DE GRUYTER OPEN

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 56 (69) 2018

DOI: 10.2478/slgr-2018-0038

Application of Chosen Data Mining Methods in Predicting Abnormal Blood Pressure in Children and Adolescents

Anna Sowińska[1], **Izabela Miechowicz**[1]

[1] Department of Computer Science and Statistics, Poznan University of Medical Sciences, Poland

Abstract. Hypertension is a common disease in highly industrialized societies, more often perceived as a health problem in adults rather than children. However, epidemiologists are currently paying more attention to the possibility of idiopathic hypertension during childhood. This article compares three classification models (logistic regression, classification trees and MARSplines) in order to determine the best classification model and distinguish the parameters that are most important in the detection of abnormal blood pressure in children. The study group consisted of 1,378 children aged between 7 and 18. After making comparisons between the methods, it was determined that MARSplines is the model that best assigns subjects to classes and can be an alternative in cases when traditional statistical methods cannot be used due to a lack of fulfillment of conditions. For prediction of abnormal blood pressure in this age group, the most important parameters were the heart rate and selected indicators of body proportions.

Introduction

Anthropometric measurements are the basic indicators of the health condition of children and adolescents, whose assessment is an essential element of pediatric examinations. This applies not only to standard anthropometric indicators such as height and body weight, but also to indicators describing the development of fat tissue (waist-to-hip and waist-to-height ratios) (Bryl, 2006; Wolański, 1975). Both measurements have been recognized as important indicators of risk factors for cardiovascular disease in adults, as well as elements of metabolic syndrome and an indicator of visceral fat (Krzyżaniak, 2004; Litwin & Niemirska, 2011).

Nowadays we can observe an increase in the number of overweight and obese children and adolescents in Poland (Felińczak & Hama, 2011; Krzyżaniak, 1999; Litwin & Niemirska, 2011; Obuchowicz, 2005) and worldwide (Jackson et al., 2018; Katta & Kokiwar, 2018). Thus, there is a need to

develop and implement preventive programs for cardiovascular disease from early childhood as well as a need for the assessment and monitoring of not only body height and weight but also body circumference (fat level) and blood pressure in children and adolescents, which can provide physicians with important information (Mikoś et al., 2010; Przybylska et al., 2012).

The aim of this study was to develop and compare mathematical models that enable classification of children and adolescents with abnormal and normal blood pressure and, based on the results of these models, to determine the values that are the most important for the detection of abnormal blood pressure among children and adolescents.

Measurement of Arterial Blood Pressure in Children and Adolescents

Measurement of blood pressure is the most important test necessary to diagnose hypertension. However, performing this measurement on children requires the use of a number of procedures necessary for the measurements to be considered reliable and clinically significant. Unfortunately, due to the number of procedures necessary to be performed to obtain correct results, such testing is rarely performed in doctor's offices. There are recommended techniques for measuring blood pressure in children and adolescents that are used to eliminate errors in measurements, while the use of standard techniques by health services allows to make comparisons between studies (National High Blood Pressure Education Program Working Group on Hypertension Control in Children and Adolescents, 1996).

Materials and Methods

The study group included 1,378 healthy children and adolescents, including 557 boys and 821 girls, aged 7 to 18, from randomly selected schools in the Wielkopolska region of Poland. The study was carried out based on a questionnaire prepared as a SCREENING TEST FOR DETERMINING HYPERTENSION. It included both general data and information on the place of residence, type of school and class, gender, and child's age. The study also included the subjects' family history and accounted for the age and education of parents as well as the history of diseases, number of hospitalizations, and any visits to specialists. The children's shoulder, thigh, waist, hips, body mass and height were also measured. Systolic, diastolic

and pulse pressure measurements were performed three times at intervals of two to three weeks with a suitable size cuff. Before the measurements were performed, the children were examined by a pediatrician to exclude any who suffered from chronic diseases of the kidneys or heart, those with endocrinology problems, as well as those having a significant degree of deformity and defects of the osteoarticular system. Anthropometric measurements of height and body weight were made according to the principles of anthropometry and included shoulder, waist and hip circumferences. The group of children qualified with elevated (abnormal) pressure values were those for whom three measurements were equal to or above the 90th percentile, $n_1 = 189$. According to published literature (Jackson et al., 2018; Kowalska et al., 2008; Wyszyńska & Litwin, 2002) and the researchers' expectations, the number of children with abnormal blood pressure was significantly lower for those under the age of 15, both boys and girls (Table 1); hence, only those children who exceeded this age were considered for further analysis.

Table 1. The number of children with normal and abnormal blood pressure

Age (years)	Boys			Girls		
	0-normal n (%)	1-abnormal n (%)	Total	0-normal n (%)	1-abnormal n (%)	Total
7	24 (4)	0 (0)	24	20 (2)	0 (0)	20
8	63 (11)	0 (0)	63	74 (9)	0 (0)	74
9	46 (8)	0 (0)	46	39 (5)	1 (0)	40
10	41 (7)	5 (1)	46	41 (5)	0 (0)	41
11	42 (8)	3 (1)	45	57 (7)	3 (0)	60
12	40 (7)	5 (1)	45	71 (9)	6 (1)	77
13	32 (6)	5 (1)	37	53 (6)	7 (1)	60
14	13 (2)	7 (1)	20	18 (2)	7 (1)	25
15	15 (3)	0 (0)	15	19 (2)	3 (0)	22
16	46 (8)	12 (2)	58	88 (11)	24 (3)	112
17	42 (8)	18 (3)	60	88 (11)	26 (3)	114
18	70 (13)	28 (5)	98	147 (18)	29 (4)	176
Total	474 (85)	83 (15)	557	715 (87)	106 (13)	821

The body proportion indexes used in this study:

$$\text{Quetelet's index: } WQ = \frac{\text{weight [g]}}{\text{height [cm]}},$$

Body Mass Index: $BMI = \dfrac{\text{weight [kg]}}{\text{height}^2\ [\text{m}^2]}$,

Rohrer's index: $WR = \dfrac{\text{weight [kg]} \cdot 10^5}{\text{height}^3\ [\text{m}^3]}$,

Corrected body mass index: $CBMI = \dfrac{\text{weight}^{1.425}\ [\text{kg}^{1.425}] \cdot 71.84}{\text{height}^{1.275}\ [\text{cm}^{1.275}]}$,

Waist-to-hip-ratio: $WHR = \dfrac{\text{waist [cm]}}{\text{hip [cm]}}$,

Waist-to-height-ratio: $WHtR = \dfrac{\text{waist [cm]}}{\text{height [cm]}}$.

To determine the optimal model that would allow the best classification of children and adolescents with normal and abnormal blood pressure, and to determine the relevant parameters, three classification methods were used. In the conducted study, the dichotomous variable is a dependent variable, therefore the following models were used: logistic regression, classification trees and the MARSplines model. In all the models used, the independent variables were: shoulder, waist, hip and thigh circumference – body aspect ratios and the heart rate. Age was omitted due to the homogeneity of the group, consisting of 16–18 year-olds whereas age does not influence other variables to a significant degree. Value *1* was used when a child had abnormal blood pressure value *0* when blood pressure was normal. All analyses were performed separately for boys and girls due to the fact that weight, height and other anthropometric parameters are proved to depend on sex.

The logistic regression model links the probability of one of two possible outcomes of variable Y with independent variables $x_1, x_2, \ldots, x_k$. These variables can be either measurable or qualitative. The Quasi-Newton method was chosen as the method of parameter estimation.

The logistic regression model for the dichotomous dependent variable is defined by the equation:

$$P(Y = 1|x_1, x_2, \ldots, x_k) = \frac{e^{\left(a_0 + \sum_{i=1}^{k} a_i x_i\right)}}{1 + e^{\left(a_0 + \sum_{i=1}^{k} a_i x_i\right)}},$$

where

Y – is a dependent variable that takes the value of 1 or 0

$x_1, x_2, x_3, \ldots, x_k$ – are used as independent variables in the logistic regression model (predictors) (Kleinbaum & Klein, 1994).

Classification trees were the second technique used in the research. To find the best tree, the C&RT method was used, and the Gini ratio was selected as the measure of accuracy, which reaches zero only when one class occurs in a given node while the maximum value is reached when the class sizes in a given node are equal. The FACT method was used as the stopping rule (Breiman et al., 1984).

The third method used was the Multivariate Adaptive Regression Splines model (MARSplines). The MARSplines model shows dependence using a set of coefficients and basis functions that are fully determined by data (Hastie et al., 2001).

$$y = f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X)$$

where

M – number of basis functions,

$h_m(X)$ – basis functions.

To find the best model in boys, 3,296 basis functions and eight interactions were used. For girls, 4,824 basis functions with a degree of interaction equal to six were needed to construct the model. The measure of adjustment in this model is the error of the Generalized Cross Validation, which takes into account the residual error and the complexity of the model and is expressed by the following formula:

$$GCV = \frac{\sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2}{\left(1 - \frac{C}{n}\right)^2},$$

where $C = 1 + cd$, n is the number of cases, d is the number of degrees of freedom equal to the number of independent basis functions. Parameter c depends on d and it is known from experience that when $2 < d < 3$ then c is the most desirable (Hastie et al., 2001).

Statistical analysis was performed using StatSoft, Inc. (2014) STATISTICA (data analysis software system), version 12 (www.statsoft.com).

Results

In the proposed logistic regression models, for both boys and girls, the heart rate was a statistically significant parameter at $p < 0.0001$. For boys, the second parameter entered into the model was waist/height

ratio at $p < 0.0001$; for girls, it was thigh circumference at $p < 0.0001$. The proposed logistic model was able to correctly predict which classes the cases belonged to for 76% of boys; for girls, the number was higher at 83%. The resulting odds ratios for both models indicate good classification performance: 6.54 for boys and 13.35 for girls. The sensitivity of the logistics model was 26% for boys, with a specificity of 95%; for girls, the sensitivity was 25%, with a specificity of 95%. Therefore, in both cases the specificity was very good. It can be said that the model based on logistic regression is suitable for hypertensive disease, due to the fact that with difficult and long-lasting pharmacological treatment, good detection of healthy children (high specificity) is of great importance (we do not want to harm the healthy ones by exposing them to possible unnecessary invasive testing or drug administration). Tables 2, 3, 4 show how the OR (Odds Ratio) changes depending on: increased heart rate in boys and girls, waist/height ratio in boys, and thigh circumference in girls. An increase in these parameters in the proposed logistic regression model increases the chance of abnormal blood pressure in children.

Table 2. Odds ratio for heart rate for girls and boys

Heart rate	Boys			Girls		
	n	OR	95% CI	n	OR	95% CI
< 65	9	references	–	4	references	–
$\langle 65 - 75)$	13	1.1	0.43–2.80	19	2.3	0.77–7.22
$\langle 75 - 85)$	27	2.9	1.22–6.77	31	4.9	1.66–14.64
$\langle 85 - 95)$	4	6.7	1.27–35.04	11	6.7	1.94–22.85
> 95	5	6.2	1.40–27.93	14	27.6	7.27–104.5

n – number of children with abnormal blood pressure

Table 3. Odds ratio for waist-to-height-ratio for boys

waist-to-height	n	OR	95% CI
$\langle 0.35 - 0.40)$	5	references	–
$\langle 0.40 - 0.45)$	27	2.7	0.98–7.67
$\langle 0.45 - 0.50)$	18	3.8	1.27–11.31
$\langle 0.50 - 0.55)$	5	4.7	1.10–20.36
$\langle 0.55 - 0.65)$	3	22.8	1.97–96.60

n – number of children with abnormal blood pressure

Table 4. Odds ratio for circumference of thigh for girls

circumference of thigh	n	OR	95% CI
$\langle 40-45)$	1	references	–
$\langle 45-50)$	13	2.2	0.28–18.10
$\langle 50-55)$	25	3.8	0.49–29.52
$\langle 55-60)$	23	10.9	1.38–86.94
$\langle 60-67\rangle$	17	26.1	3.09–98.10

n – number of children with abnormal blood pressure

The generated classification tree was able to correctly predict which classes the cases belonged to for 73% of boys; the result for girls was slightly higher at 80%. The first division criterion for boys was BMI (Figure 2), which divided the group into those who were healthy and those with abnormal blood pressure. In girls, the first variable was hip circumference (Figure 1). Another important division criterion for both boys and girls was the heart rate, followed by waist-height ratio for boys and waist circumference for girls. The sensitivity of the classification tree was 54% for boys, with a specificity of 79%; for girls, the sensitivity was 32%, with a specificity of 91%. In both cases, the specificity was good. Similarly to logistic regression, the model based on classification trees is appropriate for the detection of absence of hypertensive disease, due to its high specificity.

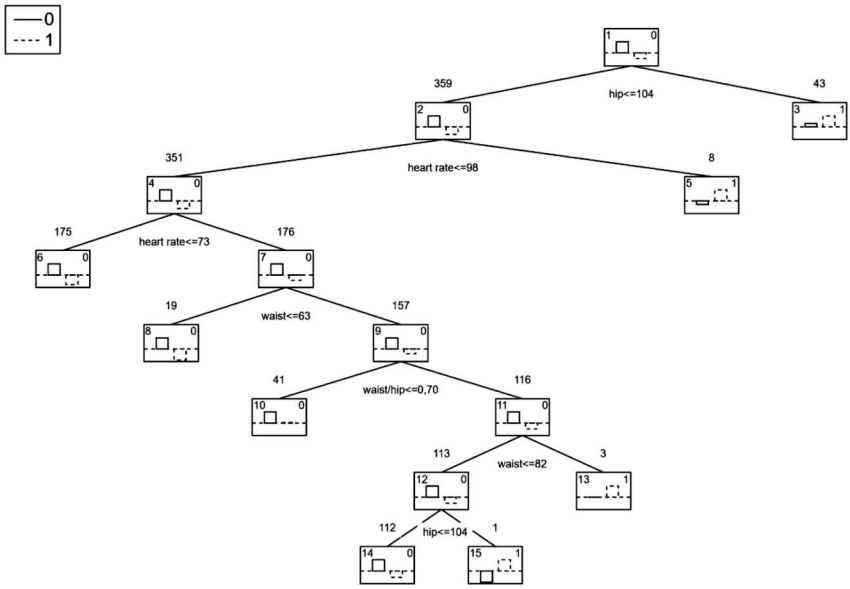

Figure 1. The classification tree for girls

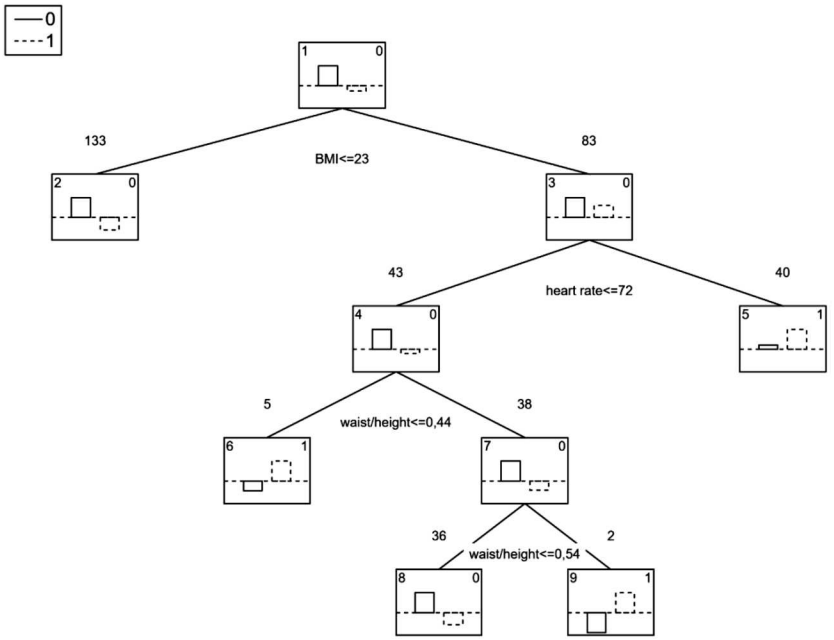

Figure 2. The classification tree for boys

Another proposed method used to classify children to either normal or abnormal blood pressure values was the MARSplines method. The following variables were used to build models for boys and girls: shoulder circumference, waist, thigh, hip and waist/height, waist/hip ratios, body mass index (BMI), Rohrer index (WR), and Quetelet index (WQ). In the resulting models, the heart rate was the predominant predictor in determining the basis functions for both boys and girls (Table 5), followed by thigh

Table 5. Participation of independent variables in MARSplines models

Boys	References to basis function	Girls	References to basis function
waist	2	shoulder	17
hip	4	hip	19
thigh	7	thigh	18
BMI	1	BMI	9
WR	1	WR	15
CBMI	6	heart rate	20
waist/height	6		
heart rate	8		

circumference for boys and hip circumference for girls. The MARSplines method was able to correctly predict case classification for boys at 69% (GCV = 0.31), and for girls at 76% (GCV = 0.24). The sensitivity of the model was 61% for boys, with a specificity of 95%; for girls, the sensitivity was 58%, with a specificity of 99%. In both cases the specificity was very high, close to 100%.

Conclusions

Cross-sectional studies are conducted in every highly developed country and are an important element of assessment of children's and adolescents' health. Nowadays, based on epidemiological studies, it is known that secondary hypertension is increasingly more common in children and adolescents. It is known nowadays that one of the main reasons for such a situation is obesity. It has been confirmed and proved in this study. Unfortunately, due to the difficulties associated with accurate measurement of blood pressure, it is still rarely measured. Therefore, the novel aspect of this study was the development of three statistical models and comparing them to find the best risk predictor of abnormal blood pressure. The authors hope that it will be an important contribution to the diagnostic process. Based on one or two significant predictive variables, physicians can exclude almost 100% of abnormal blood pressure results from the examined minor patients. In the conducted analyzes, MARSplines proved to be the best technique, with the highest sensitivity and specificity. For boys, the two most important predictors influencing abnormal blood pressure were increased heart rate and increased weight/height and waist/height ratios. In girls, the most important factors were increased heart rate and higher values of hip and thigh circumferences. In this study, the research allowed to determine a mathematical model that best classifies children as either healthy or with abnormal blood pressure. The most important variables that significantly influence blood pressure in children and can be used in the algorithm for carrying out hypertension screening test were established. The obtained results may facilitate the detection of children with abnormal blood pressure and thus prevent the consequences of cardiovascular complications.

REFERENCES

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. New York: Chapman & Hall.

Bryl, W. (2006). *Wywiad rodzinny, wskaźniki antropometryczne, wybrane parametry metaboliczne i skuteczność leczenia przeciw nadciśnieniowego u młodzieży z pierwotnym nadciśnieniem tętniczym* (pp. 6–20). Poznań.

Felińczak, A., & Hama, F. (2011). Występowanie zjawiska nadwagi i otyłości wśród dzieci i młodzieży we Wrocławiu. *Pielęgniarstwo i Zdrowie Publiczne*, *1*(1), 11–18.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction* (pp. 266–272, 283–290, 276–278). New York: Springer

Jackson, S. L., Zhang, Z., Wiltz, J. L., Lousstalot, F., Ritchey, M. D., Goodman, A. B., & Yang, Q. (2018). Hypertension Among Youths – United States, 2001–2016. *Morbidity and Mortality Weekly Report*, *67*(27), 758–762.

Katta, A. V., & Kokiwar, P. R. (2018). A Cross-Sectional Study on Correlates of High Blood Pressure among School-Going Children in an Urban Area. *Indian Journal of Community Medicine*, *43*(2), 82–85.

Kleinbaum, D. G., & Klein, M. (1994). *Logistic Regression. A Self-Learning Text.* New York: Springer.

Kowalska, M., Krzych, Ł. J., Siwik, P., & Zawiasa, A. (2008). Uwarunkowania występowania nadciśnienia tętniczego u chłopców i dziewcząt w wieku szkolnym w województwie śląskim. *Nadciśnienie tętnicze*, *12*(4), 269–276.

Krzyżaniak, A. (2004). *Ciśnienie tętnicze u dzieci i młodzieży – normy, monitorowanie, profilaktyka.* Poznań: Wydawnictwo Akademii Medycznej w Poznaniu.

Krzyżaniak, A. (1999). *Ciśnienie tętnicze krwi dzieci i młodzieży miasta Poznania w latach 1986 i 1996. Uwarunkowania, kierunek zmian, normy.* Poznań: Wydawnictwo Akademii Medycznej w Poznaniu.

Litwin, M., & Niemirska, A. (2011). Nadciśnienie tętnicze pierwotne i zaburzenia metaboliczne u dzieci i młodzieży. *Forum Zaburzeń Metabolicznych*, *2*(2), 124–131.

Mikoś, M., Mikoś, M., Mikoś, H., Obara-Moszyńska, M., & Niedziela, M. (2010). Nadwaga i otyłość u dzieci i młodzieży. *Nowiny Lekarskie*, *79*(5), 397–402.

National High Blood Pressure Education Program Working Group on Hypertension Control in Children and Adolescents. (1996). Update on the 1987 Task Force Report on High Blood Pressure in Children and Adolescents. *Pediatrics*, *98*(4), 649–658.

Obuchowicz, A. (2005). Epidemiologia nadwagi i otyłości – narastającego problemu zdrowotnego w populacji dzieci i młodzieży. *Endokrynologia, Otyłość i Zaburzenia Przemiany Materii*, *1*(3), 9–12.

Przybylska, D., Kurowska, M., & Przybylski, P. (2012). Otyłość i nadwaga w populacji rozwojowej. *Hygeia Public Health*, *47*(1), 28–35.

Wolański, N. (1975). *Metody kontroli i normy rozwoju dzieci i młodzieży.* Warszawa: PZWL.

Wyszyńska, T., & Litwin, M. (2002). *Nadciśnienie tętnicze u dzieci i młodzieży* (pp. 9–60). Warszawa: PZWL.