, 29 had too many missing cases to be useful in the analysis. The list of all variables showing if they are included in the analysis and the reasons for the exclusion is given in Table 1. That left 58 variables in the initial set; all were included in the preliminary analysis.

Table 1 – List of all variables
Таблица 1 – Список всех переменных
Табела 1 – Листа свих променљивих

* LF (low frequency)
** HF (high frequency)
*** LC (low contribution)

No	Variable (number of categories)	Step 1		Steps 2 & 3	
		Included	Missing cases	Included	Reason for exclusion
1	Sex (2)	Yes		No	LF* (3/37)
2	Age (4)	Yes		Yes	
3	Malar rash (2)	Yes		Yes	
4	Discoid rash (2)	Yes		No	LF (3/37)
5	Photosensitivity (2)	Yes		Yes	
6	Oral ulcers (2)	Yes		No	LC***

No	Variable (number of categories)	Step 1		Steps 2 & 3	
		Included	Missing cases	Included	Reason for exclusion
7	Dryness of the mouth (2)	Yes		Yes	
8	Arthralgia (2)	Yes		No	HF**(34/37)
9	Arthritis (3)	Yes		No	LC
10	Dryness of eyes (2)	Yes		Yes	
11	Proximal scleroderma (2)	Yes		Yes	
12	Sclerodactyly (2)	Yes		Yes	
13	Digital ulcers (2)	Yes		Yes	
14	Raynaud phenomenon (2)	Yes		No	LC
15	Livedo reticularis (2)	Yes		No	LF (4/37)
16	Dysphagia (2)	Yes		Yes	
17	Teleangiectasia (2)	Yes		No	LC
18	Fever (2)	Yes		No	LC
19	Weight loss (2)	Yes		No	LC
20	Malaise (2)	Yes		No	LC
21	Hair loss (2)	Yes		No	LC
22	Lymphadenopathy (2)	Yes		No	LF (4/37)
23	Epilepsy (2)	Yes		No	LC
24	Psychiatric (2)	Yes		No	LC
25	Psychologic (2)	Yes		No	LC
26	Cerebrovascular disease (2)	Yes		No	LF (2/37)
27	Miscarriage (2)	Yes		No	LC
28	Thrombosis (2)	Yes		No	LF (4/37)
29	Embolism (2)	Yes		No	LF (1/37)
30	Pleural effusion (2)	Yes		Yes	
31	Pulmonary fibrosis (2)	Yes		Yes	
32	Calcinosis (2)	Yes		No	LF (2/37)
33	Blood pressure (3)	Yes		No	LC
34	Erythrocyte sedimentation rate (4)	Yes		No	LC
35	Fibrinogen (3)	Yes		No	LC
36	Anemia (3)	Yes		No	LC
37	Leucopenia (3)	Yes		No	LC
38	Lymphopenia (3)	Yes		Yes	
39	Thrombocytopenia (3)	Yes		No	LC
40	Iron (3)	No	7		
41	Erythrocyturia (3)	Yes		No	LC
42	Cylindruria (2)	Yes		No	LF (4/37)

No	Variable (number of categories)	Step 1		Steps 2 & 3	
		Included	Missing cases	Included	Reason for exclusion
43	Proteinuria (4)	Yes		Yes	
44	Leukocyturia (2)	Yes		Yes	
45	Coombs test (2)	No	5		
46	RF (2)	No	8		
47	CRP (3)	No	5		
48	ANA (3)	Yes		Yes	
49	HEp-2 ANA (2)	No	12		
50	Anticentromere antibody (2)	No	23		
51	ANCA (2)	No	15		
52	MPO (2)	No	32		
53	PR3 (2)	No	33		
54	Anti Sm (2)	No	30		
55	RNP (2)	No	19		
56	Anti ds DNA (3)	Yes		Yes	
57	SSA (3)	No	14		
58	SSB (2)	No	24		
59	SCl 70 (2)	Yes		Yes	
60	AclA IgG (2)	No	13		
61	AclA IgM (2)	No	13		
62	B2GPI IgG (2)	No	31		
63	B2GPI IgM (2)	No	31	No	
64	LA (2)	No	32		
65	VDRL (2)	No	25		
66	KCT (2)	No	18		
67	Lowered complement (2)	Yes		No	LC
68	Elevated IgG IgM (3)	Yes		Yes	
69	Cryoglobulins (2)	No	16		
70	Paraprotein (2)	Yes		No	LC
71	Keratoconjunctivitis sicca (2)	Yes		Yes	
72	Funduscopic abnormalities (2)	No	12		
73	Other eye symptoms (2)	No	9		
74	Capillaroscopy (2)	Yes		Yes	
75	Diffusing capacity (2)	Yes		Yes	
76	Pericardial effusion (2)	Yes		Yes	
77	Pulmonary hypertension (2)	Yes		Yes	
78	Pulmonary scintigraphy (2)	No	32		

No	Variable (number of categories)	Step 1		Steps 2 & 3	
		Included	Missing cases	Included	Reason for exclusion
79	Salivary scintigraphy (4)	No	27		
80	Endocranial NMR (3)	No	22		
81	Chest x ray (2)	Yes		Yes	
82	Hand x ray (2)	No	30		
83	Esophageal dysfunction (2)	Yes		Yes	
84	Lupus band test (2)	No	33		
85	Labial salivary gland histopathology (4)	Yes		Yes	
86	Kidney histopathology (3)	Yes		Yes	
87	Electroneuromyography (3)	No	30		

Two-dimensional solution

Table 2 presents the results of the MCA in the two-dimensional space. Cronbach's alpha is very high for both dimensions, confirming their validity and importance for the interpretation. The first dimension explains 30.698% of the total variability, while the second one explains 25.603%. In the two-dimensional space, the total of 56.301% of the variance is explained. Even though more than 40% of the variability is not explained in this solution, reducing the dimensionality from 27 (number of variables entered in the final analysis) to only two is a very good result and worth further discussion and interpretation.

Table 2 – 2D results of the MCA
Таблица 2 – Результаты 2Д анализа множественной корреспонденции
Табела 2 – Резултати 2Д мултикореспонденционе анализе

Dimension		Variance Accounted For		
	Cronbach's Alpha	Inertia	% of Variance	Total
1	.913	8.289	.307	30.698
2	.888	6.913	.256	25.603
Total		15.201	.563	56.301

The object scores represent positioning of the patients in the two-dimensional space, the objects are labeled by the diagnosis. Figure 1 plots the objects (in our case, they are the patients with the connective tissue disease diagnosis), using their scores along the first two dimensions. At the first glance, it is obvious that the first dimension separates PSS on the right (higher values of the scores) from other

patients, positioned at the left (lower score values). The separation is very clean along the line of x approximately equal to 0.5. The grouping along the second dimension is also very interesting, although the separation is not so clean. Positioned high are PSS, Sy Sjögren, the cases with both Sy Sjögren and SLE and several of the SLE cases. Most of the SLE cases are positioned lower. The second dimension shows both that the SLE cases are more heterogonous than the PSS or Sy Sjögren cases, and that the separation between SLE and Sy Sjögren is not clean.

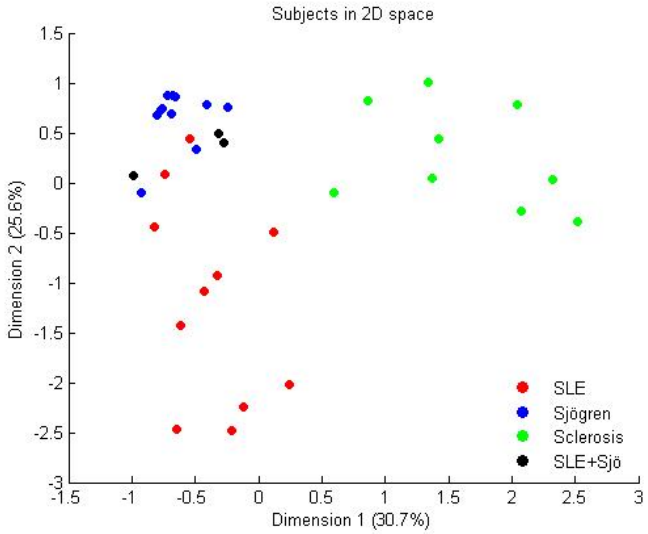

Figure 1 – 2D objects plot
Рис. 1 – 2 Д изображение объектов на участке
Слика 1 – Приказ објеката у равни

The clinical medical experience is in accordance with this result. The PSS patients are usually easy to distinguish by their characteristics from the SLE or Sy Sjögren patients, who are more similar regarding their clinical and biochemical characteristics.

In order to better understand the first two virtual dimensions, we are going to analyze the discrimination measures of all 27 variables (Table 3). The discrimination measures are the squared component loadings along the two virtual axes, and have the meaning of the variance of the quantified variables. As previously explained, in the last step of the variables selection, all variables with the mean discrimination measure in the 2D solution less than 0.1 were excluded from the final analysis.

Table 3 – Discrimination measures of the variables, 2D solution
Таблица 3 – Дискриминативные значения переменных, 2Д решение
Табела 3 – Дискриминационе мере променљивих, 2Д решење

	Dimension		Mean
	1	2	
Age	.064	.466	.265
Malar rash	.106	.137	.122
Photosensitivity	.178	.122	.150
Dryness of the mouth	.208	.506	.357
Dryness of the eyes	.152	.616	.384
Proximal scleroderma	.674	.002	.338
Sclerodactyly	.840	.026	.433
Digital ulcers	.468	.004	.236
Dysphagia	.507	.052	.280
Pleural effusion	.027	.408	.218
Pulmonary fibrosis	.488	.011	.249
Lymphopenia	.030	.434	.232
Proteinuria	.034	.719	.376
Leukocyturia	.572	.176	.374
ANA	.097	.275	.186
Anti ds DNA	.044	.746	.395
Scl-70	.512	.001	.256
Elevated IgG IgM	.114	.338	.226
Keratoconjunctivitis sicca	.131	.545	.338
Capillaroscopy	.472	.052	.262
Diffusing capacity	.601	.065	.333
Pericardial effusion	.139	.205	.172
Pulmonary hypertension	.507	.012	.260
Chest x ray	.254	.117	.185
Esophageal dysfunction	.673	.036	.354
Labial salivary gland histopathology	.376	.316	.346
Kidney histopathology	.018	.526	.272
Active Total	8.289	6.913	7.601
% of Variance	30.698	25.603	28.150

The discrimination measure can take values between 0 and 1. The discrimination measure plot (Figure 2) is very helpful in the interpretation of the virtual space.

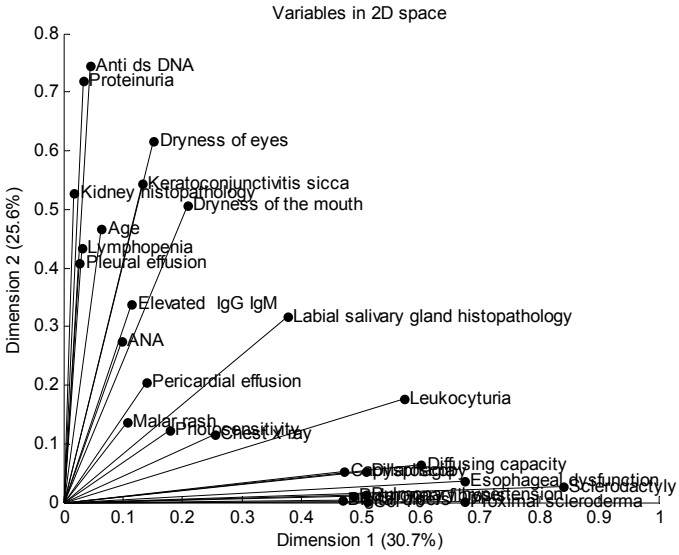

Figure 2 – Discrimination measure plot, 2D solution
Рис. 2 – Изображение 2Д решения, при применении дискриминативных значений
Слика 2 – Приказ 2Д решења применом дискриминационе мере

There are a number of variables with a relatively high value of the discrimination measure along the first, but very low value along the second virtual dimension. In Figure 2, they are positioned very low, close to the x axis. The variables from this group are Diffusing capacity, Esophageal dysfunction, Proximal scleroderma, Sclerodactyly, Digital ulcers, Dysphagia, Pulmonary fibrosis, Scl-70, Capillaroscopy, and Pulmonary hypertension and they can be used to explain the role of the first virtual dimension in the solution. These variables are typical for the PSS patients; some of them like Proximal scleroderma and Esophageal dysfunction are used as the diagnostic criteria for PSS. Therefore, the first dimension was named ‘Sclerosis’.

There are also several variables with relatively low values of the discrimination measure along the first, but quite high values along the second virtual dimension. In the discrimination measure plot (Figure 2), they are positioned very close to the y axis. Kidney histopathology, Proteinuria, Anti ds DNA, Pleural effusion and Lymphopenia are in this

group. These are the variables important for the diagnosis of SLE and lupus nephritis. The variables characteristic for Sy Sjögren (Dryness of eyes, Dryness of mouth, keratoconjunctivitis sicca) are also positioned relatively high and close to the y axis, but not as close as the SLE group of the variables. Some of the variables are characteristic for both SLE and Sy Sjögren (Elevated IgG i IgM, ANA). They are also leaning towards the y axis, but have lower discrimination measures. The second dimension was accordingly named 'SLE and/or Sy Sjögren'.

The rest of the variables have similar contributions towards both virtual dimensions. Most of the variables in the middle, especially the ones with relatively low discrimination measures, are typically seen in both SLE and Sy Sjögren. The variables like Malar rash, Photosensitivity and ANA are positioned closer to the coordinate center and not too close to any of the axes,since they can be observed in both SLE and Sy Sjögren, as is known from the clinical practice.

It is important to understand that the 2D solution explains only 56.301% of the total variability contained in the data, and that it is quite likely that some of these variables highly contribute towards the third (or a higher ranked) dimension, which would not be shown in the 2D representation. In order to better understand the role and importance of different variables for the connective tissue disease diagnosis, we are also going to look at the three-dimensional solution.

Three-dimensional solution

The three-dimensional solution keeps the first two dimensions described in Section 4.1 and adds one more dimension to the preexisting two-dimensional solution. The third dimension also has a relatively high value of Cronbach's Alpha (0.696) and adds 11.221% to the explained variability (Table 4). The first three dimensions together explain 67.522% of the total variance.

Table 4 – 3D results of the MCA
Таблица 4 – 3Д результаты применения анализа множественной корреспонденции
Табела 4 – 3Д резултати примене мултикорепонденционе анализе

Dimension		Variance Accounted For		
	Cronbach's Alpha	Inertia	% of Variance	Total
1	.913	8.289	.307	30.698
2	.888	6.913	.256	25.603
3	.696	3.030	.112	11.221
Total		18.231	.675	67.522

The positions of the objects in the 3D space (Figure 3) are revealing that the PSS patients are clearly separated from others, while the SLE and Sy Sjögren patients are not clearly separated from each other. However, the patients with both diagnoses (SLE and Sy Sjögren) are correctly positioned in the area where the two diagnoses are overlapping. It is also noticeable that the PSS and Sy Sjögren patients do not vary much along the third dimension. The SLE cases, however, are showing significant heterogeneity along the third dimension, as well as along the second dimension. As it was mentioned in the previous discussion, the SLE patients tend to be more different between them and more heterogonous, while the PSS and Sy Sjögren patients are more homogenous in their groups. The third dimension may give more insight in the causes of the SLE heterogeneity.

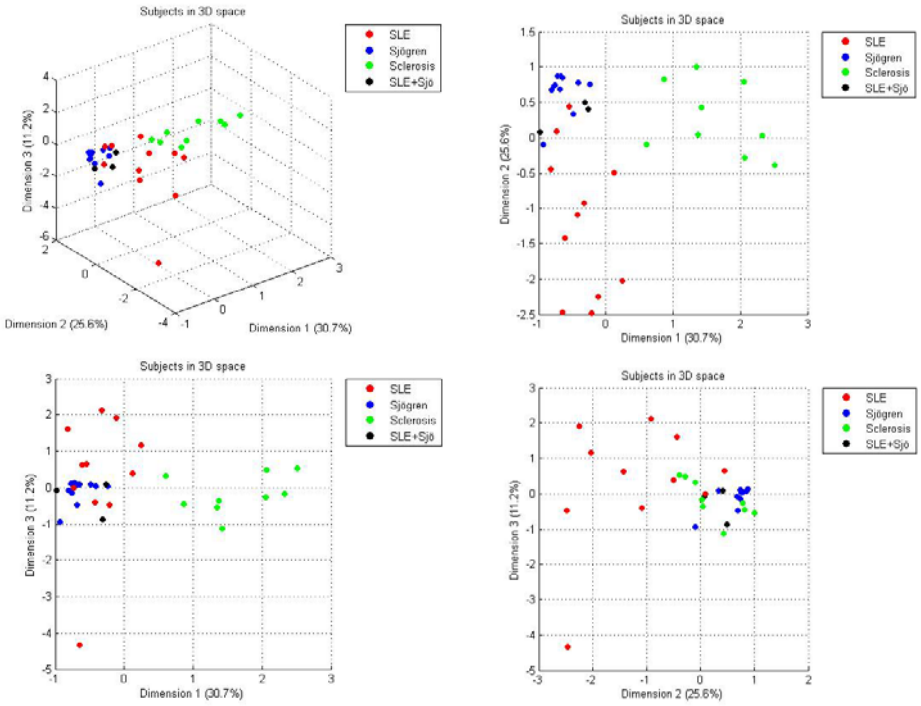

Figure 3 – 3D objects plot
Рис. 3 – 3Д изображение объектов
Слика 3 – 3Д приказ објеката

The variables with very high values of the discrimination measure along the third dimension are Age, Kidney histopathology and

Leukocyturia, while the values of Proteinuria, Anti ds DNA are also relatively high. These variables are responsible for the variations among the SLE cases, and are indicating some level of the kidney dysfunction. This is why the third dimension was named 'Renal Impairment'.

Lupus is a chronic inflammatory autoimmune disease which can affect any organ system, but mainly involves the skin, joints, kidneys and the nervous system (Ching et al, 2012), (Edworthy, 2005), (Hahn et al, 2005), (Hahn et al, 2012), (Muscal & Brey, 2010). SLE has a multitude of presentations ranging from mild, localized disease to severe multi-organ involvement abruptly or sequentially over the course of months to even years. Some patients can have only 4 diagnostic criteria, but many of patients can have more, between 4 and all 11 criteria. This poses a challenge to practitioners as SLE can be a great mimicker of many diseases.

One of the first steps in evaluating a patient with lupus is to recognize that there are various subtypes of lupus (Arbuckle et al, 2009), (Melba & Ovalle, 2013). Autoantibodies alone would not be sufficient to diagnose SLE because these autoantibodies are also present in other rheumatologic diseases (Arbuckle et al, 2009), (Shiboski et al, 2012), (Heaton, 1959), (Tan et al, 2005), (Manoussakis et al, 2004). Sy Sjogren and SLE do have similarities. Their autoantibody profiles are similar. They effect women more than men and have similar HLA haplotypes and autoantibodies. Most likely this is not a coincidence, but it may not be clinically relevant (Manoussakis et al, 2004), (Scheinfeld, 2006).

Sjögren's syndrome may occur in patients with systemic lupus erythematosus (SLE). The subset of patients with SLE and SY Sjögren has a distinct clinical and laboratory phenotype, with a higher frequency among older white women with photosensitivity, oral ulcers, Raynaud's phenomenon, anti-Ro antibodies, anti-La antibodies and a lower frequency of renal disease, anti-dsDNA antibodies and anti-RNP antibodies.

Classification using the Discriminant Analysis

As it was already mentioned, the diagnosis of the patients was never used during the MCA analysis. Since the positions of the objects in the virtual space (Figure 1) indicate that there is a natural grouping of the patients with the same diagnosis, it was necessary to check if that grouping is good enough to be used for the purpose of diagnosis, learning and automated separation of the objects. To accomplish that, the linear discriminant analysis was used.

The object scores from the two-dimensional MCA were used as an input to the discriminant analysis. The grouping variable was the diagnosis, consisting of four different classes: SLE, PSS, Sy Sjögren and SLE + Sy Sjögren. The number of predictors (virtual numerical variables obtained as the result of the MCA analysis) was two, so the number of discriminant functions was also two - equal to the min(number of classes – 1, number of predictors).

Table 5
Таблица 5
Табела 5

Diagnosis	Object scores, dimension 1			Object scores, dimension 2		
	Mean	Std. Deviation	Valid N	Mean	Std. Deviation	Valid N
SLE	-.372118	.3471989	11	-1.188039	1.0299628	11
Sjögren	-.635476	.1850666	14	.695795	.2670113	14
PSS	1.617873	.6588988	9	.260829	.5162426	9
SLE+Sjö	-.523629	.3994424	3	.326613	.2201460	3
Total	.000000	1.0137938	37	.000000	1.0137938	37

Table 5 presents the group means of both variables, while the results of the equality of means test are given in Table 6. The low values of Wilk's Lambda indicate that both variables are very important for the classification and are significantly contributing towards the objects separation (the significance asymptotically converging towards zero in both tests).

Table 6
Таблица 6
Табела 6

	Wilks' Lambda	F	df1	df2	Sig.
Object scores dimension 1	.147	63.775	3	33	.000
Object scores dimension 2	.372	18.570	3	33	.000

The eigenvalues of both discriminant functions with their corresponding canonical correlations are given in Table 7; the first function explains 79.9%, and the second 20.1% of the total variability.

Table 7
Таблица 7
Табела 7

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	6.392(a)	79.9	79.9	.930
2	1.605(a)	20.1	100.0	.785

Based on the MCA virtual dimensions, the DA algorithm was very successful in predicting the group membership (Table 8). 86.5% of the cases were classified correctly. All of the PSS and SLE+Sy Sjögren patients were correctly classified. The only misclassifications were 3 of the SLE and 2 of the Sy Sjögren cases, all predicted as being SLE+Sy Sjögren patients.

Table 8
Таблица 8
Табела 8

Diagnosis	Predicted group membership				Total
	SLE	Sy Sjögren	PSS	SLE+Sjö	SLE
SLE	8	0	0	3	11
Sy Sjögren	0	12	0	2	14
PSS	0	0	9	0	9
SLE+Sjö	0	0	0	3	3
SLE	72.7	.0	.0	27.3	100.0
Sy Sjögren	.0	85.7	.0	14.3	100.0
PSS	.0	.0	100.0	.0	100.0
SLE+Sjö	.0	.0	.0	100.0	100.0

The explanation could be that SLE and Sy Sjögren are frequently overlapping diseases; at the moment we see the patients for the first time it might not be obvious that they can have a mixed form of the disease, named the overlap syndrome. Also, patients with diagnoses of SLE can have some characteristics of Sy Sjögren, (such as dryness of mouth and eyes), but without enough criteria for both diagnoses. A number of patients who seem to have only Sy Sjögren can develop some

manifestations of SLE (eg lymphopenia, ds DNA). The border between the diagnoses of SLE and Sy Sjögren is very subtle, and could be the explanation of the aforementioned misclassifications. Figures 4 and 5 show the corresponding discrimination measure plot for the 3D solution.

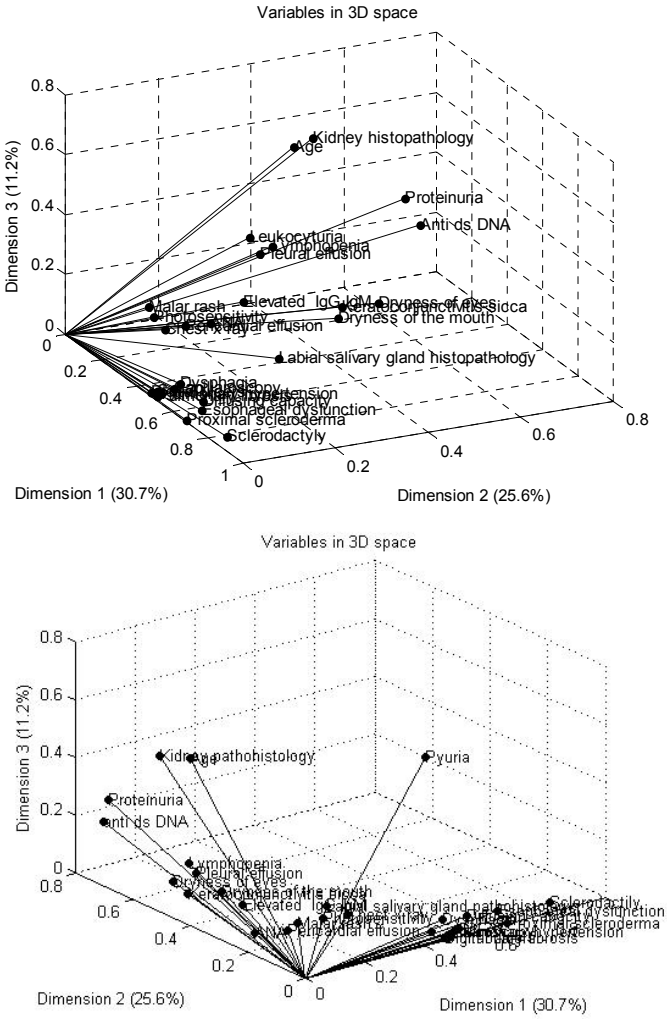

Figure 4
Рис. 4
Слика 4

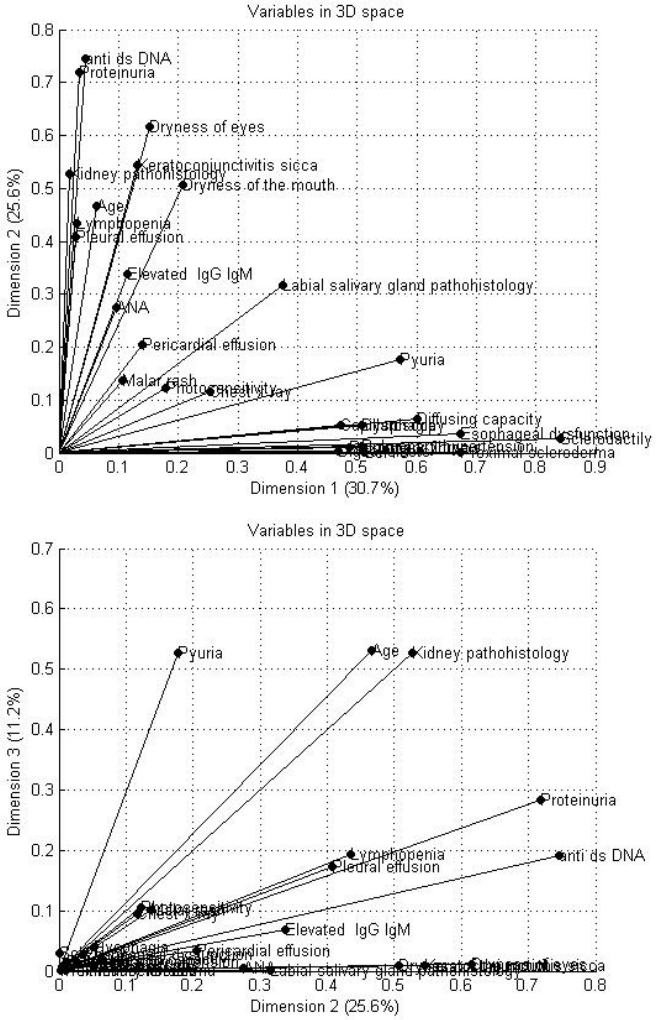

Figure 5
Рис. 5
Слика 5

Conclusion

This study has demonstrated that it is possible to significantly lower the number of parameters needed to diagnose the connective tissue diseases. Out of 87 available variables, 60 were discarded in the three-step eliminatory process. The remaining 27 variables were analyzed

using the multiple correspondence analysis. The three-dimensional solution was enough to identify the most important parameters related to different diseases and clearly separate the cases. Even the two-dimensional solution was enough to give a significant insight into the relationships among the variables and spatial positioning of the patients. The close proximity of some of the variables in the three-dimensional solution might indicate that a further dimension reduction is possible, which can be the subject of a separate study.

The importance of the results is in a possible successful application of the methods of advanced statistics in the medical practice, especially in the process of learning. The discriminant analysis classification was based on the two-dimensional MCA solution. The high rate of correctly classified objects indicates that it would be possible to automate the diagnostic processes, which would lead towards development of decision support systems in this area of medicine. In addition to their knowledge and experience, clinical experts would have further help in decision support systems. Thist can allow easier learning, faster checking of the diagnostic steps, lower rates of misdiagnosed cases, and easier communication with experts from other medical centers.

References

Almeida, R.M.V.R., Infantosi, A.F.C., Suassuna, J.H.R., & Costa, J.C.G.D. 2009. Multiple correspondence analysis in predictive logistic modelling: Application to a living-donor kidney transplantation data. *Computer Methods and Programs in Biomedicine,* 95(2), pp.116-128. Available at: https://doi.org/10.1016/j.cmpb.2009.02.003.

Arbuckle, M.R., McClain, M.T., Rubertone, M.V., Scofield, R.H., Dennis, G.J., James, J.A., & Harley, J.B. 2003. Development of Autoantibodies before the Clinical Onset of Systemic Lupus Erythematosus. *New England Journal of Medicine,* 349(16), pp.1526-1533. Available at: https://doi.org/10.1056/nejmoa021933.

Armañanzas, R., Calvo, B., Inza, I., Lopez-Hoyos, M., Martinez-Taboada, V., Ucar, E.,…, Zubiaga, A.M. 2009. Microarray Analysis of Autoimmune Diseases by Machine Learning Procedures. *IEEE Transactions on Information Technology in Biomedicine,* 13(3), pp.341-350. Available at: https://doi.org/10.1109/titb.2008.2011984.

Benzécri, J.P. 1973. Correspondances. In *L'Analyse des Données*.Dunod. Tome 2.

Ching, K.H., Burbelo, P.D., Tipton, C., Wei, C., Petri, M., Sanz, I., & Iadarola, M.J. 2012. Two Major Autoantibody Clusters in Systemic Lupus Erythematosus. *PLoS ONE,* 7(2), p.32001. Available at: https://doi.org/10.1371/journal.pone.0032001.

Crichton, N.J., & Hinde, J.P. 1989. Correspondence analysis as a screening method for indicants for clinical diagnosis. *Statistics in Medicine,* 8(11), pp.1351-1362. Available at: https://doi.org/10.1002/sim.4780081107.

Edworthy, S.M. 2005. Clinical Manifestations of Systemic Lupus Erythematosus. In: Harris, E.D., & et al. Eds., *Kelley's Textbook of Rheumatology*.Philadelphia, Pa: WB Saunders, pp.1201-1224; 7th ed.

Gifi, A. 1990. *Nonlinear Multivariate Analysis*.John Wiley and Sons.

Greenacre, M.J. 1984. *Theory and Applications of Correspondence Analysis*.Academic Press.

Greenacre, M. 1992. Correspondence analysis in medical research. *Statistical Methods in Medical Research,* 1(1), pp.97-117. Available at: https://doi.org/ Available at: https://doi.org/10.1177/096228029200100106.

Greenacre, M., & Blasius, J. 2006. *Multiple correspondence analysis and related methods*.Chapman and Hall/CRC.

Hahn, B.H., Karpouza, G.A., & Tsao, B.P. 2005. Pathogenesis of systemic lupus erythematosus. In: Harris, E.D., & et al. Eds., *Kelley's Textbook of Rheumatology*.Philadelphia, Pa: WB Saunders, pp.1174-1200; 7th ed.

Hahn, B.H., Mcmahon, M.A., Wilkinson, A., Walace, W.D,…,& Grossman, J.M. 2012. American College of Rheumatology guidelines for screening, treatment, and management of lupus nephritis. *Arthritis Care & Research,* 64(6), pp.797-808. Available at: https://doi.org/10.1002/acr.21664.

Heaton, J.M. 1959. Sjogren's Syndrome and Systemic Lupus Erythematosus. *BMJ,* 1, pp.466-469. Available at: https://doi.org/10.1136/bmj.1.5120.466.

Hochberg, M.C. 1997. Updating the American college of rheumatology revised criteria for the classification of systemic lupus erythematosus. *Arthritis and Rheumatism,* 40(9), pp.1725-1725. Available at: https://doi.org/10.1002/art.1780400928.

Hoogen, F.v.d., Khanna, D., Fransen, J., Johnson, S.R., Baron, M., Tyndall A,…, & Pope, J.E. 2013. 2013 classification criteria for systemic sclerosis: an American college of rheumatology/European league against rheumatism collaborative initiative. *Annals of the Rheumatic Diseases,* 72(11), pp.1747-1755. Available at: https://doi.org/10.1136/annrheumdis-2013-204424.

Klecka, W. 1980. *Discriminant Analysis*.Teller Road, Thousand Oaks California, United States of America: SAGE Publications. Available at: https://doi.org/10.4135/9781412983938.

Le Roux, B., & Rouanet, H. 2004. *Geometric Data Analysis*.Kluwer Academic Publishers.

Manoussakis, M.N., Georgopoulou, C., Zintzaras, E., Spyropoulou, M., Stavropoulou, A., Skopouli, F.N., & Moutsopoulos, H.M. 2004. Sjögren's syndrome associated with systemic lupus erythematosus: Clinical and laboratory profiles and comparison with primary Sjögren's syndrome. *Arthritis and Rheumatism,* 50(3), pp.882-891. Available at: https://doi.org/10.1002/art.20093.

McLachlan, G.J. 1992. *Discriminant Analysis and Statistical Pattern Recognition*. NJ, USA: Wiley. Available at: https://doi.org/10.1002/0471725293.

Melba, I., & Ovalle, M.D. 2013. The Many Faces of Lupus: An Approach to the Assessment of a Lupus Patient. *Clinical Medicine and Diagnostics,* 3(2), pp.11-17. Available at: http://article.sapub.org/10.5923.j.cmd.20130302.01.html.

Muscal, E., & Brey, R.L. 2010. Neurologic Manifestations of Systemic Lupus Erythematosus in Children and Adults. *Neurologic Clinics,* 28(1), pp.61-73. Available at: https://doi.org/10.1016/j.ncl.2009.09.004.

Nadashkevich, O., Davis, P., & Fritzler, M.J. 2004. A proposal of criteria for the classification of systemic sclerosis. *Med. Sci. Monit,* 10(11), pp.615-621.

Petri, M., Orbai, A.M., Alarcón, G.S., Gordon, C., Merrill, J.T., Fortin, P.R.,..., & Magder, L.S. 2012. Derivation and validation of the Systemic Lupus International Collaborating Clinics classification criteria for systemic lupus erythematosus. *Arthritis Rheum,* 64(8), pp.2677-2686. Available at: https://doi.org/10.1002/art.34473.

Scheinfeld, N. 2006. Sjögren syndrome and systemic lupus erythematosus are distinct conditions. *Dermatol Online J.,* 12(1). Available at: https://escholarship.org/uc/item/0jp529zq.

Shiboski, S.C., Shiboski, C.H., Criswell, L.A., Baer, A.N., Challacombe, S., Lanfranchi, H., & Daniels T. E. Sjögren's International Collaborative Clinical Alliance (SICCA) Research Groups. 2012. American College of Rheumatology Classification Criteria for Sjögren's Syndrome: A Data-Driven, Expert Consensus Approach in the SICCA Cohort. *Arthritis Care & Research,* 64(4), pp.475-487. Available at: https://doi.org/10.1002/acr.21591.

Tan, E.M., Cohen, A.S., Fries, J.F., Masi, A.T., Mcshane, D.J., Rothfield, N.F., . . . Winchester, R.J. 2005. The 1982 revised criteria for the classification of systemic lupus erythematosus. *Arthritis and Rheumatism,* 25(11), pp.1271-1277. Available at: https://doi.org/10.1002/art.1780251101.

Vitali, N., Bombardieri, S., Jonsson, R., Moutsopoulos, H.M., Alexander, E.L., Carsons, S.E., Daniels, T.E.,..., Weisman, M.H. & European Study Group on Classification Criteria for Sjögren's Syndrome, 2002. Classification criteria for Sjögren's syndrome: a revised version of the European criteria proposed by the American-European Consensus Group. *Annals of the Rheumatic Diseases,* 61(6), pp.554-558. Available at: https://doi.org/10.1136/ard.61.6.554.

СТАТИСТИЧЕСКИЙ МЕТОД ВЫБОРА ОПТИМАЛЬНЫХ ПАРАМЕТРОВ ДЛЯ ДИАГНОСТИКИ НЕКОТОРЫХ ЗАБОЛЕВАНИЙ СОЕДИНИТЕЛЬНОЙ ТКАНИ

Мира Й. Паскота[а], *Санвила* С. Рашкович[б],
Александра Ж. Перич-Попадич[б], *Воислав* Д. Джурич[б],
Жикица М. Йовичич[б], *Александар* М. Перович[а]

[а] Белградский университет, Факультет транспорта и связи, г. Белград, Республика Сербия

[б] Белградский университет, Медицинский факультет, Клинический центр Республики Сербия, Клиника аллергологии и иммунологии, г. Белград, Республика Сербия

РУБРИКА ГРНТИ: 27.00.00 МАТЕМАТИКА;
27.43.17 Математическая статистика
ВИД СТАТЬИ: оригинальная научная статья
ЯЗЫК СТАТЬИ: английский

Резюме:

В данной работе представлена так называемая редукция размерности данных, проведенная методом селекции и экстракции характерных атрибутов, с целью выбора оптимальных параметров для диагностики заболеваний иммунной системы. Анализ множественной корреспонденции проведен не только при экстракции, но и при исследовании самой структуры данных, а также при диагностике латентных переменных. Благодаря проведенному анализу множественной корреспонденции на материале экстрагированных латентных переменных с максимальной точностью было классифицировано 86,5% наблюдаемых случаев. Высокий уровень точно классифицированных заболеваний свидетельствует о реальных возможностях автоматизации диагностических процессов, которая поможет в усовершенствовании системы поддержки диагностики системных заболеваний соединительной ткани. Данные системы отличаются надежностью и скоростью диагностики, они легко осваиваются и облегчают коммуникацию специалистов из различных медицинских учреждений.

Ключевые слова: анализ множественной корреспонденции, редукция размерности, дискриминативный анализ, заболевания соединительной ткани, автоиммунные заболевания, диагностика.

СТАТИСТИЧКИ ПРИСТУП ИЗБОРУ ОПТИМАЛНИХ ПАРАМЕТАРА У ДИЈАГНОСТИЦИ НЕКИХ БОЛЕСТИ ВЕЗИВНОГ ТКИВА

Мира Ј. Паскота[а], *Санвила* С. Рашковић[б],
Александра Ж. Перић-Попадић[б], *Војислав* Д. Ђурић[б],
Жикица М. Јовичић[б], *Александар* М. Перовић[а]

[а] Универзитет у Београду, Саобраћајни факултет,
Београд, Република Србија
[б] Универзитет у Београду, Медицински факултет, Клинички центар
Србије, Клиника за алергологију и имунологију,
Београд, Република Србија

ОБЛАСТ: математика
ВРСТА ЧЛАНКА: оригинални научни рад
ЈЕЗИК ЧЛАНКА: енглески

Сажетак:

Ради избора оптималних параметара у дијагностици системских аутоимуних болести, аутори су се у овом раду фокусирали на тзв. редуцкију димензионалности података употребом метода селекције и екстракције карактеристичних атрибута. Вишеструка анализа кореспонденције коришћена је не само за екстракцију, већ и за испитивање саме структуре података, као и за детекцију кључних латентних променљивих. На екстраховане латентне променљиве је, применом дискриминантне анализе, коректно класификовано 86,5% посматраних случајева. Висока успешност класификације упућује на реалне могућности аутоматизације дијагностичког процеса, што би резултирало развојем система за подршку у дијагностици системских болести везивног ткива. Овакви системи омогућују лакше учење, бржу и поузданију дијагностику и лакшу комуникацију са експертима из других медицинских центара.

Кључне речи: вишеструка анализа кореспонденције, редукција димензионалности, дискриминантна анализа, болести везивног ткива, аутоимуне болести, дијагностика.

Paper received on / Дата получения работы / Датум пријема чланка: 21.03.2019.
Manuscript corrections submitted on / Дата получения исправленной версии работы / Датум достављања исправки рукописа: 01.06.2019.
Paper accepted for publishing on / Дата окончательного согласования работы / Датум коначног прихватања чланка за објављивање: 03.06.2019.