Journal of the Serbian Chemical Society
JSCS@tmf.bg.ac.rs • www.shd.org.rs/JSCS
J. Serb. Chem. Soc. *87* (9) 1025–1033 (2022)
JSCS–5575
Original scientific paper
Published 22 July 2022

Laboratory data clustering in defining population cohorts: Case study on metabolic indicators

IVAN D. PAVIĆEVIĆ[1], GORAN MILJUŠ[2#] and OLGICA NEDIĆ[2*#]

[1]*Institute of Public Health of Belgrade, Belgrade, Serbia and* [2]*University of Belgrade, Institute for the Application of Nuclear Energy (INEP), Belgrade, Serbia*

(Received 6 January, revised 6 April, accepted 27 April 2022)

Abstract: The knowledge on the general population health is important for creating public policies and organization of medical services. However, personal data are often limited, and mathematical models are employed to achieve a general overview. Cluster analysis was used in this study to assess general trends in population health based on laboratory data. Metabolic indicators were chosen to test the model and define population cohorts. Data on blood analysis of 33,049 persons, namely the concentrations of glucose, total cholesterol and triglycerides, were collected in a public health laboratory and used to define metabolic cohorts employing computational data clustering (CLARA method). The population was shown to be distributed in 3 clusters: persons with hypercholesterolemia with or without changes in the concentration of triglycerides or glucose, persons with reference or close to reference concentrations of all three analytes and persons with predominantly elevated all three parameters. Clustering of biochemical data, thus, is a useful statistical tool in defining population groups in respect to certain health aspect.

Keywords: computational model; blood analytes; dependent variables; community health groups.

INTRODUCTION

The knowledge on the general population health is important for government bodies responsible for creating public policies, social organizations, and medical institutions. However, personal data are often limited, and mathematical models are employed to achieve a general overview. Computational data clustering is a method for dividing data into sets with similar characteristics,[1] rooting back to the mid twentieth century.[2] Cluster analysis defines natural groupings of objects based on certain relationships and may identify specific patterns, suggesting the

* Corresponding author. E-mail: olgica@inep.co.rs
# Serbian Chemical Society member.
https://doi.org/10.2298/JSC220106037P

classification of members of the dataset. The clustering large applications (CLARA) algorithm was developed as an appropriate replacement for memory demanding and computationally intensive PAM (partitioning around medoids) algorithm, suitable for the comparison of medians.[3] The CLARA method can deal with data consisting of a large number of objects (thousands) reducing the computing time and memory storage problem.

The cluster analysis was tested in this study using basic biochemical parameters obtained by a routine blood analysis, namely the concentrations of glucose, total cholesterol and triglycerides. These indicators are useful for monitoring metabolic homeostasis or the assessment and following-up of metabolic disorders, such as metabolic syndrome and diabetes mellitus type 2 (DM2).[4,5] Insulin endocrine interplay has wide influence on both catabolic and anabolic processes, from the primary role in glucose homeostasis, across regulation of the levels of fasting free fatty acids in plasma, to the exertion of storage signals for glycogen and lipids.[6–8] Moreover, recent studies on the application of statins have shown backward relationship, from the disruption of cholesterol synthesis to erroneous glucose metabolism caused by the insulin resistance and the development of DM2.[4,9,10] The population studies on high levels of blood glucose and triglycerides have shown that strong correlation between the two parameters exists, as expected from the interconnected metabolic pathways.[8,11] All these findings suggest unique patterns formed between levels of glucose, triglyceride and cholesterol in blood long before the onset of the metabolic disorder.

The aim of this study was to apply data clustering model to define population cohorts using laboratory results without detailed personal (demographic and/or clinical) information. Having in mind the above-mentioned metabolic relations, it seemed relevant to test the model using these three health indicators and define the population cohorts that depend on their interrelation. As far as we are aware, clustering analysis was not applied for such purpose before. The analytical approach described in this article can contribute to the increase of capacity for the surveillance of public health and the creation of action plans which can help to manage health issues in practice.

EXPERIMENTAL

Data collection

Data on blood analysis of 33,049 persons were collected and conceded for this study in a public health laboratory of the Institute for the Application of Nuclear Energy (INEP) in 2019 (this institution is also an owner of the database). The concentrations of glucose, total cholesterol and triglycerides from serum were determined applying IFCC approved methods, on the day of blood collection (after an overnight fasting of the participants). Glucose was measured by GOD-PAP, cholesterol by CHOD-PAP and triglycerides by GPO-PAP method, using commercial reagents (Human GmbH, Wiesbaden, Germany) and automated analyzer (Konelab 20, Thermo Fisher, Finland).

Data analysis

Original data were converted and reorganized from the Microsoft Excel tables to the PostgreSQL relational database model with three tables: participants, type of the analysis and reports (results). After gender assignment, the names and other personal data were removed from data frames, thus protecting participants' anonymity. For statistical analysis, R statistical framework was used (R version 3.6.1 for 64-bit gnu/linux). The analysis was performed taking into consideration the reference ranges of three parameters (3.9–6.1 mM for glucose, < 6.2 mM for total cholesterol and < 2.3 mM for triglycerides), not clinically recommended values (which are lower than reference in the case of cholesterol and triglycerides). The overall data were analyzed both as a single set and divided into subgroups with respect to age, gender, and season (Spring–Summer from April to September and Autumn–Winter from October to March). The results were shown as median values and ranges (2.5^{th}–97.5^{th} percentile).

Statistical tests used in this study were χ^2-test for independence (in testing cluster performance), Mann–Whitney U test for the comparison of two groups within clusters, Kruskal––Wallis H test for the comparison of three groups between clusters, Spearman's rank correlation coefficient for the determination of the correlation between parameters, principal component analysis (PCA) for the reduction of dimensions and variability analysis, and CLARA clustering for the defining of cluster groups. The CLARA method was applied using the clara function from the cluster package with 100 runs, Euclidean distance and PAM-like parameter. The analysis and visualization of the clustered data was performed using the clusplot function from the cluster package.[12]

RESULTS

Data interpretation and formation of clusters

The primary criterion for data collection was the selection of participants with all three parameters analyzed in the same run. Since large clinical data sets seldom follow normal distribution,[13] especially in randomized population studies, the skewness was observed for all three analytes.

Although the connections between gender, age, season and biochemical parameters may be expected, there was no practical reason to form clusters around factor data. However, before clustering biochemical parameters (concentrations of glucose, total cholesterol and triglycerides), a convenient way to test the grouping of variables is to apply the principal component analysis, PCA.[14] The PCA allows data reduction to very few components, if data exhibit acceptable correlations and shows yields of the variables to variances.[14,15] When biochemical parameters were extracted for PCA, the yields of glucose and cholesterol had different directions with triglycerides in-between (Fig. 1). With 80.4 % of variability explained by the first two principal components, the data set used in this study made a promising case for the subsequent cluster analysis.

Clustering only laboratory results opened the possibility of exploring the connections between these three analytes and the formation of clusters. The analysis of the data has shown interesting distribution across clusters (Fig. 2). The distribution of persons was as follows: 15214 in cluster 1, 13873 in cluster 2 and 3962 in cluster 3.

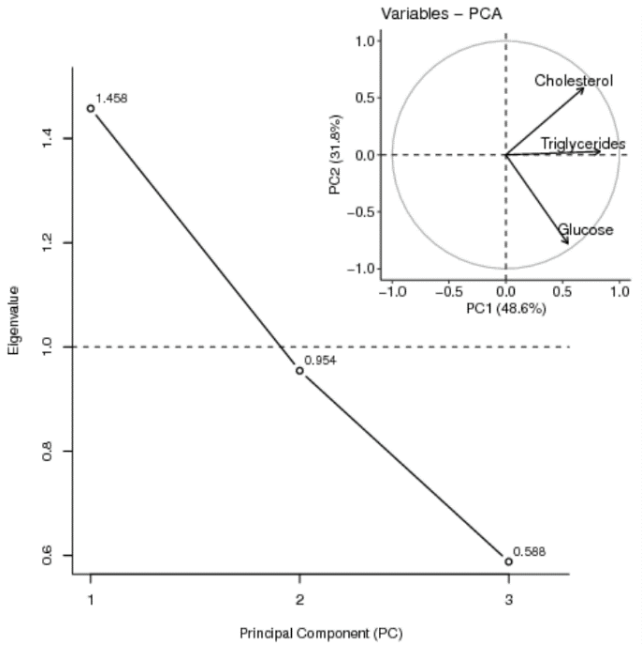

Fig. 1. The PCA variable yields of the biochemical parameters.

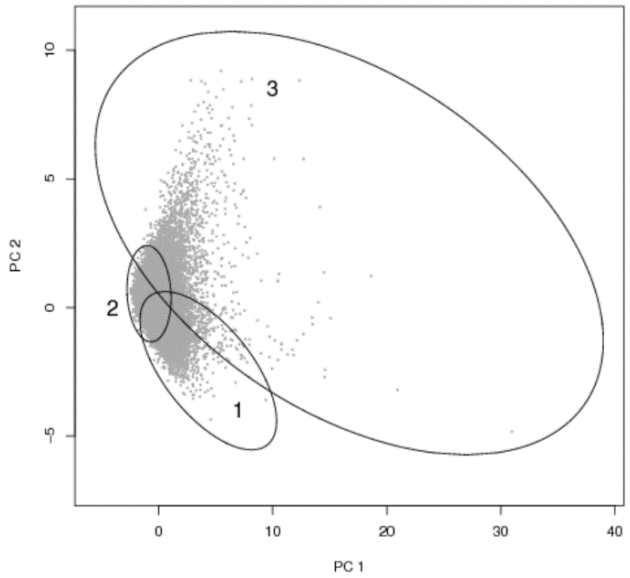

Fig. 2. Clustering of dataset after PCA dimension reduction to 2D. Graphical presentation of three clusters was made using the clusplot funtion from the cluster package. Since the clustering data has three parameters, and function uses PCA for the reduction of dimensions, the corresponding plot shows that two principal components explain 80.4 % of the point variability.

Assigning cluster values to the entire data set revealed the connections between these three blood analytes and health condition, based on the reference borders between health and a disease. Table S-I of the Supplementary material to this paper reviews median values, ranges and a number of persons in each cluster group formed a subgroup taking into consideration gender, season and age.

Without deeper analysis, the results from Table S-I suggested that the concentration of total cholesterol seemed to be independent from the concentrations of glucose and triglycerides in clusters 1 and 3 (different trend in distribution). The greatest data dispersion was seen in the third cluster, whereas the most concentrated data were in the second cluster (Fig. 3).

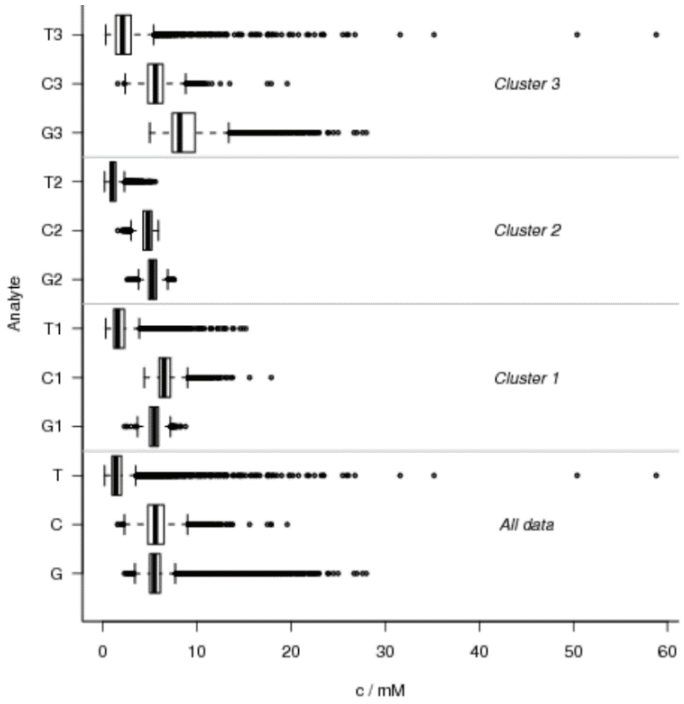

Fig. 3. Boxplots of the original data frame and data sets of the clusters. G, C, T represent glucose, total cholesterol and triglycerides, respectively.

Examination of the data

Kruskal–Wallis H test for the comparison of the concentrations of glucose, total cholesterol and triglycerides between clusters gave $P < 0.005$ in each case, indicating that each cluster represented its own population. Further analysis, taking into consideration gender, age and season, was performed within a cluster and then summed-up for the comparative study.

Gender difference was explored using Mann–Whitney U test. The significant difference was found ($P < 0.005$) for all parameters and in all three clusters. The gender distribution was similar in clusters 1 and 2 (40–45 % of males), which implicated that, at this level, genders tend to cluster at the same ratio around clusters' medians. The third cluster represented a set of outliers from the reference ranges for the investigated parameters, thus, defining strictly a disease group (Table S-I) and within this group gender distribution was slightly changed (55 % males). Furthermore, the second variable was introduced, the age, and three groups were formed: 20–40, 41–60 and > 60 years old. The difference between age groups was analyzed using Kruskal-Wallis H test and the significance was found for each parameter ($P < 0.005$). The most discernible result of this examination was that majority of persons in cluster 3 were more than 60 years old (58 % of males and 75 % of females).

The comparison of the data sorted by the season was made in the same manner. The seasonal difference was significant only for glucose in clusters 1 and 2 (*i.e.*, those without extremely high concentrations) and triglycerides in cluster 1.

The correlation between parameters within one cluster was defined using Spearman's rank correlation coefficient. A correlation was found between the concentrations of total cholesterol and triglycerides in cluster 3 ($r = 0.38$), where most of the persons with high concentrations of these analytes were grouped. The concentrations of glucose and triglycerides exhibited moderate correlation in clusters 1 and 2, but not in cluster 3.

In order to visually overview relations between all the examined characteristics of the population groups in three clusters (*i.e.*, gender, age and season), the simultaneous presentation of the data is given in Fig. 4. This type of presentation can aid in the estimation of general trends in the population health.

DISCUSSION

As already said in Introduction, the aim of this study was to show the applicability of cluster analysis in the assessment of general trends in population health having only limited amount of data, namely laboratory results (without demographic and/or clinical data on persons/patients). We have chosen to demonstrate the usefulness of the model in defining metabolic clusters, thus the results will be only briefly discussed in the context of metabolic diseases without deeper implications.

The grouping of the population by clustering concentrations of glucose, total cholesterol and triglycerides imposed a question whether total cholesterol is a dependent variable or a satellite element to a general lipid metabolism, regulated by insulin. Cholesterol clustering can possibly be explained by specific factors, and in favour of that hypothesis was an unique cholesterol group defined in clus-

ter analysis, which could not be easily linked with evident DM2 and metabolic syndrome. Thus, the population in cluster 1 could be seen as a specific "cholesterol group" independent, or "peculiarly" related to glucose and triglycerides levels. Cluster 3, on the other hand, contained versatile cases with predominantly elevated all three parameters, and most likely associated with metabolic syndrome and/or diabetes. Finally, cluster 2 consisted mostly of persons with reference or close to reference concentrations of all three analytes.

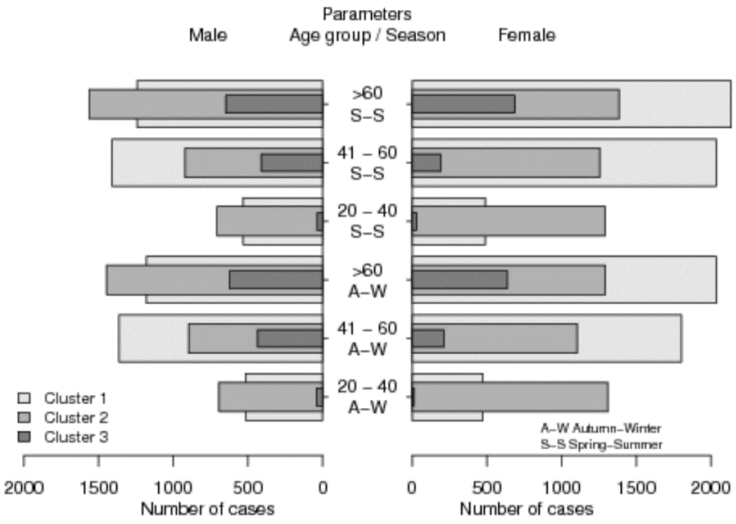

Fig. 4. Comparative graph presenting population in clusters according to the concentration of glucose, total cholesterol and triglycerides, gender, age and season.

Several studies have shown that the concentrations of glucose and triglycerides can be nominated as prediction factors for the assessment of insulin resistance, which guides into obesity, metabolic syndrome and DM2.[8,16,17] Population studies confirmed tight connection between elevated levels of fasting glucose and total cholesterol.[18–20] A new mathematical model for insulin resistance based on those parameters was proposed.[20,21] Our study based on the large data set demonstrated that general population could be divided statistically using cluster analysis into distinguishable groups according to these three parameters. Without going into deeper pathophysiological analysis, as that was not the intention of this article, it is evident that the results reported can serve as a screening method to overview the population health in respect to the basic metabolic indicators. Periodical investigations can demonstrate changes over time, offering a tool to create preventive strategies, or adjust medical capacities for the treatment.

CONCLUSIONS

The proposed model can offer an estimation of the health status of the population in respect to specific parameters (*i.e.*, specific health aspect) having only laboratory results, and can be of a significant assistance in creating public health policies.

SUPPLEMENTARY MATERIAL

Additional data and information are available electronically at the pages of journal website: https://www.shd-pub.org.rs/index.php/JSCS/article/view/11549, or from the corresponding author on request.

Acknowledgments. This work was supported in part by the Ministry of Education, Science and Technological development of the Republic of Serbia (contract number 451-03--9/2021-14/). Data on blood analysis were collected and thereby conceded for this study in the laboratory of INEP (included in the health service in Serbia).

ИЗВОД

КЛАСТЕР АНАЛИЗА ЛАБОРАТОРИЈСКИХ ПОДАТАКА У ЦИЉУ ДЕФИНИСАЊА ПОПУЛАЦИОНИХ ГРУПА: ТЕСТ МОДЕЛ СА МЕТАБОЛИЧКИМ ИНДИКАТОРИМА

ИВАН Д. ПАВИЋЕВИЋ[1], ГОРАН МИЉУШ[2] и ОЛГИЦА НЕДИЋ[2]

[1]Градски завод за јавно здравље Београд, Београд и [2]Универзитет у Београду, Институт за примену нуклеарне енергије (ИНЕП), Београд

Познавање општег здравственог стања популације је важно за креирање јавних политика, као и за организацију медицинске службе. Лични подаци су, међутим, често ограничени и да би се стекао општи увид, примењују се математички модели. Кластер анализа је примењена у овој студији за утврђивање општег здравственог стања популације на основу лабораторијских података. Основни метаболички индикатори су изабрани за тестирање модела. Подаци о концентрацији глукозе, укупног холестерола и триглицерида у крви 33,049 особа су сакупљени у лабораторији јавног здравственог система и коришћени су за дефинисање метаболичких група применом компјутерске кластер анализе (CLARA метод). Сви испитаници су се разврстали у 3 кластера: особе са хиперхолестеролемијом, са или без одступања у концентрацији триглицерида или глукозе, особе са референтним или скоро референтним концентрацијама сва три аналита и особе са претежно повећаним концентрацијама сва три параметра. Показано је да је формирање кластера, користећи биохемијске податке, користан статистички алат за дефинисање популационих група у погледу одређеног здравственог аспекта.

(Примљено 6. јануара, ревидирано 6. априла, прихваћено 27. априла 2022)

REFERENCES

1. N. Sauvageot, A. Schritz, S. Leite, A. Alkerwi, S. Stranges, F. Zannad, S. Streel, A. Hoge, A.-F. Donneau, A. Albert, M. Guillaumeet, *Nutr. J.* **16** (2017) 4 (https://doi.org/10.1186/s12937-017-0226-9)
2. L. Kaufman, P. J. Rousseeuw, *Finding groups in data: An introduction to cluster analysis*, Wiley, New York, 1990 (ISBN 978-0-471-73578-6)
3. C. W. Hu, H. Li, A. A. Qutub, *BMC Bioinform*. **19** (2018) 19 (https://doi.org/10.1186/s12859-018-2022-8)

4. K. G. Parhofer, *Diab. Metab. J*. **39** (2015) 353 (https://doi.org/10.4093/dmj.2015.39.5.353)
5. G. H. Tomkin, D. Owens, *Diab. Metab. Syndr. Obes. Targ. Ther*. **10** (2017) 333 (https://doi.org/10.2147/DMSO.S115855)
6. J. Jensen, P. I. Rustad, A. J. Kolnes, Y. C. Lai, *Front. Physiol*. **2** (2011) 112 (https://doi.org/10.3389/fphys.2011.00112)
7. M. P. Czech, M. Tencerova, D. J. Pedersen, M. Aouadi, *Diabetologia* **56** (2013) 949 (https://doi.org/10.1007/s00125-013-2869-1)
8. C. J. Toro-Huamanchumo, D. Urrunaga-Pastor, M. Guarnizo-Poma, H. Lazaro-Alcantara, S. Paico-Palacios, B. Pantoja-Torres, V. Del Carmen Ranilla-Seguin, V. A. Benites-Zapata, *Diab. Metab. Syndr*. **13** (2019) 272 (https://doi.org/10.1016/j.dsx.2018.09.010)
9. S. Huptas, H. C. Geiss, C. Otto, K. G. Parhofer, *Am. J. Cardiol*. **98** (2006) 66 (https://doi.org/10.1016/j.amjcard.2006.01.055)
10. H. Cederberg, A. Stancakova, N. Yaluri, S. Modi, J. Kuusisto, M. Laakso, *Diabetologia* **58** (2015) 1109 (https://doi.org/10.1007/s00125-015-3528-5)
11. S. B. Lee, M. K. Kim, S. Kang, K. Park, J. H. Kim, S. J. Baik, J. S. Nam, C. W. Ahn, J. S. Park, *Endocrinol. Metab.* **34** (2019) 179 (https://doi.org/10.3803/EnM.2019.34.2.179)
12. A. Kassambara, *Practical guide to cluster analysis in R Unsupervised machine learning*, STHDA, 2017 (www.sthada.com)
13. J. Krithikadatta, *Conserv. Dent*. **17** (2014) 96 (https://doi.org/10.4103/0972-0707.124171)
14. I. T. Jolliffe, *Principal component analysis*, Springer, New York, 2002 (https://doi.org/10.1007/b98835)
15. I. T. Jolliffe, J. Cadima, *Phil. Trans., A* **374** (2016) 20150202 (https://doi.org/10.1098/rsta.2015.0202)
16. F. Abbasi, G. M. Reaven, *Metab. Clin. Exp*. **60** (2011) 60 (https://doi.org/10.1016/j.metabol.2011.04.006)
17. L. E. Simental-Mendia, G. Hernandez-Ronquillo, R. Gomez-Diaz, M. Rodriguez-Moran, F. Guerrero-Romero, *Pediatr. Res.* **82** (2017) 920 (https://doi.org/10.1038/pr.2017.187)
18. T. Du, G. Yuan, M. Zhang, X. Zhou, X. Sun, X. Yu, *Cardiovasc. Diabetol*. **13** (2014) 146 (https://doi.org/10.1186/s12933-014-0146-3)
19. F. Guerrero-Romero, R. Villalobos-Molina, J. R. Jimenez-Flores, L. E. Simental-Mendia, R. Mendez-Cruz, M. Murguia-Romero, M. Rodríguez-Morán, *Arch. Med. Res*. **47** (2016) 382 (https://doi.org/10.1016/j.arcmed.2016.08.012)
20. J. Salazar, V. Bermudez, M. Calvo, L. C. Olivar, E. Luzardo, C. Navarro, H. Mencia, M. Martinez, J. Rivas-Rios, S. Wilches-Duran, M. Cerda, M. Graterol, R. Graterol, C. Garicano, J. Hernandez, J. Rojas, *F1000Res*. **6** (2017) 1337 (https://doi.org/10.12688/f1000research.12170.3)
21. L. E. Simental-Mendia, M. Rodriguez-Moran, F. Guerrero-Romero, *Metab. Syndr. Rel. Disord*. **6** (2008) 299 (https://doi.org/10.1089/met.2008.0034).