Applied Medical Informatics
Original Research Vol. 40, No. 3-4 /2018, pp: 54-62

Hierarchical and Partitional Cluster Analysis of Glucose and Insulin Data from the Oral Glucose Tolerance Test

Miguel ALTUVE

Faculty of Electrical and Electronic Engineering, Pontifical Bolivarian University, Bucaramanga, Colombia.
E-mail: miguel.altuve@upb.edu.co

Received: August 11, 2018/Accepted: December 23, 2018/ Published online: December 27, 2018

Abstract
The body's ability to regulate glucose homeostasis is commonly assessed through the oral glucose tolerance test (OGTT). Several variations of OGTT exists, but the most used in clinical practice is the 2-sample 2-hour OGTT, in which glucose is measured in fasting and two hours after a glucose load. In the 5-sample 2-hour OGTT, glucose is measured in fasting and every 30 minutes after a glucose load, during two hours. In these tests, besides glucose, insulin level can also be measured from the blood samples, increasing thus the number of variables to analyze and perform a better metabolic assessment. In this paper, a cluster analysis is carried using the levels of glucose and insulin from the 2-sample 2-hour OGTT and from the 5-sample 2-hour OGTT, from subjects with metabolic syndrome and professional marathon runners. Different configurations of k-means and agglomerative hierarchical clustering were used to perform the clustering of data and analyze the relationships between clusters with the study groups. Results show that the k-means clustering algorithm performs better than the agglomerative hierarchical clustering, and, with the Manhattan distance measure, k-means perfectly groups subjects using the ten variables from the 5-sample 2-hour OGTT.

Keywords: Data analysis; Metabolism; Information retrieval; Medical technology; Computer applications

Introduction

Normal metabolism in the human body is assured by an appropriate concentration of glucose, the main source of energy for cells, in the blood. Insulin and glucagon pancreatic hormones are in charge of the regulation of plasma glucose concentration through a negative feedback mechanism: insulin secretion is stimulated by an increase in plasma glucose concentration whereas glucagon secretion is stimulated by a reduction of it [1, 2].

The alteration of plasma glucose concentration and the deficiency in insulin secretion may cause distinct symptoms and diseases, compromising the quality of life and well-being of people with these pathological conditions. For instance, low plasma glucose concentration affects the brain and neuronal functioning [3, 4] whereas high plasma glucose concentration may cause atherosclerosis, kidney failure, nerve damage and blindness [5, 6]. The deficiency in insulin secretion may lead to type one diabetes, but the elevated plasma insulin concentration has been associated with diabetes and hypertension [7, 8].

Genetic factors, as well as poor nutrition and a sedentary lifestyle, may cause or aggravate abnormal glucose metabolism. Obesity, fasting hyperglycemia, hypertension, elevated triglycerides and decreased level of HDL cholesterol are characteristic disorders of the metabolic syndrome that

increase the chances of having diabetes and cardiovascular disease [9]. Changes in feeding behavior and doing regular physical activity are powerful actions to take to prevent and treat these disorders [10].

The oral glucose tolerance test (OGTT) is used to check glucose tolerance by measuring the body's ability to metabolize glucose. In the 2-sample 2-hour OGTT, levels of plasma glucose are measured after fasting (8-10 hours of not eating) and then again two hours after ingestion of a dose of glucose, whereas in the 5-sample 2-hour OGTT, five drawings of blood are performed: one after fasting, and four others after glucose intake, at intervals of 30 minutes each draw. Plasma insulin levels may also be measured at these time instants. In this sense, four different variables are available from the 2-sample 2-hour OGTT (two levels of glucose and two levels of insulin, taken in fasting and two hours after glucose intake), and ten variables are available from the 5-sample 2-hour OGTT (five levels of glucose and five levels of insulin). These metabolic indicators are not currently used in the clinical practice neither to provide better follow-up care to patients nor to characterize subjects from distinct populations or metabolic conditions. It has been recently shown that the insulin hormone in marathon runners removes excess glucose from the blood faster than subjects with the metabolic syndrome [11]. In the future, it could be possible to extract the levels of glucose and insulin from the blood using wearable technology and by processing those samples using smartwatches or smartphones alerting the user in case of possible danger can become reality.

Discovering patterns and retrieving clinically useful information manually in glucose and insulin data is a complex and tedious task when the amount of data is considerably large as the obtained in studies for population characterization. The first approach to try out for exploratory data analysis in such cases is cluster analysis, in which data are partitioned into clusters to find natural groups in the dataset. This process uses a measure of similarity (e.g., distance measures) to regulate the aggregation of objects into many clusters, such as objects in the same cluster are similar but different from objects in other clusters [12, 13]. Besides, since the data is not labeled, the clustering process does not use any a priori knowledge for data clustering. That is why this kind of approach received the name of unsupervised classification [14].

The most common approaches for cluster analysis in biomedical applications are hierarchical and partitional clustering [15, 16, 17]. However, both approaches could lead to different clustering results depending on the biomedical application [18]. Hierarchical clustering produces nested sets of clusters by grouping objects over a variety of scales with a nested sequence of partitions following a hierarchical tree structure or dendrogram (a two-dimensional diagram). In the agglomerative hierarchical clustering algorithm, the two closest clusters on a level are merged to form a new larger single cluster on the next level. Partitional clustering, like the k-means clustering algorithm, aggregates objects into a prespecified number of mutually exclusive clusters without any hierarchical taxonomy. k-means algorithm partitions the data into k disjoint clusters so that the within-cluster sum of squares is minimized by iteratively moving points from one cluster to another.

The cluster analysis carried out in this work uses plasma glucose and insulin concentrations as attributes to hierarchical and partitional clustering algorithms to find groups in the data. A first cluster analysis uses the four variables available from the 2-sample 2-hour OGTT, and a second cluster analysis uses the ten variables available from the 5-sample 2-hour OGTT. The goal was to explore the similarity of the groups obtained using four and ten variables and to investigate whether the resulted clusters are population specific.

Material and Method

Glucose and Insulin Data

Subjects were selected in the study if they met with the following criteria: male, ages ranging from 18 to 45 years, non-smoker, good overall health (no evidence of physical disabilities or cardiovascular disease), not taking any medications, and professional marathon runner with a weekly training of 180-240 km or people diagnosed with the metabolic syndrome according to the National Cholesterol Education Program's Adult Treatment Panel III [19]. Exclusion criteria were: unable to meet the

inclusion criteria, current alcohol or drug abuse, and diabetes. In consequence, the sample data was divided into two groups: people with the metabolic syndrome and professional marathon runners.

In the morning, after a night of fasting, the 5-sample 2-hour OGTT was performed on the subjects. The levels of plasma glucose and insulin concentrations were measured before (G_0 and I_0 at 0 min) and after (G_{30} and I_{30} at 30 min, G_{60} and I_{60} at 60 min, G_{90} and I_{90} at 90 min, and G_{120} and I_{120} at 120 min) the oral intake of 75 grams of liquid glucose.

The clinical protocol was carried out by a physician between 2009 and 2013 at the Laboratory of Clinical Investigations, Caracas University Hospital, Venezuela [20, 21]. All procedures performed in the study were in accordance with the ethical standards of the Caracas University Hospital, and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

Each attribute of the dataset was standardized using the standard score (z-score). This standardization allows preserving the range by rescaling all the data to the same range. Median and interquartile range (IQR, the difference between the 75th and the 25th percentiles of the data) values of the plasma glucose and insulin concentrations were reported. The two independent groups were compared with the Mann-Whitney U-test.

Data Clustering

In this paper, hierarchical and partitional clustering algorithms were carried out using the attributes at 0 min and 120 minutes (G_0, G_{120}, I_0, and I_{120}), and using all the attributes from the five-time instant (G_0,…, G_{120}, I_0,…, I_{120}). The agglomerative hierarchical clustering was chosen as the hierarchical clustering technique whereas the k-means clustering algorithm was chosen as the partitional clustering one.

Euclidean (ℓ2 norm), Manhattan (ℓ1 norm), Chebyshev (ℓ∞ norm), Mahalanobis and Cosine distance metrics were used in the agglomerative hierarchical clustering with average, complete and single linkage methods for all distance metrics and, also, with centroid and Ward linkage methods for the Euclidean distance. On the other hand, in the k-means clustering approach, the Squared Euclidean, Manhattan and Cosine distance measures were used.

Evaluation of Clustering

In the k-means clustering algorithm, k=2 clusters were chosen in advance to perform the cluster assignment given that the dataset consists of two groups (people with metabolic syndrome and marathon runners). Similarly, two clusters were selected from the dendrogram plot of the hierarchical cluster tree as the two last clusters formed just before these clusters merged into the single cluster at the top of the tree.

Even if the class of subjects from the database was labeled in advance, this information was not used during the clustering task. The goal was to test whether clusters have relationships with these profiles when two clusters are used to group the data. In this sense, for each clustering experiment, a confusion matrix was used to visualize the results and compare clusters from groups of the database. The accuracy of the group assignment was computed as the sum of the diagonal of the confusion matrix divided by the total population and expressed in percentage. In addition, to check the assignment of the data to the cluster, the silhouette coefficient (SC) was used in both clustering approaches, where a higher SC relates to a better cluster assignment [22]. The 95% confidence interval (CI) boundaries for the mean SC value were also computed.

Results

Fifteen people aged 31.40±6.97 years diagnosed with the metabolic syndrome according to the National Cholesterol Education Program's Adult Treatment Panel III and 15 professional marathon runners aged 33±8.21 years were included in the analysis.

The summary of the plasma glucose and insulin concentrations are shown in Tables 1 and 2, respectively.

Table 1. Median (IQR) of the glucose concentration during the OGTT. Plasma glucose levels are given in mmol/mol. p-values were obtained from the Mann-Whitney U-test. IQR stands for interquartile range, G_0, G_{30}, G_{60}, G_{90}, and G_{120} correspond to the levels of plasma glucose concentration at 0, 30, 60, 90 and 120 minutes of the OGTT, respectively.

	G_0	G_{30}	G_{60}	G_{90}	G_{120}
Metabolic syndrome	103 (10.25)	161 (37)	165 (51.5)	146 (40.5)	131 (30)
Marathon runners	86 (11)	112 (39.75)	90 (34.75)	84 (34.25)	71 (27.25)
p-value	6×10^{-6}	0.0005	4.8×10^{-5}	1.9×10^{-5}	4.1×10^{-6}

Table 2. Median (IQR) of the insulin concentration during the OGTT. Plasma insulin levels are given in μIU/mL. p-values were obtained from the Mann-Whitney U-test. IQR stands for interquartile range, I_0, I_{30}, I_{60}, I_{90}, and I_{120} correspond to the levels of insulin concentration at 0, 30, 60, 90 and 120 minutes of the OGTT, respectively.

	I_0	I_{30}	I_{60}	I_{90}	I_{120}
Metabolic syndrome	11 (6.5)	70 (85.5)	70 (143)	75 (91.75)	94 (93)
Marathon runners	2 (1.41)	28.79 (12.5)	24 (14.93)	21.4 (17.97)	16.6 (12.11)
p-value	7.9×10^{-5}	0.0011	8.8×10^{-5}	2.3×10^{-5}	3.4×10^{-5}

Examples of dendrogram plots are shown in Figure 1. These plots represent the agglomerative hierarchical clustering results for four (G_0, G_{120}, I_0, and I_{120}) and ten (G_0, …, G_{120}, I_0,…, I_{120}) attributes, and were obtained using the Cosine distance metric and Average linkage method. Only one subject (subject #2 in the metabolic syndrome group) was misclassified in both cases.

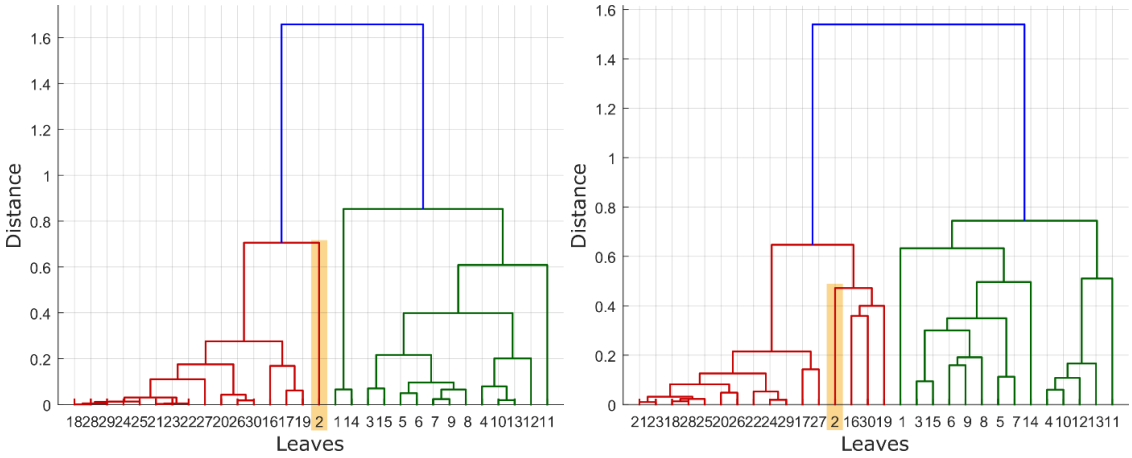

Figure 1. Dendrogram plots using four (left-hand graph) and ten (right-hand graph) attributes for the agglomerative hierarchical clustering using Cosine distance metric and Average linkage method. Subject #2 (highlighted in both plots) from the metabolic syndrome group was incorrectly clustered in both cases.

Examples of silhouette plots using ten attributes are shown in Figure 2 for k-means clustering with Manhattan distance metric and agglomerative hierarchical clustering with Cosine distance metric and Average linkage method. Ten subjects in the metabolic syndrome group showed silhouette coefficients lower than 0.5, and among them, two subjects (#1 and #3) showed negative silhouette coefficients, and one subject (#2) was incorrectly classified when agglomerative hierarchical clustering method was used. Ten subjects in the metabolic syndrome group showed silhouette coefficients lower than 0.5, and among them, three subjects (#1, #2 and #3) showed negative silhouette coefficients when k-means clustering method was applied. In this case, contrary to agglomerative hierarchical clustering, k-means clustering correctly classified subject #2. In addition, the mean value SC was 0.5822 both for k-means and for agglomerative hierarchical clustering and

the accuracy was 100% for k-means and 96.67% for the agglomerative hierarchical clustering. Moreover, silhouette coefficients were statistically significantly higher in the marathon runner group than in the metabolic syndrome group (0.8760 vs. 0.3790, $p=4.14\times10^{-6}$, using k-means; 0.8841 vs. 0.3581, $p=4.14\times10^{-6}$, using agglomerative hierarchical clustering), and silhouette coefficients of the marathon runner group were higher, but not statistically significant, using agglomerative hierarchical clustering than using k-means (0.8841 vs. 0.8760, p=0.8357).

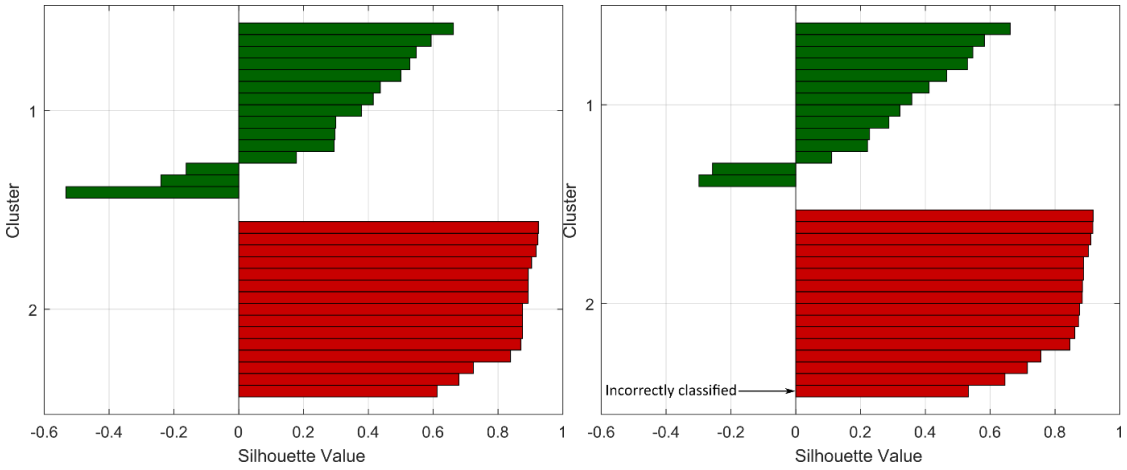

Figure 2. Silhouette plots using ten attributes for the k-means clustering algorithm using Manhattan distance measure (a) and the agglomerative hierarchical clustering using Cosine distance metric and Average linkage methods (b). The metabolic syndrome group is represented with green bars whereas the marathoner group is shown in red.

Clustering performances using k-means and agglomerative hierarchical clustering algorithms, using four and ten attributes, for different configuration of the algorithms, are shown in Tables 3 and 4. From these tables, we can see that k-means clustering correctly groups subjects using the Manhattan distance measure with ten attributes (all levels of glucose and insulin from the OGTT) although this result was not necessarily associated with the highest mean silhouette coefficient. The highest mean silhouette coefficient (0.6991) was obtained using k-means with four attributes using both Squared Euclidean and Manhattan distance measures. The worst clustering performances (53.33%) were achieved using agglomerative hierarchical clustering with four attributes, and the lowest mean silhouette coefficient values (0.3877) were obtained using agglomerative hierarchical clustering with ten attributes using Single linkage methods. Generally, for different configurations of the clustering algorithms, k-means performed better than agglomerative hierarchical clustering.

Table 3. Clustering performance using the k-means algorithm for different distance measures, for four and ten attributes. Accuracy values greater than 90% and mean SC values greater than 0.6 are highlighted in bold. SC stands for silhouette coefficient and CI stands for the 95% confidence interval.

Distance	Number of attributes	Accuracy (%)	Mean SC (CI)
Squared Euclidean	4	**93.33**	**0.6991 (0.6142 to 0.7839)**
	10	**93.33**	0.5881 (0.4714 to 0.7048)
Manhattan	4	**93.33**	**0.6991 (0.6142 to 0.7839)**
	10	**100.00**	0.5632 (0.4219 to 0.7046)
Cosine	4	**96.67**	**0.6873(0.5820 to 0.7926)**
	10	**96.67**	0.5822 (0.4545 to 0.7099)

Table 4. Clustering performance using the agglomerative hierarchical clustering algorithm for different distance metrics and linkage methods, for four and ten attributes. Accuracy values greater than 90% and mean SC values greater than 0.6 are highlighted in bold. SC stands for silhouette coefficient and CI stands for the 95% confidence interval.

Distance	Linkage	Number of attributes	Accuracy (%)	Mean SC (CI)
Euclidean	Average	4	53.33	0.5438 (0.4209 to 0.6667)
		10	66.67	**0.6670 (0.5886 to 0.7454)**
	Complete	4	90.00	**0.6749 (0.5845 to 0.7654)**
		10	66.67	**0.6670 (0.5886 to 0.7454)**
	Single	4	53.33	0.5438 (0.4209 to 0.6667)
		10	53.33	0.3877 (0.2604 to 0.5151)
	Centroid	4	53.33	0.5438 (0.4209 to 0.6667)
		10	56.67	**0.6486 (0.5084 to 0.7887)**
	Ward	4	**96.67**	**0.6873 (0.5820 to 0.7926)**
		10	90.00	0.4803 (0.2976 to 0.6630)
Manhattan	Average	4	90.00	**0.6749 (0.5845 to 0.7654)**
		10	63.33	**0.6624 (0.5738 to 0.7511)**
	Complete	4	90.00	**0.6749 (0.5845 to 0.7654)**
		10	66.67	**0.6670 (0.5886 to 0.7454)**
	Single	4	53.33	0.5438 (0.4209 to 0.6667)
		10	53.33	0.3877 (0.2604 to 0.5151)
Cosine	Average	4	**96.67**	**0.6873 (0.5820 to 0.7926)**
		10	**96.67**	0.5822 (0.4545 to 0.7099)
	Complete	4	**96.67**	**0.6873 (0.5820 to 0.7926)**
		10	**96.67**	0.5822 (0.4545 to 0.7099)
	Single	4	**96.67**	**0.6873 (0.5820 to 0.7926)**
		10	**96.67**	0.5822 (0.4545 to 0.7099)
Chebyshev	Average	4	60.00	0.5086 (0.3717 to 0.6456)
		10	60.00	**0.6297 (0.5095 to 0.7498)**
	Complete	4	60.00	0.5086 (0.3717 to 0.6456)
		10	60.00	**0.6297 (0.5095 to 0.7498)**
	Single	4	53.33	0.5438 (0.4209 to 0.6667)
		10	53.33	0.3877 (0.2604 to 0.5151)
Mahalanobis	Average	4	53.33	0.5438 (0.4209 to 0.6667)
		10	53.33	0.3877 (0.2604 to 0.5151)
	Complete	4	53.33	0.5438 (0.4209 to 0.6667)
		10	60.00	**0.6297 (0.5095 to 0.7498)**
	Single	4	53.33	0.5438 (0.4209 to 0.6667)
		10	53.33	0.3877 (0.2604 to 0.5151)

Discussion

Dendrogram plots (see Figure 1) have helped to visualize how data are grouped (merged) throughout the agglomerative hierarchical clustering process. In contrast, given the high dimensionality of the data (4 or 10 attributes), the grouping of the data using k-means clustering was not possible to visualize. On the other hand, silhouette coefficients (see Figure 2) have helped to visualize tightness and separation of objects within clusters and the quality of clustering achieved, where better mean silhouette coefficients were obtained for k-means. Professional marathon runners showed more consistent and homogenous levels of glucose and insulin (see Tables 1 and 2), and thus higher values of silhouette coefficients, than subjects with the metabolic syndrome, as can be seen in Figure 2. This could be the result of more efficient and stable glucose metabolism in marathon runners thanks to an improvement of the insulin-glucagon negative feedback mechanism as a consequence of exercise training.

The use of 10 attributes instead of 4 has provided the clustering algorithms with more information when clustering the data and lead to better classification results, mainly using k-means, as can be observed in Tables 3 and 4. The use of different distance measures has resulted in similar clustering for k-means, but different clustering results were obtained with the use of different distance metrics and linkage methods for the agglomerative hierarchical clustering, where Chebyshev and Mahalanobis distance metrics and single linkage method lead to the worst classification results.

Mean silhouette coefficients per clustering experiment were not very high (ranging from 0.3877 to 0.6991) and were not associated with classification accuracy results (ranging from 53.33% to 100%). The principal reason of that was that, compared with marathon runners, silhouette coefficients for subjects with metabolic syndrome were lower given the high variability among attributes as consequences of metabolic disorders; indeed, three subjects in the metabolic syndrome group were difficult to classify.

The significance of this study lies in the analysis of different machine-learning-based clustering approaches for grouping subjects with different metabolic conditions on the basis of the levels of plasma glucose and insulin concentrations obtained at different time instants. Nevertheless, there were a number of practical difficulties and limitations that should be highlighted. One challenge was the recruitment of subjects since financial constraints and the willingness of subjects to undertake the OGTT test led to a sample size of fifteen subjects per group. The use of a clustering algorithm to cluster labeled data can be seen as another limitation of this study; however, the label of the data was not used during the classification task and was only used to estimate the accuracy of the created clusters. Using a larger database with a greater number of attributes, a deeper exploratory analysis and clustering methodology could be carried out to unveil unknown subgroups in the data, even unrelated to metabolic profiles; our research team is currently working on this direction.

Conclusions

Using insulin and glucose data from the OGTT, k-means and agglomerative hierarchical clustering strategies with different configuration have grouped individuals according to their metabolic profile: subjects with metabolic syndrome and professional marathon runner. However, only k-means with Manhattan distance and ten attributes provided a perfect grouping of the data (100% accuracy). According to the experiments carried out, k-means performed better than agglomerative hierarchical clustering and using the data from five blood samples (spaced in time by 30 minutes) instead of using the data from two blood samples (spaced in time by two hours) lead to better classification performances.

While recognizing the study limitations, the results of this work showed that using a clustering machine learning approach, it was possible to automatically differentiate two groups of people based on metabolic variables, namely, plasma glucose and insulin concentrations, taken at different time instants. In this sense, through a larger database and by including other populations in the study, a future work would be the determination of cut-off values of the levels of glucose and insulin concentrations to characterize different populations. This would help in the early diagnosis of metabolic diseases and in the reduction of the cost of healthcare and medical treatment.

List of abbreviations

OGTT – Oral glucose tolerance test.
HDL cholesterol – High-density lipoprotein cholesterol.
IQR – Interquartile range.
G_0 – Level of plasma glucose concentration at 0 min of the OGTT.
G_{30} – Level of plasma glucose concentration at 30 min of the OGTT.
G_{60} – Level of plasma glucose concentration at 60 min of the OGTT.
G_{90} – Level of plasma glucose concentration at 90 min of the OGTT.
G_{120} – Level of plasma glucose concentration at 120 min of the OGTT.

I_0 – Level of plasma insulin concentration at 0 min of the OGTT.
I_{30} – Level of plasma insulin concentration at 30 min of the OGTT.
I_{60} – Level of plasma insulin concentration at 60 min of the OGTT.
I_{90} – Level of plasma insulin concentration at 90 min of the OGTT.
I_{120} – Level of plasma insulin concentration at 120 min of the OGTT.
SC – Silhouette coefficient.
CI – Confidence interval.

Conflict of Interest

The author declares that he has no conflict of interests.

References

1. Widmaier EP, Raff H, Strang KT. Vander's human physiology, 14th ed. McGraw-Hill Education; 2015.
2. Gylfe E. Glucose control of glucagon secretion – 'There's a brand-new gimmick every year'. Upsala journal of medical sciences. 2016;121(2):120-32.
3. Rooijackers HM, Wiegers EC, Tack CJ, van der Graaf M, de Galan BE. Brain glucose metabolism during hypoglycemia in type 1 diabetes: insights from functional and metabolic neuroimaging studies. Cellular and Molecular Life Sciences. 2016;73(4):705-22.
4. Languren G, Montiel T, Julio-Amilpas A, Massieu L. Neuronal damage and cognitive impairment associated with hypoglycemia: an integrated view. Neurochemistry International. 2013;63(4):331-43.
5. Kaul K, Hodgkinson A, M Tarr J, M Kohner E, Chibber R. Is inflammation a common retinal-renal-nerve pathogenic link in diabetes?. Current Diabetes Reviews. 2010;6(5):294-303.
6. Bornfeldt KE, Tabas I. Insulin resistance, hyperglycemia, and atherosclerosis. Cell Metabolism. 2011;14(5):575-85.
7. American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care. 2014;37(Supplement 1):S81-90.
8. Xun P, Liu K, Cao W, Sidney S, Williams OD, He K. Fasting insulin level is positively associated with incidence of hypertension among American young adults: a 20-year follow-up study. Diabetes Care. 2012;35(7):1532-7.
9. Ruderman NB, Carling D, Prentki M, Cacicedo JM. AMPK, insulin resistance, and the metabolic syndrome. The Journal of Clinical Investigation. 2013;123(7):2764-72.
10. Roberts CK, Hevener AL, Barnard RJ. Metabolic syndrome and insulin resistance: underlying causes and modification by exercise training. Comprehensive Physiology. 2013;3(1):1-58.
11. Altuve M, Perpiñan G, Severeyn E, Wong S. Comparing glucose and insulin data from the two-hour oral glucose tolerance test in metabolic syndrome subjects and marathon runners. 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2016, pp. 5290-5293.
12. Slonim DK. From patterns to pathways: gene expression data analysis comes of age. Nature Genetics. 2002;32:502.
13. Bandyopadhyay S, Saha S. Unsupervised classification: similarity measures, classical and metaheuristic approaches, and applications. Springer Science & Business Media; 2012.
14. Saxena A, Prasad M, Gupta A, Bharill N, Patel OP, Tiwari A, Er MJ, Ding W, Lin CT. A review of clustering techniques and developments. Neurocomputing. 2017;267:664-81.
15. Bruse JL, Zuluaga MA, Khushnood A, McLeod K, Ntsinjana HN, Hsia TY, et al. Detecting clinically meaningful shape clusters in medical image data: metrics analysis for hierarchical clustering applied to healthy and pathological aortic arches. IEEE Transactions on Biomedical Engineering. 2017;64(10):2373-83.

16. Rosati S, Agostini V, Knaflitz M, Balestra G. Muscle activation patterns during gait: A hierarchical clustering analysis. Biomedical Signal Processing and Control. 2017;31:463-9.
17. Salem SB, Naouali S, Chtourou Z. A fast and effective partitional clustering algorithm for large categorical datasets using a k-means based approach. Computers & Electrical Engineering. 2018;31;68:463-83.
18. Xu R, Wunsch DC. Clustering algorithms in biomedical research: a review. IEEE Reviews in Biomedical Engineering. 2010;3:120-54.
19. Grundy SM, Brewer Jr HB, Cleeman JI, Smith Jr SC, Lenfant C. Definition of metabolic syndrome. Arteriosclerosis, Thrombosis, and Vascular Biology. 2004;24(2):e13-e18.
20. Severeyn E, Wong S, Passariello G, Cevallos JL, Almeida D. Methodology for the study of metabolic syndrome by heart rate variability and insulin sensitivity. Revista Brasileira de Engenharia Biomédica 2012;28(3):272-77.
21. Ledezma CA, Severeyn E, Perpiñan G, Altuve M, Wong S. A new on-line electrocardiographic records database and computer routines for data analysis. 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2014, pp. 2738-2741.
22. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987;20:53-65.