Visual Assessment of Endemic Nephropathy Markers Relationship

Zdenko SONICKI[a,1], Ante CVITKOVIĆ[b], Karen L. EDWARDS[c], Marica MILETIĆ-MEDVED[b], Dubravka ČVORIŠĆEC[d], Vladimir BABUŠ[a], Bojan JELAKOVIĆ[d]
[a] *Andrija Štampar School of Public Health, Medical School, University of Zagreb, Croatia*
[b] *Institute of Public Health Brodsko Posavska County, Slavonski Brod, Croatia*
[c] *School of Public Health, University of Washington, USA*
[d] *Zagreb University Hospital Center, Medical School, University of Zagreb, Croatia*

Abstract. The aim of this paper was to assess relationship between possible endemic nephropathy (EN) markers visually by the CoPlot methodology, and to illustrate this promising data analysis approach. From 912 screened persons in 3 Croatian endemic villages, 25 persons were diagnosed as confirmed EN patients, 371 as non-EN, and the remainder were classified as suspected of having EN, or at risk. Data on 25 confirmed EN patients were matched with appropriate non-EN examinees. All records with missing data were excluded, resulting in 35 subjects with complete data on the 13 key EN variables for CoPlot mapping. CoPlot solution met the accepted goodness of fit measure thresholds. Result showed relationship between EN markers, identifying some nearly duplicated variables, and possible outliers needing some subsequent analysis.

Keywords. visualization, multivariate analysis, endemic nephropathy, markers

Introduction

Endemic (Balkan) nephropathy (EN) is chronic tubulointerstitial nephritis with insidious onset, slow progression to end stage renal disease, and strong association with urothelial carcinoma of the upper urinary tract. It was first described 50 years ago in several small, discrete farming communities along the Danube River and its major tributaries. EN occurs more frequently in some households, but is not an inherited disease. Despite 50 years of research, the etiology of this chronic kidney disease remains unknown [1, 2, 3]. An effort to diagnose the disease early in its course is challenged by the fact that there are no specific diagnostic markers. Thus, several diagnostic markers are required to make the presumptive diagnosis. Different EN centers are using diagnostic criteria which utilize similar diagnostic elements, but with different combinations and cut-off values [4]. To address this problem, potential EN markers, their relationship, interactions, and possible redundancy should be explored. This task is straightforward if sufficient data are available.

Regression analysis and other multivariate methods are one approach to this problem. However, existing datasets are limited by small numbers of observations and

[1] Corresponding Author: Zdenko Sonicki, Andrija Štampar School of Public Health, Rockefellerova 4, 10000 Zagreb, Croatia; e-mail: Zdenko.Sonicki@snz.hr

key variables for EN, resulting in inadequate power. To explore interactions and relationships between variables, an alternative approach for reducing multidimensional data into lower dimensional structure that uncovers the hidden structure is needed. Plotting a map where similarities between observations are shown is a way to illustrate the relationship between all available observations. Multidimensional Scaling (MDS) facilitates that kind of analysis, but has some key limitations for this task. Final MDS map does not display observations, just variables and its relationship, and axes used to create a map have no inherent meaning, limiting map's interpretability.

Adaptation of MDS, called CoPlot, addresses these limitations [5]. Importantly, CoPlot allows analysis of a dataset where the number of variables is greater then the number of observations. This is essential for studying rare diseases, such as EN. CoPlot map could also be used to identify outliers and errors in the data, assessment of the relationships within the data, and for selection of key variables for subsequent analysis. It also identifies redundant variables and observations. These results can then be used to inform development of a standardized definition for EN.

The aim of this paper was to assess the relationship between possible EN markers visually by the CoPlot methodology, and to illustrate this promising data analysis approach.

1. Examinees and Methods

In 2005 a population-based field survey was performed in 3 of the 14 Croatian endemic villages, and 1 control village [6]. During that survey 1081 persons underwent screening, 912 from endemic region: 450 from Kaniža (village with 824 inhabitants), 254 from Bebrina (village with 521 inhabitants) and 208 from Banovci (village with 400 inhabitants). Epidemiologic, laboratory and clinical data were collected. According to WHO diagnostic criteria [5], 25 persons were diagnosed as confirmed EN patients, 371 as non-EN, and the remainder were classified as suspected of having EN, or at risk.

1.1. EN Markers Selection

According to previous research, information on possible environmental exposures, including number of years spent in endemic area, positive family history on EN, as well as information on age, and gender were considered as possible epidemiological markers [1]. Serum creatinine levels and albuminuria were included as potential markers for early kidney damage, while alpha 1 microglobulinuria was included as a marker for tubular proteinuria. Glomerular filtration rate was estimated using Cockroft-Gault equation and abbreviated MDRD formula adjusted to age and gender. Glomerular filtration rate was considered as a marker of global renal function. In addition, as anemia is considered one of important characteristics of EN, red blood cell count, hemoglobin and hematocrite were also included. This resulted in 13 potential markers for diagnosing EN.

1.2. Sampling Procedure

Twenty-five non EN examinees were matched to the 25 confirmed EN patients for age, gender, and village for visual assessment of EN markers. However, complete data is needed to plot observations. Therefore all records with missing data were excluded,

resulting in 35 subjects with complete data on the 13 key EN variables for CoPlot mapping. Twelve subjects (7 women, 5 men) were confirmed EN patients, while 23 (11 women, 12 men) were non-EN examinees. Median age for confirmed EN patients was 71.5 years, and interquartile range (IQR) was 7 years, while non-EN group was 71 years median age, and 9 years IQR. Median of exposure in endemic region was 71.5 years for confirmed EN patients, IQR was 9 years, while exposure median for non-EN group was 71 years, and IQR was 10 years.

1.3. The CoPlot Methodology

CoPlot applies superimposing of two plots in sequence. First plot uses MDS to present distances between observations, while second plot, conditioned on the first plot, displays vectors of relationships among variables. Those vectors describe correlations among the variables. Distances between each of the observations are calculated with the distance metrics called city-block distance. Instead of distance matrix, CoPlot uses non-metric MDS based on Smallest Space Analysis [8, 9], to produce a two-dimensional plot of n observations. A second plot, superimposed on first, consisting of vectors for each variable, is calculated by least-squares regression, so that the correlation of the values of variable and projections of each observation is maximized. The length of each vector is proportional to the correlation between the original data of that variable and the projections of the observations onto the vector. In order to describe how good the plot represents the observations and how good another plot represents the variables goodness of fit diagnostics are applied. Relative loss of information that appears when the multidimensional data are transformed into two dimensions is measured by coefficient of alienation [9]. Coefficients of alienation less then 0.15 provide maps that fit the data well. Also, a magnitude of maximal correlations could be calculated, and upon that, variables that have low magnitudes of correlations could be eliminated from the analysis.

2. Results

First, the CoPlot solution for the 13 potential EN markers met the accepted goodness of fit measure thresholds, coefficient of alienation was 0.122, and average of correlations was 0.796. CoPlot map, where confirmed EN patients are color coded as black, and non-EN examinees as white circle, is shown in Figure 1. Similar observations are located on map close to one another. Each vector represents one of thirteen EN marker variables. Highly correlated variables are described by vectors close together, and if their correlation is negative, the vectors are going in opposite directions.

It is obvious that markers associated with possible anemia: red blood cell count, hemoglobin and hematocrite are nearly duplicated in providing information for making presumptive EN diagnosis. Similarly, age and time spent in endemic area also show redundancy in our data sample, and they are negatively correlated with variable describing positive family history on EN.

Observations for confirmed EN patients and non-EN examinees are clustered, showing more variability between confirmed EN-patients then non-EN examinees. The analysis also identifies examinees labeled 6, and 30, and patient labeled 24 as possible outliers, needing some subsequent reevaluation.

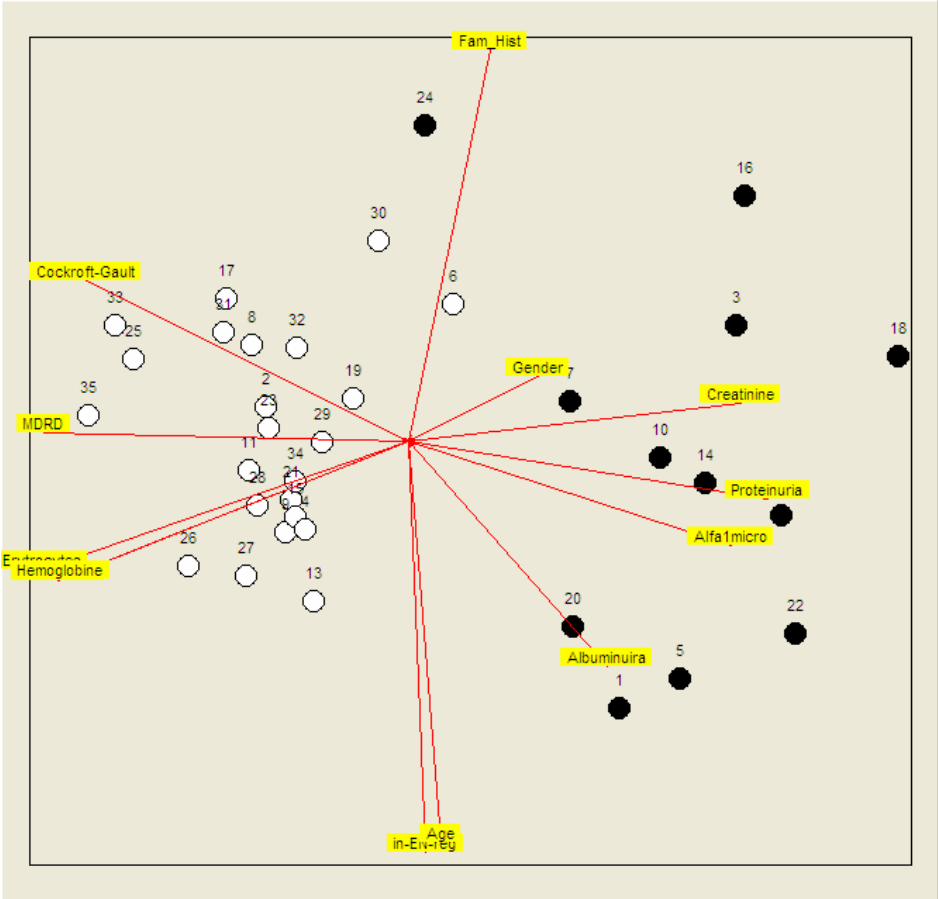

Figure 1. CoPlot Map of 13 EN Markers (Black Circles = confirmed EN; White Circles = non-EN, Vectors = variables). Similar observations are located on map close to one another, highly correlated variables are described by vectors close together, and negatively correlated variables are oriented in opposite directions.

Correlations and orientations of each EN marker are shown in Table 1. Gender can be identified as a variable contributing smallest correlation to the map.

Table 1. Correlations and Orientations of 13 Possible EN Markers

Variable	Orientation	Correlation	Variable	Orientation	Correlation
Gender	21°	0.34	Erythrocytes	-165°	0.94
Age	-84°	0.77	Hemoglobin	-163°	0.90
Time spent in EN area	-87°	0.78	Hematocrite	-163°	0.91
Family History on EN	75°	0.77	Proteinuria	-7°	0.89
Albuminuria	-41°	0.65	MDRD	179°	0.90
Alpha 1 microglobulinuria	-14°	0.82	Cockroft-Gault	159°	0.85
Creatinine	5°	0.82			

3. Conclusion

Visual assessment of EN markers using CoPlot mapping shows good discrimination between confirmed EN patients and non-EN examinees, identifying redundancy of markers associated with possible anemia, and redundancy of markers describing period of exposure in endemic region. Result is suggesting interesting concept, that positive family history on EN is negatively correlated with time spent in endemic area, suggesting that if there is underlying genetic susceptibility, exposure time could be shorter for some individuals. Identification of variables contributing small correlation to the map, and duplicated or nearly duplicated variables, gives possibility to exclude these variables in subsequent analysis in order to identify a more parsimonious set of variables.

Beside other advantages, CoPlot methodology shows possibility to identify relationship in the data which were not detected from exploratory univariate analyses or from regression analyses. This is promising analytical methodology, especially for medical problems with restricted data availability.

Acknowledgements

We thank prof. Ingram Olkin for focusing our attention on CoPlot methodology, and prof. Arthur P. Grollman for his guidance and help in EN research.

Supported by the Croatian Ministry of Science, Education and Sport; research project "Predictive Models in Health Care" No.108-0982560-0257, and research project "Endemic Nephropathy in Croatia; epidemiology, diagnostics, and etiopathogenesis" No.108-0000000-0329.

References

[1] A.P. Grollman, S. Shibutani, M. Moriya, *et al.,* Aristolochic acid and the etiology of endemic (Balkan) nephropathy, *Proc. Natl. Acad. Sci. U.S.A.* ***104*** (2007), 12129–12134.

[2] A.P. Grollman, B. Jelaković, Role of environmental toxins in endemic (Balkan) nephropathy, *J Am Soc Nephrol.* ***18*** (2007), 2817-2823.

[3] T. Hranjec, A. Kovač, J. Kos J, *et al*., Endemic nephropathy: the case for chronic poisoning by Aristolochia, *Croat Med J* ***46*** (2005), 116-125.

[4] D. Čvorišćec, Early diagnosis of endemic nephropathy, *Clin Chim Acta* ***297*** (2000), 85-91.

[5] D.M. Bravata, K.G. Shojania, I. Olkin, A. Raveh, CoPlot: A Tool for Visualizing Multivariate Data in Medicine, *Statistics in Medicine* ***27*** (2007), 2234-2247.

[6] M. Miletić-Medved, B. Jelaković, D. Bistrović, N. Leko, Z. Marić, Epidemiološke osobitosti endemske nefropatije u Hrvatskoj 2005. godine, *Acta Med Croatica* ***61*** (2007), 141-148. (in Croatian)

[7] WHO, The Endemic Nephropathy of South-Eastern Europe, *Bull World Health Organ* ***32*** (1965), 441-448.

[8] L. Guttman, A general non-metric technique for finding the smallest coordinate space for a configuration of points, *Psychometrika* ***33*** (1968), 469-506.

[9] A. Raveh, On measures of monotone association, *Am Statistician* ***40*** (1986), 117-123.