Data mining analysis of factors influencing children's blood pressure in a nation-wide health survey

Piotr Wąsiewicz[1], Zbigniew Kułaga[2]
Mieczysław Litwin[3]

[1]Institute of Electronic Systems, Warsaw University of Technology
[2] Public Health Division, the Children's Memorial Health Institute,
[3]Department of Research, the Children's Memorial Health Institute, Warsaw, Poland

ABSTRACT

Blood pressure in childhood and adolescents is important indicator of good health and strong predictor of BP in adulthood. Genetic susceptibility, environmental and socioeconomic factors are related both with life style, obesity and cardiovascular risk including elevated BP. Increased body mass index is strictly correlated with BP, and obesity and overweight is main intermediate phenotype of childhood hypertension. However, despite current obesity epidemic available data do not fully support the hypothesis that it has resulted in increase of BP in children. We analysed data obtained from 7591 children participating in nation-wide health survey using data mining methodology. Results reveal relationships of obesity and high blood pressure with school environment characteristics.

Keywords: blood pressure (BP), obesity, socioeconomic factors, health survey, multidimensional scaling, decision tree

1. INTRODUCTION

Health is determined by human biology, health care systems, environment, and lifestyle of which lifestyle and environmental determinants have major impact and are both potentially modifiable[1] and it is well established that health depends on socioeconomic circumstances[2]. Blood pressure measurement is commonly used health indicator.Hypertension is one of the most important risk factors of of cardiovascular disease. Although clinically overt cardiovascular disease is rare in childhood, elevated blood pressure in youth may lead to target organ damage [3]. On the other hand level of blood pressure in childhood is strong predictor of adult blood pressure level. In childhood and adolescence body mass has large and positive effect on blood pressure[5]. Obesity and overweight are also the dominant intermediate phenotype of childhood hypertension[6]. Prevalence of obesity among children and adolescents is increasing worldwide[7] and the process is highly dynamic in some countries[8]. International definition for children and adolescents overweight and obesity, introduced in 2000[9], based on age and sex specific body mass index (BMI) cut off points, provides opportunity to study blood pressure according to the system which connects definitions of overweight and obesity in childhdren and adult. Underweight in childhood, as opposite to obesity, is not less (compared to obesity) important issue concerning health of children. Introduction of underweight in children and adolescents definition[10], based on the same system of BMI cut of points and logic assumption - linking child's BMI with BMI in adult - enables international comparisons of, not only weight extremes prevalence, but also high blood pressure prevalence in connection with body weight trends over time and between countries. Increased odds of high blood pressure in connection with overweight and obesity in children and adolescents were reported by Urrutia-Rojas X. et al[11] and Salvadori M et al.[12] however recent meta-analysis on obesity epidemic in children and adolescents and blood pressure has questioned influence of obesity on elevated BP in children and adolescents[13]. It seems that relation between BMI and blood pressure is also confounded by other factors related to life style and socioeconomic status. To evaluate this hypothesis we measured associations between underweight, normal weight, overweight, obesity socioeconomic factors, blood pressure and prevalence of elevated blood pressure obtained on a single occasion in representative sample of children and adolescents 7-18 years of age in Poland. We used classical methods of data mining: decision trees and Kruskal multidimensional scaling (MDS). Data mining is a new discipline, which emerged at the intersection of statistics, databases, pattern recognition, visualization, artificial intelligence and many other methodologies. It relies very heavily on statistical models and methods, but its purpose is quite another one. It is often used for very large databases quick analysis, where to research, what to focus on, what data sector to choose for futher investigation.

2. SUBJECTS AND METHODS

The data for this cross-sectional study came from OLAF study – blood pressure reference ranges establishment for polish children and adolescents. Study participants (children and adolescents 6.5-18.5 years of age) were randomly selected in two-stage sampling. Primary units, schools were sampled from all schools in Poland sampling frame with probability proportional to size. On the second stage, study participants were sampled from all pupils at school sampling frame. All subjects and theirs' parents (in case of subjects under 18 years of age) gave their informed consent to participate in the study (parents and subjects over 16 years of age gave consent in writing). Subjects examinations were conducted from November 2007 till June 2008. Before study commencement approval of The Children's Memorial Health Institute Ethics Committee to conduct the study was obtained.

Medical history of study participant including past, present diseases and medication was taken from parents. Examinations were conducted in school nurses' offices. Each subject's general health status was assessed by physician. Height was measured by a stadiometer (SECA 214) in standing position (with no shoes) to the nearest millimetre as is seen on the right in Fig.1. Weight was measured in light underwear using calibrated, digital, medical scale with 50g accuracy (Radwag WPT 100/200). In case of weight and height two measurements were taken from each subject. If difference between two measurement was 5 mm or more (in case of height) and 300g or more (in case of weight) third measurement was done.

Blood pressure was measured by shown in Fig. 1 Datascope Accutor Plus automated, oscillometric device validated for use in children [14]. Four cuff sizes were available (child cuff, small adult cuff, adult cuff, and large adult cuff). The appropriate cuff size was determined by measurement of the mid-upper arm circumference. The cuff was applied to the right arm, wrapped to a tightness allowing two fingers to be inserted under the cuff. Triplicate blood pressure measurement in 1-2 minutes intervals was taken after 5-10 minute rest in sitting position with arm supported and feet on the floor. The mean of second and third measurement was used in analysis.

Waist and hip circumference were measured in duplicate with flexible anthropometric tape. Skinfolds: triceps, subscapular and abdominal were measured in duplicate with skin fold caliper (Harpenden) depicted in the middle of Fig. 1.

Fig. 1 On the left oscyllometric BP, in the middle skin folds calipers, on the right a stadiometer.

3. LIFE STYLE AND SOCIO-ECONOMIC DETERMINANTS QUESTIONNNAIRE; SCHOOL ENVIRONMENT QUESTIONNAIRE

Questionnaire on socio-economic factors: family structure, parents education, source of income, income category was filled by parents. The questionnaire included also information on child's overall physical activity and exposure to television and computer. In turn, school environment questionnaire was filled by school master or his/her delegate. The questionnaire included information on availability of energy-dense food(s) at school, food advertisements, school canteen and percentage of pupils who eat lunch at school, physical activity infrastructure and its usage.

4. STATISTICAL ANALYSIS: DATA MINING METHODS

Data Mining is the analysis of often large observational data sets to discover unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner[17]. It is an all-purpose toolbox, which contains such methods that allow of revealing new patterns and dependencies in researched data among them classical ones: decision trees and multidimensional scaling (MDS).

A decision tree is more a data mining visualisation tool than a classifier. It is described in a internet dictionary as „a predictive model; that is, a mapping from observations about an item to conclusions about its target value". In computer science it uses a easy to understand graph of node decisions (attribute tests, which each branch is equal to one attribute value) and their possible consequences at tree leaves. It is generated in learning process with a help of a training set.

Another method of classification visualization only with the help of raw data and special distance measures between examined children – points in the multivariable space is multidimensional scaling (MDS). Such the multidimensional space can be visualized in two "artificial" dimensions after scaling based e.g. on the general dissimilarity coefficient of Gower[15] and Kruskal nonmetric MDS[16], where each point dissimilarities with other ones are transferred into two dimensional spacial distances between them. MDS participates in the new emphasis on methods of data analysis which are exploratory. Its value is not helping to measure something accurately, nor in determining how accurate a measurement is. Instead, it helps provide insight into relationships among the objects of the domain.[16]

5. DATA MINING RESULTS

From November 2007 till June 2008 7591 children and adolescents 6.5-18.5 years of age were enrolled to the study all over Poland. The number of girls was 3946 and number of boys was 3645. 6925 children and adolescents were healthy (free of chronic diseases and current medication influencing blood pressure) and were included with the rest of children in the analyses. BMI was calculated by dividing weight in kilograms by height in meters squared. To compare blood pressure between the categories of: underweight, overweight, obesity and normal weight, blood pressure measurements and other numeric variables were converted to z-scores per month and pooled according to sex.

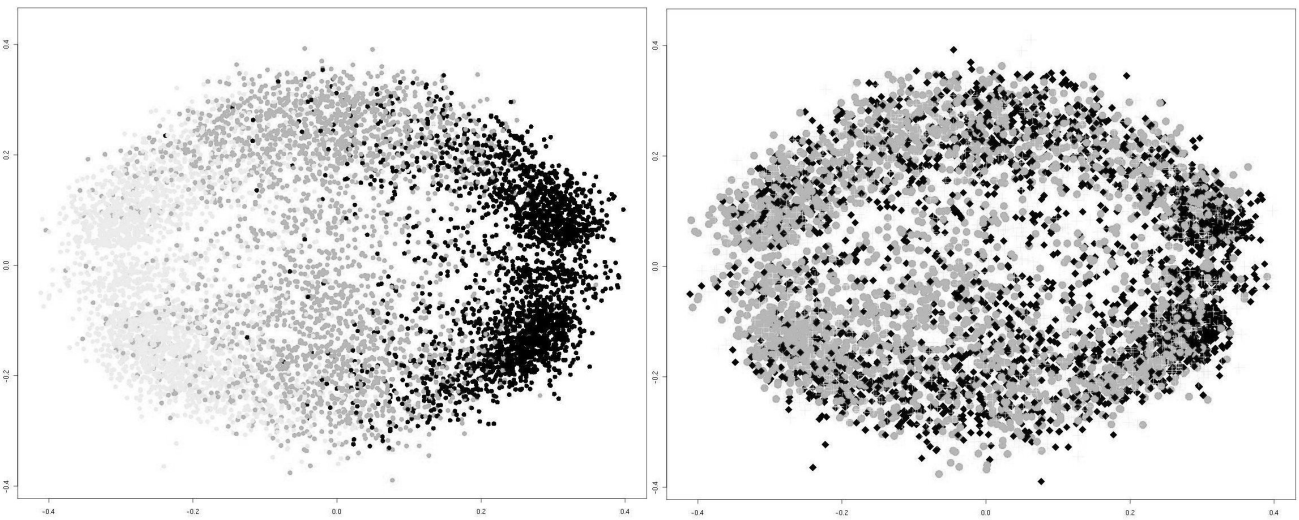

Fig. 2 Children (points in multidimensional space, each variable is one dimension) dissimilarities scaled to two "artificial" dimensions. On the left three levels of BMI: black – higher BMI, dark gray – about BMI mean, light gray – lower BMI. On the right two levels of mean pressure: black – higher mean pressure, dark gray – lower mean pressure.

Visualization multidimensional scaling and decision trees were used to show the odds of high blood pressure resulting from BMI, family background, habitual residence, parent income, parent education level. Normal BMI and mean pressure were reference categories. Analyses were conducted by using R environment (wiki.r-project.org).

With the help of the Grover function dissimilarities between chidren were calculated based on data with the following variables: sex, birthweigth, arm circumference, cuff, distance to a school, footbicycleuse, footbicycledistance, lift to a school, physical education, mother education level, father education level, family income per person, room(own or with other), computer(with internet, without internet), tv, height, weight, waist circumference, hip circumference, skinfolds: triceps, abdomen, subscapular, systolic pressure, diastolic pressure, mean pressure, pulse, children number in a family, BMI, age in months, codeplace, measurement hour. Before calculating dissimilarities all continuous variables were converted to zscore in intervals (made for each sex and each month of child age) and discretized into 3 levels: lower, about mean, high.

Dissimilarities data set was used by the Kruskal function to place children as points on eight two-dimensional pictures (Fig.2-5), where spacial distances are directly proportional to the dissimilarities and gray grades of points are depended on special factor made of some mentioned earlier variables or additional binary ones such as school vending machines, a school shop, advertisements in a school, family income origin, school canteen.

On the left in Fig. 2 three levels of discretized BMI are depicted in a range from black coloured higher values to light gray coloured lower values. On the right side of Fig. 2 two levels of discretized mean pressure are depicted: black coloured higher values and dark gray coloured lower values, mean values of mean pressure are excluded. The higher BMI the higher mean pressure. The correlation was proved later by generating decision trees seen in Fig. 6.

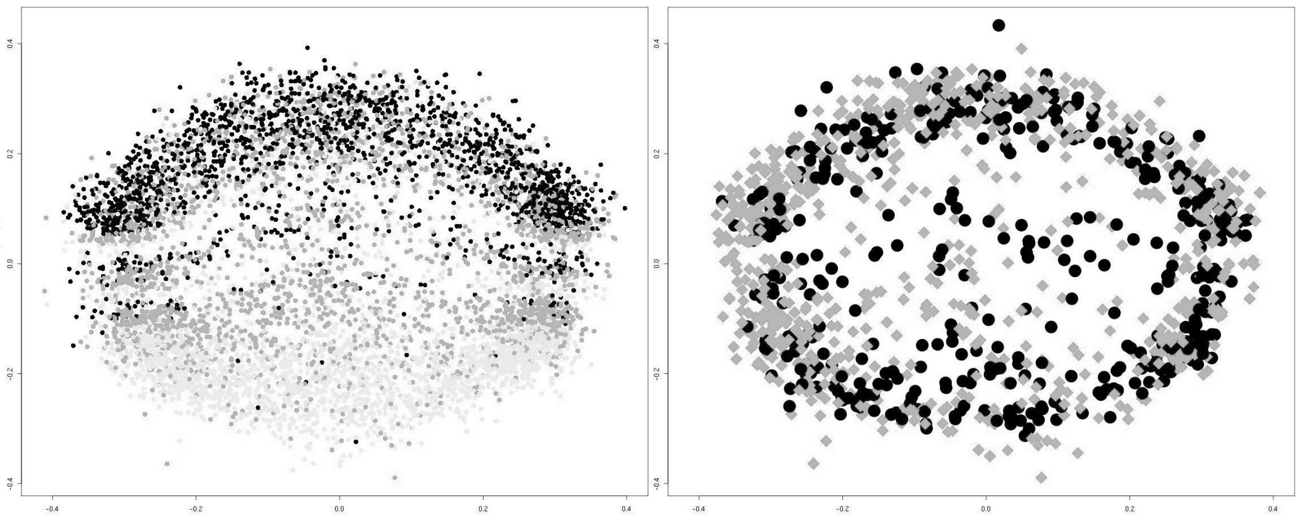

Fig. 3 Children (points in multidimensional space, each variable is one dimension) dissimilarities scaled to two “artificial” dimensions. On the left three levels of distance to a school: light gray – near, dark gray – not near, black – far away. On the right black circles – both mother and father have high education and less than 3 children; dark gray rhombs - both mother and father have only primary education and more than 3 children.

In Fig. 3 (on the left) three levels of discretized distance to a school are depicted in a range from black coloured higher values to light gray coloured lower values. Distance gray grade layers are placed horizontally and BMI layers are vertical. Here distance has no significant influence on BMI or mean pressure and its higher third value is correlated with secondary schools. Data is expressed as zscore for each month and sex, so primary and high primary school pupils are uniform scattered on two-dimensional surface except secondary school pupils. However, creation of decision trees from Fig. 6 proved that distance was chosen from only discretized data (no zscore) as the first node for recognizing BMI and mean pressure levels. So maybe strong correlation between distance and BMI or mean pressure was lost during zscoring and scaling in MDS.

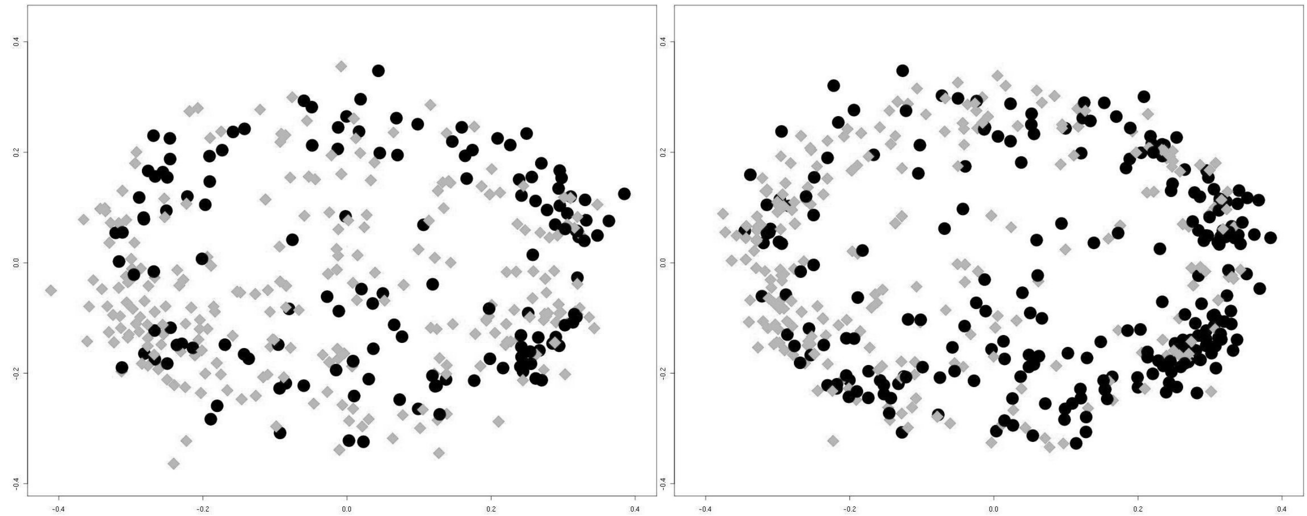

Fig. 4 Children (points in multidimensional space, each variable is one dimension) dissimilarities scaled to two "artificial" dimensions. On the left black circles – in a school there are vending machines, a shop, advertisements; dark gray rhombs – there is no vending machine, no shop, no advertisement, no canteen. On the right black circles – at home the only child has its own room, tv, computer with internet connection; dark gray rhombs – more than 3 children have no own room, no tv, no computer.

In the right picture of Fig. 3 the influence of father and mother education level and the number of their children is confirmed. Gray rhombs are in majority at the left side of the mentioned picture where BMI and mean pressure were at lower level than their means (thin, often with lower pressure children). At the right side of the picture there is more black circles.

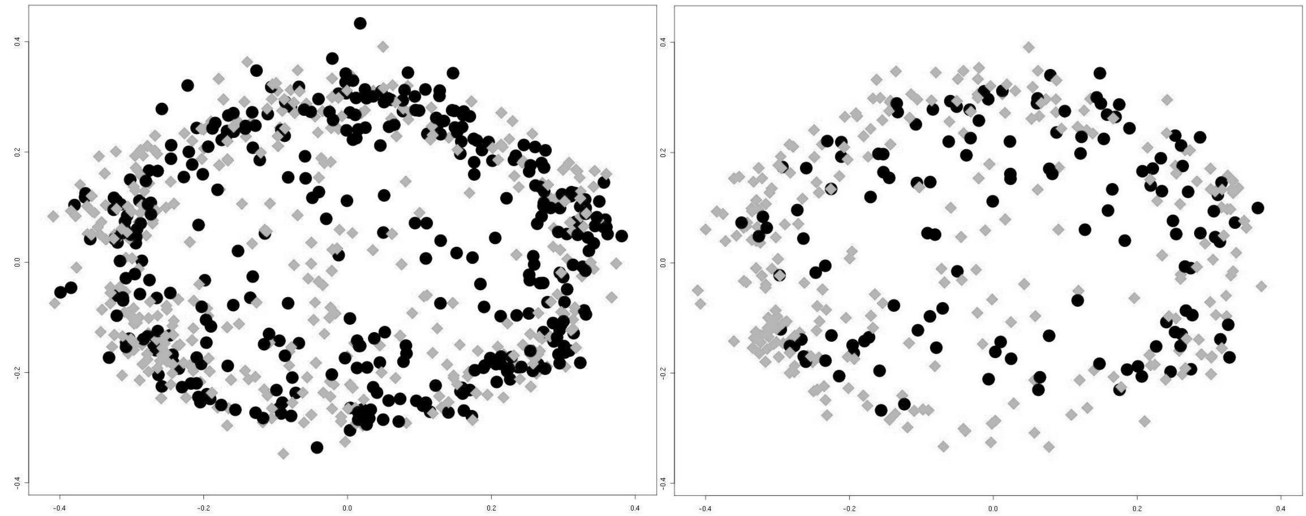

Fig. 5 Children (points in multidimensional space, each variable is one dimension) dissimilarities scaled to two "artificial" dimensions. On the left black circles – parents have high income and work on their own or for somebody; dark gray rhombs - parents have low income and work on their own or for somebody. On the right black circles – rural farm income; dark gray rhombs - parents have a low rent.

The same rule of rhombs and circles positions can be applied to next four pictures from Fig. 4 and 5. Thus, in Fig. 4 obesity is more common among children from schools with vending machines, shops, advertisements or among children who have no sibling, own room, computer with internet connection and tv. In Fig. 5 obesity is more common among

children, whose parents have high income and work on their own or for somebody (both work kinds generated the same orientation of gray and black points) or among children whose parents are farmers.
In turn, children whose schools have no vending machine, no shop, no advertisement, no canteen or who have their room together with their siblings and have no tv, no computer are more often thin (Fig. 4). Underweight is more common among children whose parents have low income (they work on their own or for somebody) or low rents (Fig. 5).

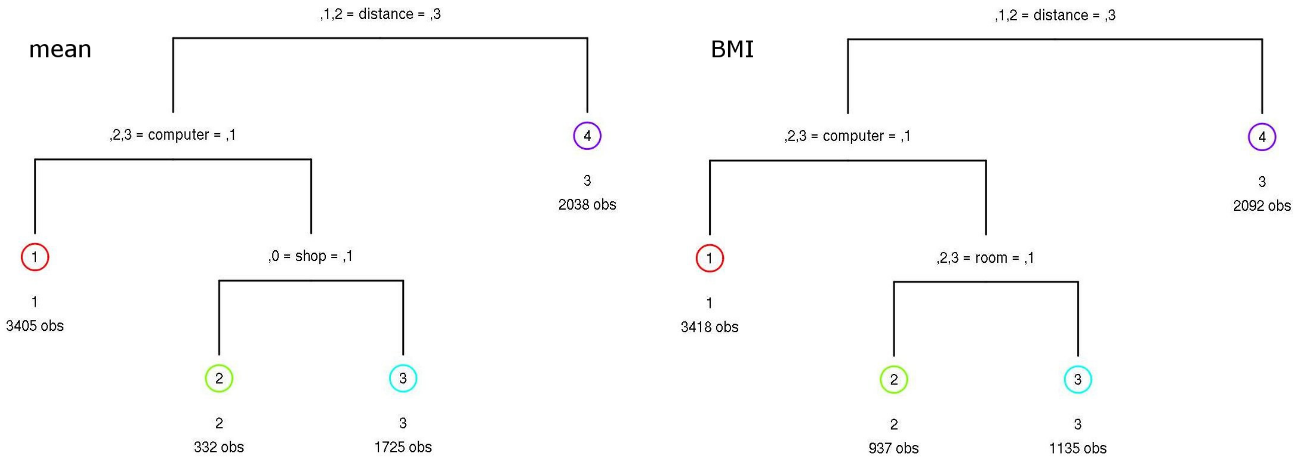

Fig. 6 Decision trees made for mean pressure and BMI. They are quite similar and for the first node distance was chosen from discretized (only) variables used in MDS but age in months, pulse, mean, BMI, sex and strongly correlated with BMI all circumferences. Computer levels: 1 – with internet connection, 2 – without internet, 3 – no computer. Distance levels: 1 – near, 2 – not near, 3 – far away. Room levels: 1- own room, 2 – with siblings, 3 – no room.

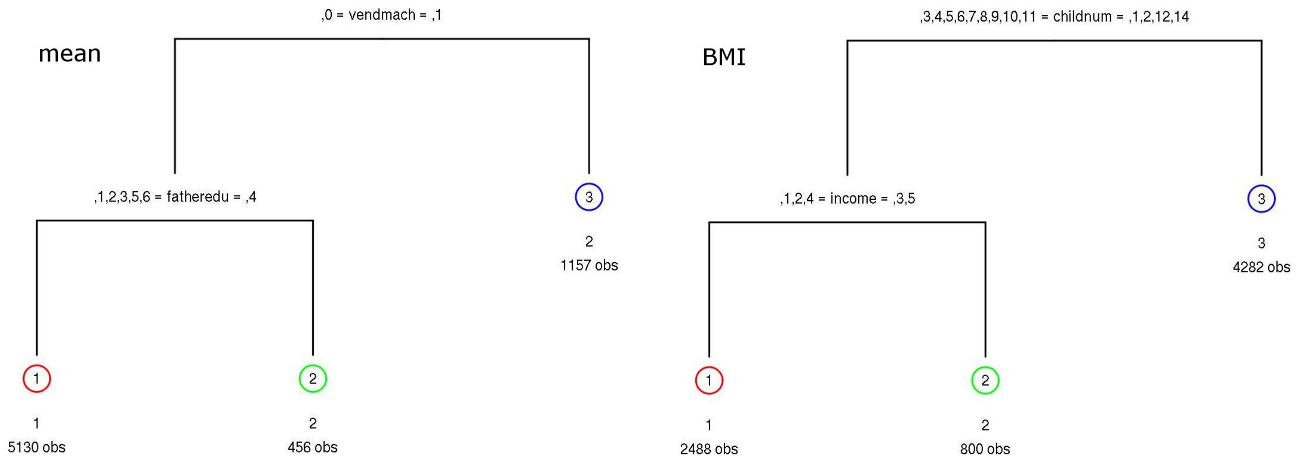

Fig. 7 Decision trees made for mean pressure and BMI. Nodes were chosen from converted to zscore by month, discretized variables used in MDS but age in months, pulse, mean, BMI, sex and strongly correlated with BMI all circumferences. Father education levels: 1 – high education, 2 – secondary, 3 – technical primary, 4 – primary, 5 – part primary, 6 – uknown. Income per person levels: 1 – up to 250 zlotys (4 zlotys about 1 euro), 2 – 251-500, 3 – 501-1000, 4 – 1001-1500, 5 – more than 1500 zlotys.

Decision trees were generated with typical rpart procedure for discretized data (Fig. 6) and for each month and sex converted to zscore, and discretized data (Fig. 7). First ones reveal mean pressure and BMI dependance on a distance to a school and computer with internet connection. Additionaly, high level of mean pressure depends on shop existence at school and high level of BMI on a child own room. Second ones (Fig.7) confirmed lost correlations of distance, but added as the most important new ones for mean pressure vendor machines at school and father education and for BMI – children number in a family and family income per person level.

6. DISCUSSION SUMMARY

Our study results confirm already known associations of higher BMI and higher blood pressure or association of underweight with low income. Revaled associations of obesity and high blood pressure with school environment characteristics are of great interest from public health perspective. An ecological approach to the obesity pandemic (also called "globesity") suggest changes in the environment as being (partially) responsible for the global health problem [18]. Components of obesogenic environment are (among others) availability of energy dense foods and beverages [19]. These factor have been identified by our analysis. Vending machines located at school provide unlimited access to energy dense snaks and high sugar drinks. In addition similar range of products including also fast foods is provided by schools' shops. Association of vending machines at school with higher blood pressure is for the first time (to authors knowledge) reported and justifies further research.

The health survey will proceed until all 15000 participants take part. Data retrieval with a help of data mining techniques increases project added values. Obtained results describe an influence of participants life environment on pressure level, in consequence on their health.

7. ACKNOWLEDGEMENT

The study has been supported by European Economic Area Financial Mechanism and Norwegian Financial Mechanism, and Ministry of Science and Higher Education – Poland; number of grant: PL0080.
Following investigators have participated as OLAF study team and conducted study procedures in the regions: Białystok: Agnieszka Rybi-Szumińska, Michał Szumiński, Katarzyna Taranta-Janusz, Edyta Tenderenda, Anna Wasilewska, Gdańsk: Piotr Czarniak, Dominik Świętoń, Przemysław Szcześniak, Kraków: Monika Miklaszewska, Anna Moczulska, Katarzyna Wilkosz, Katarzyna Zachwieja, Iwona Ogarek, Lublin: Marek Majewski, Aleksandra Sobieszczańska-Droździel, Izabela Szlązak, Paweł Szlązak, Małgorzata Zajączkowska, Łódź: Anna Półtorak-Krawczyk, Marcin Tkaczyk, Opole: Danuta Gmyrek, Poznań: Tomasz Krynicki, Jolanta Sołtysiak, Toruń: Roman Stankiewicz, Sława Zbucka, Warszawa: Jan Szpor, Robert Pietruczuk, Wrocław: Jacek Kleszczyński, Magdalena Naleśniak, Anna Wawro, Irena Wikiera-Magott, Danuta Zwolińska, Zabrze: Piotr Adamczyk, Tatiana Augustyn-Iwachów, Beata Banaszak, Omar Bjanid, Katarzyna Broll-Waśka, Aurelia Morawiec-Knysak.

8. REFERENCES

1. Lalonde M. A New Perspective on the Health of Canadians. Ottawa, Ontario, Canada: Minister of Supply and Services; 1974
2. Davey Smith G, Hart C, Blane D, Gillis C, Hawthorne V. 1997. Lifetime socio-economic position and mortality: prospective observational study. Brit Med J 314: 547-52
3. Litwin M et al. Left ventricular hypertrophy and arterial wall thickening in children with essentials hypertension. Pediatr Nephrol. 2006;21:811-9.
4. Chen X, Wang Y. Tracking of Blood Pressure From Childhood to Adulthood. A Systematic Review and Meta–Regression Analysis. Circulation 2008;117:3171-3180.
5. Jackson L, Thalange NKS, Cole TJ. Blood pressure centiles for Great Britain. Arch. Dis. Child. 2007;92:298-303.
6. Flynn JT, Alderman MH. Characteristics of children with primary hypertension seen at a referral center. *Pediatr Nephrol* 2005; 20:961–966
7. Lobstein T, Baur L, Uauy R. Obesity in children and young people: a crisis in public health. Obes Rev 2004;5(suppl 1):4-85.
8. Kułaga Z, Krzyżaniak A, Palczewska I, Barwicka K. Rapid westernization of Warsaw (Poland) children obesity prevalence comparison with Australia and USA population. Ann Diag Paed Pathol 2007, 11:71-75.
9. Cole TJ, Bellizzi MC, Flegal KM, Dietz WH. Establishing a standard definition for child overweight and obesity worldwide: international survey. BMJ 2000;320:1240-3.
10. Cole TJ, Flegal KM, Nicholls D, Jackson AA. Body mass index cut offs to define thinness in children and adolescents: international survey. BMJ 2007;335:194.

11. Urrutia-Rojas X, Egbuchunam CU, Bae S et al. High blood pressure in school children: prevalence and risk factors. BMC Pediatrics 2006;6:32doi:10.1186/1471-2431-6-32.
12. Salvadori M, Sontrop JM, Garg AX et al. Elevated Blood Pressure in Relation to Overweight and obesity Among Children in a Rural Canadian Community. Pediatrics 2008;122;e821-e827.
13. Chiolero A, Bovet P, Paradis G, Paccaud F. Has Blood Pressure Increased in Children in Response to the Obesity Epidemic? Pediatric 2007;119:544-553.
14. Wong Sik-Nin, Yn Tz Sung R, and Chuk-Kwan Leung L. Validation of three oscillometric blood pressure devices against auscultatory mercury sphygmomanometer in children. Blood Pressure Monitoring 2006;11:281-291.
15. Gower, J. C. A general coefficient of similarity and some of its properties, Biometrics 1971, 27, 623-637.
16. Kruskal, J B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29:1-27, 1964.
17. Hand, D., Mannila, H., & Smyth, P. .Principles of data mining. Adaptive Computation and Machine Learning Series. MIT Press 2001.
18. Egger G, Swinburn B. An "ecological" approach to the obesity pandemic. BMJ 1997;315:477-480.
19. Swinburn B, Egger G. The runaway weight gain train: too many accelerators, not enough brakes. BMJ 2004;329:736-739.