ASSESSMENT OF HYPERTENSION RISK IN CHILDREN BY CHEMOMETRICAL TECHNIQUES

Viera Mrázová,[a] Miroslava Makohusová,[a]* László Kovács[b] and Katarína Babinská[b]

Keywords: hypertension, prediction, classification, multivariate data analysis

The present work deals with the multivariate data analysis to elucidate the hypertension risk factors in children, which may contribute in some extent to their prospective medical treatment. Results of laboratory tests together with the data obtained from medical documentation were used to indicate as well as to predict the hypertension diagnosis in children. The best diagnostic classification outputs were obtained using artificial neural networks, the K-th nearest neighbour technique, general and linear discriminant analysis. In contrast to the assessment of single laboratory test results, a combination of several tests enables more comprehensive information that can help the physician in diagnosis. This work exemplifies a possible approach to computer-aided medical diagnosis.

* Corresponding Authors
E-Mail: miroslava.makohusova@gmail.com
[a] Department of Chemistry, Faculty of Science, University of SS Cyrilus and Methodius , Trnava, Slovak Republic.
[b] 2nd Department of Paediatrics, Faculty of Medicine of Comenius University, Bratislava, Slovak Republic.

Introduction

The very important advantage of chemometrics and its diverse techniques is versatility, facilitated by its high level of abstraction, characteristic for the scientific disciplines extensively utilizing mathematics. When processing various data of everyday life, exploitable e.g. in medicine or pharmacy, but also in food control or environmental monitoring, resembling algorithms and similar ways of the data processing and evaluation are implemented for different objects under investigation. Considering medicinal data, chemometrical approach allows characterization and quantification of the role of the influencing factors and may effectively support diagnosing and monitoring the progression of a disease. Moreover, this approach enables a total, multicomponent insight into diagnosis **1, 2, 3** instead of sequential view on laboratory results.

Paediatric hypertension is a field of increasing interest and importance. Early identification of children at risk for hypertension is important **4** to prevent serious, long-term complications associated with undetected disease. Although children do not typically suffer from hypertensive disease, an accumulation of medical data suggests that the systolic blood pressure elevation is as an important factor in the morbidity of hypertension in children, as in adults **5**.

Primary hypertension in childhood and adolescence is not a benign disease and cause significant damage of the target organ. Even though clinically evident cardiovascular disease is rare in childhood, the extent of target organ damage is the main risk factor for future cardiovascular events **6**. Patients with moderate to severe hypertension, as well as those with left ventricular hypertrophy, diabetes, cardiovascular and renal disease are considered to be at higher risk and should be considered for prompt pharmacological therapy **7**.

In children, blood pressure measurement is cumbersome and its categorization is based upon the normative data universally defined by National High Blood Pressure Education Program Working Group **8**. The prevalence of childhood obesity is increasing at an alarming rate in developed countries **9**. The prevalence of hypertension is higher in overweight and obese adolescents, and their health state appears to be independent of a family history of hypertension **10**. Obese children are approximately at a three-fold higher risk for hypertension than non-obese children **11**.

Diagnostic criteria for elevated blood pressure in children are based on the concept that blood pressure in children increases with age and body size, making it impossible to utilize a single blood pressure level to define hypertension, as done in adults **12**.

Optimal nutrition during the first year of life is critical to infants' healthy growth and development, and breastfeeding might be a key component. Its duration had no significant effect on infant weight within the first 6 months, but after the 7th month the breastfed infants were lighter than the ones not breastfed. Similarly, the height of the breastfed children is slightly smaller **13**.

Chemometrical multivariate data analysis may elucidate the risk factors in children, which contribute in some extent to prospective treatment of hypertension. The aim of this work is selection of the factors significantly influencing hypertension in children. It is based on exploitation of the results of clinical laboratory tests, as well as the data obtained from medical documentation; further descriptors like age, weight, height, period of breastfeeding are used to indicate or predict the hypertension diagnosis.

Experimental

Data description

The analysed data originate from the 2nd Department of Paediatrics, Faculty of Medicine of Comenius University in Bratislava. The data matrix involves 70 patients aged 3-18 years, among them 29 girls and 41 boys. Three categories of probands were created: (1) the probands without hypertension - 42 patients, (2) the patients with conditional hypertension caused by another disease - 17 non-obese patients with kidney disorders and obese patients associated with other diseases, (3) the patients with confirmed hypertension - 11 probands.

The data with potential effect on hypertension were obtained from the clinical laboratory tests, medical documentation and the data about family history. They represent a set of the following chemometrical descriptors, which all are continuous variables:

Haemoglobin content (g L^{-1}), *HB*; haematocrit (red blood cell volume fraction in the whole blood volume), *HCT*; mean corpuscular volume (fL), *MCV*; content of alanine aminotransferase (μkat L^{-1}), *ALT*; creatinine content (μmol/L), *CR*; sodium concentration (mmol L^{-1}), Na; potassium concentration (mmol L^{-1}), *K*; child's age (year), *AGE*; body mass index, *BMI*; actual weight during hospitalization (kg), *AW*; actual height during hospitalization (cm), *AH*; birth weight (g), *BW*; height at birth (cm), *BH*; period of breastfeeding (months), *BF*. The abbreviated descriptor names in italics were used in computations supported by software specified below.

Hypertension HT was used as the target categorical variable. The probands were categorized into three categories: 1-without hypertension, 2-with conditional hypertension caused by another illness, 3-the patients with hypertension.

Calculation techniques and software

Basic data processing was performed by MS Excel spreadsheet. The following techniques were used for calculations and graph preparations: principal component analysis (PCA), cluster analysis (CLU), correlation analysis (CA), analysis of variance (ANOVA), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), general discriminant analysis (GDA), the K-th nearest neighbour classification (KNN), logistic regression (LR) and artificial neural networks (ANN).

Data exploration and analysis was performed by using advanced contemporary software packages: SAS ver. 9.1.3 (SAS Institute, Inc. 2004), SAS JMP ver. 7.0 (SAS Institute, Inc. 2007), SPSS ver. 15.0 (SPSS, Inc. 2006), Statgraphics Centurion XV ver. 15.1.02 (Stat-Point, Inc. 2006).

Results and discussion

Principal components analysis and cluster analysis

Both techniques, PCA and CLU, enable better displaying of the investigated problem and demonstrate natural grouping of similar objects under study. In this work, all fourteen continuous descriptors were used in calculations of principal component analysis. Acquired principal components result from a linear combination of the descriptors, optimized for maximum data variation. The rays corresponding to individual descriptors are depicted in the PCA loadings plot (Figure 1); the loadings mean in fact the ray end points. The risk factors predisposing the development of hypertension are located along the increasing values of the first principal component (PC1), which may be therefore indicated the quantity reflecting the hypertension risk. It is consistent with an axiom that the PC1 usually signifies the main effect of the given study (under condition that the descriptors are rationally selected).

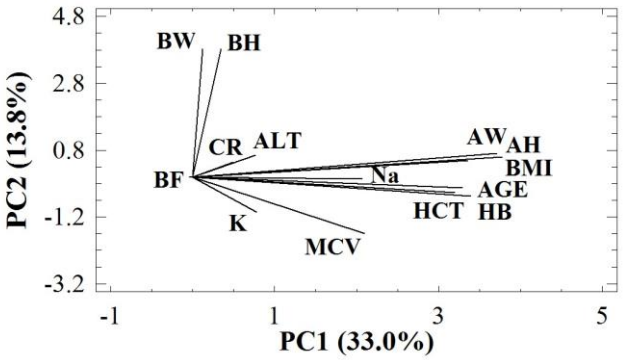

Figure 1. PCA loading plot illustrates 14 descriptors and 70 objects - the patient samples. Software Statgraphics Centurion XV.

The largest loadings with respect to PC1 exhibit *AH*, *AW*, *BMI*, *HB*, *AGE* and *HCT*; a lower effect upon PC1 has Na and *MCV*. The descriptors of the birth weight and height at birth are located along the PC2 axis, perpendicular to the PC1 axis. These two descriptors together with the breastfeeding duration should not have any significant effect upon hypertension in children.

The cluster analysis is a technique that sorts objects into groups by the greatest similarity within the groups and the maximum variability between the groups. The method is suitable for data where the observed variables (descriptors) are quantitative. CLU output is a dendrogram; in this work clustering among the descriptors was evaluated.

Figure 2 shows that the descriptors *HB* with *HCT*, then *AW* with *BMI*, and also *AGE* with *AH* are closely related and form the first common cluster. These six descriptors also most substantially contribute to the PC1 (Fig. 1) therefore they may be considered as most hazardous factors. The second cluster contains descriptors BW and BH, which are similar in some extent to *ALT*; the first two ones exclusively represent PC2 in PCA and their loadings versus PC1 are negligible therefore they have probably no effect upon hypertension. The third cluster is formed by the couples *MCV* and *Na*, *K* and *BF*, plus *CR*; these descriptors except the first two exhibit a small or even insignificant impact on hypertension risk. With regard to the problem of hypertension in children, the PCA and CLU techniques provide synergistic outputs and interpretation of the role of descriptors by their position in the PC2 – PC1 plane is similar to the way how they form the clusters.

Table 1. The correlation table with pairs correlation coefficients *r* for continual variables. Critical value r_{crit} = 0.306. Software SPSS.

	AGE	***AW***	***AH***	***HB***	***HMT***	***MCV***	***Na***	***BMI***
AGE	1.000							
AW	0.654	1.000						
AT	0.850	0.810	1.000					
HB	0.409	0.502	0.543	1.000				
HMT	0.472	0.563	0.592	0.964	1.000			
MCV	0.331	0.296	0.333	0.378	0.417	1.000		
Na	0.294	0.391	0.369	0.278	0.292	0.285	1.000	
BMI	0.519	0.895	0.662	0.475	0.517	0.201	0.344	1.000

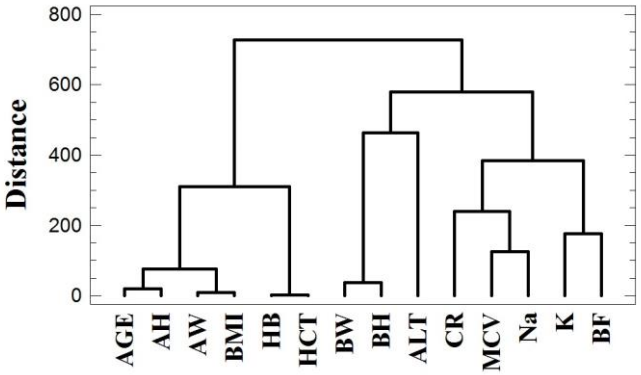

Figure 2. Cluster analysis of 14 investigated descriptors using Ward's method and squared Euclidean distance. Software Statgraphics Centurion XV.

Correlation analysis and analysis of variance

Correlation analysis demonstrates mutual correlations of all pairs of descriptors. Its output is a correlation table containing the pair correlation coefficients R. A part of the descriptors (*AH*, *AW*, *BMI*, *HB*, *HCT*, *Na*) is strongly and positively correlated to *AGE* and their values increase directly with the patient's age.

Found strong correlations between *BMI*, actual weight (*AW*) and height (*AH*) are expected since *BMI* is calculated from these descriptors. A high correlation between the patient's height at birth (*BH*) with the birth weight (*BW*) is also logical. The sodium content considerably increases with the patient's weight, height and *BMI* and is significantly but less strongly correlated to *AGE*, *MCV*, *HCT* and *HB*. Very strong positive correlation (R = 0.964) was found between *HCT* and *HB*, which is related to their shortest distance in the dendrogram and a close position of the corresponding rays in the PC2 - PC1 plane. The results of correlation analysis are consistent with the results of cluster analysis and PCA and are also in agreement with the expected relationships between some descriptors.

Analysis of variance (ANOVA) is a statistical method, which allows to find whether the value of a random variable (descriptor) is significantly affected by some categorical variable, called factor that has two or more levels – categories. In this work the effects of only one factor were investigated - factor *HT* with three categories reflecting different level of hypertension risk. The achieved ANOVA results point out that the strongest impact of hypertension (*HT*) to the investigated descriptors exhibits BMI, then AGE, *AW*, *AH* and *ALT*; in case of other descriptors the effect is smaller or even insignificant. Except *ALT* this is fully consistent with the PCA output. The only descriptor which is different at the significance level p = 0.05 for all three categories of the *HT* factor is *BMI*. Further four above mentioned descriptors reflect only the differences between two category combinations (1 vs. 3, 2 vs. 3) but do not reflect the change between categories 1 and 2. On the other hand, the potassium content (*K*) is not significantly influenced by the change between the category 2 and 3, and finally, breastfeeding (*BF*) is not related to the distinction between the *HT* categories 1 and 3 even though the exchange between the categories 1 and 2 or 2 and 3 shows the significant relation. These findings were obtained by means of least significant difference (LSD) test as well as Bonferroni test in ANOVA.

Discriminant analysis and artificial neural networks

The aim of multidimensional classification in medicinal chemistry, decomposed into four steps, is creation of diagnostic categories, development of classification model by classifying the patients according to known relevancy to the given category, validation of the established model and, finally, categorization of the not yet classified subjects into one of possible classes. The classification efficiency is defined by the ratio of the successfully classified patient samples to the number of all samples; it is calculated in percentage for the training set of the patient samples as well as for the independent validation set.

The patients classified into three categories by the hypertension risk *HT* are displayed in Figure 3, which shows the graphical output of linear discriminant analysis. The samples of probands without hypertension (class 1) are located at negative values of the first discriminant function (DF1), while the patients with confirmed hypertension (class 3) are located at high DF1 values. The samples of the patients with hypertension as a result of another disease (class 2) are located between the groups 1 and 3.

Only two of seventy samples were incorrectly classified by LDA. The first one regards the patient with diagnosed secondary hypertension (class 2) but his values of biochemical tests were affected by metabolic disturbances and therefore he was classified into class 1. The second case regards the patient diagnosed into class 1, but due to gastrointestinal disease the tests values were altered and he was classified into class 2.

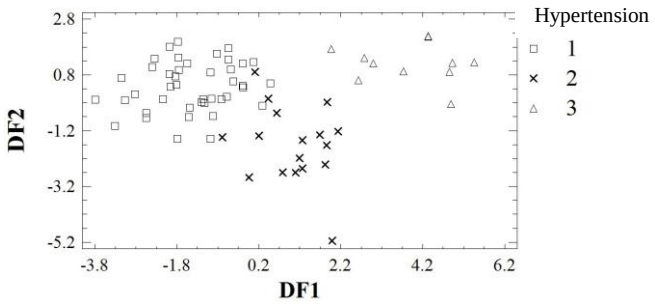

Figure 3. Classification of 70 patient samples by linear discriminant analysis according to hypertension diagnosis *HT*: 1 - patients without hypertension, 2 - patients with a conditional hypertension caused by another disease, 3 - patients with hypertension. Software Statgraphics Centurion XV.

Advanced approach to classification of the categorized samples involves application of several techniques of multidimensional data analysis and selection of that with the best classification efficiency, based mainly on the results achieved for the validation set. This way of valuation has already been successfully applied for diagnostic purposes and disease monitoring **1, 3, 14, 15**. In addition to LDA further classification techniques were used in this work – QDA, GDA, the non-metric KNN technique with different number of nearest neighbours, and logistic regression, which all belong to the discriminant analysis family **16**. Their results are compared to the outputs of ANN.

Table 2 summarizes the classification efficiency of 70 patient samples categorized into 3 classes using 14 continuous descriptors and different classification techniques. The LDA results were obtained by the software packages Statgraphics, SPSS and SAS, and by all software the same results were obtained. The quadratic discriminant analysis (QDA) and the K-th nearest neighbour method (KNN) were performed by using SAS software. The calculations by logistic regression were performed by means of JMP software. Most efficient ways of classification can be selected by the cross validation results in the last column of Table 2.

Table 2. Classification results (in %) in diagnosing hypertension in children.

Classification method	Training set	Cross-validation
LDA	97.1	90.0
QDA	98.6	58.6
GDA	96.5	91.2
KNN $K = 3$	82.9	77.1
KNN $K = 5$	91.4	87.1
KNN $K = 7$	91.4	88.6
KNN $K = 9$	95.7	91.4
KNN $K = 11$	94.3	91.4
LR	100.0	83.8
ANN*)	100.0	92.9*)

*)In ANN classification the test set results are shown instead of the cross-validation output.

Analysis by artificial neural networks was performed by using software Statistica 7.0. A three-layer perceptron with back error propagation was used. The whole data set was randomly divided into a training set (46 samples), validation set (10 samples) and a test set (14 samples). The best results were achieved using 12 optimally selected continuous descriptors at the input: *AGE*, *AH*, *ALT*, *AW*, *BF*, *BH*, *BMI*, *BW*, *HB*, *K*, *MCV*, *Na* and 8 hidden neurons. The selection of the best network was made by comparing the classification efficiency of many networks evaluated for the validation data set, which was performed automatically by software using the option Intelligent Problem Solver).

The inspection of Table 2 reveals very good results of several techniques – over 90 % efficiency. The best results provide ANN with 92.9 % of the uncategorized patient samples classified into the appropriate category. Slightly lower, but still very successful are the classification efficiencies observed for KNN (for K = 9 or 11), GDA and LDA. All of them allow a valuable prediction of the hypertension risk in children.

Conclusions

Hypertension is not only a serious disease of adults, but as demonstrated in recent years, it is now increasingly encountered in children. The progress of this latent disease leads to damage and subsequent failure of vital organs. Diagnosis of hypertension in children is very difficult. The problems with the blood pressure measurement require a long-term monitoring of the state of the patient via biochemical tests leading the physician to diagnose. There exist several hidden risk factors that contribute to the development of hypertension and a possibility of their coincidence increases a real risk for its occurrence. Chemometrical approach, which makes complex analysis of the impact of all risk factors, may effectively support diagnosing and monitoring this disease. It was found that the factors most influencing hypertension risk in children are actual height and weight, body mass index, and age; the biochemical tests mostly reflecting children hypertension are haematocrit, haemoglobin and in a lower extent also sodium content and mean corpuscular volume. Applied multidimensional classification techniques enable a successful prediction of the hypertension risk in children. For this purpose the most successful are artificial neural networks but also simpler techniques – K-th nearest neighbour, linear and general discriminant analysis provide valuable risk prediction. They may considerably facilitate the diagnosis, start the therapy earlier and avoid complications from hypertension.

Acknowledgements

This study was supported by the grant VEGA 1/0073/13 and APVV-0014/11. The analysed data were provided by 2nd Department of Paediatrics, Faculty of Medicine of Comenius University in Bratislava.

References

[1]Balla, B., Mocák, J., Pivovarníková, H., Balla, J., *Chem-Eur. J.*, **2004**, *72*, 259-267.

[2]Kavková, D., Varmusová, E., Tudík, I., Mocák, J., Balla, B. *Stud. Pneumol. Phtisiol*., **2007**, *5*, 199-203.

[3]Mrázová, V., Mocák, J., Varmusová, E., Kavková, D., Bednárová, A., *J. Pharm. Biomed*., **2009**, *50*, 210-215.

[4]Harrabi, I., Belarbia, A., Gaha, R., Essoussi, A. S., Ghannem, H., *Can. J. Cardiol*., **2006**, *22*, 212-216.

[5]Sorof, J. M., *Am. J. Hypertens*., **2002**, *15*, 57-60.

[6]Litwin, M., Niemirska A., Śladowska-Kozlowska, J., Wierbicka, A., Janas, R., Wawer, Z. T., Wisniewski, A., Feber, J., *Pediatr. Nephrol.*, **2010**, *25*, 2489-2499.

[7]Pierdomenico, S.D., Lapenna, D., Bucci, A., Di Iorio, A., Neri, M., Cuccurullo, F., Mezzetti, A., *Am. J. Hypertens*., **2004**, *17*, 876-888.

[8]Falkner B, Gidding SS, Portman R, Rosner B., *Pediatrics*, **2008**, *122*, 238-242.

[9]Salvadori, M, Sontrop, J. M., Garg, A. X., Truong, J, Suri R. S., Mahmud, F. H., Macnab, J. J., Clark, W. F., *Pediatrics*, **2008**, *122*, 821-827.

[10]Benmohammed, K., Nguyen, M. T., Khensal, S., Valensi, P., Lezzar, A., *Diabetes Metab*., **2011**, *37*, 291-297.

[11]Sorof, J., Daniels, S., *Hypertension*, **2002**, *40*, 441-447.

[12]Lurbe, E., Cifkova, R., Cruickshank, J. K., Dillon, M. J., Ferreira, I., Invitti, C., Kuznetsova, T., Laurent, S., Mancia, G., Morales-Olivas, F., Rascher, W., Redon, J., Schaefer, F., Seeman, T., Stergiou, G., Wühl, E., Zanchetti, A., *J. Hypertens*., **2009**, *27*, 1719-1742.

[13]Li, S. C., Kuo, S. C., Hsu, Y. Y., Lin, S. J., Chen, P. C., Chen, Y. C. J., *J. Exp. Clin. Med*., **2010**, *2*, 165-172.

[14]Mocák, J., Balla, B., *Chem. Listy.*, **2003**, *97*, 736-737.

[15]Ďurčeková, T., Mocák, J., Boronová, K., Balla, J., *J. Pharm. Biomed. Anal.* **2011**, *54*, 141-147.

[16]Khattree, R., Naik, N. N., *Multivariate data reduction and discrimination with SAS software*, **2002.**

Received: 31.03.2014.
Accepted: 26.04.2014.