

Clinical values dataset processing through cluster analysis to find cardiovascular risk

C M Bucci^{1,2}, W E Legnani^{1,2}, R L Armentano^{1,2,3}

¹ Biologic-Engineering Research and Development Group, Electronic Department, Buenos Aires Regional Faculty, National Technological University, Medrano 951 (C1179AAQ), Buenos Aires, Argentina.

² Non-Linear Systems and Modelling Group, Centre for Signal and Image Processing, National Technological University, Medrano 951 (C1179AAQ), Buenos Aires, Argentina.

³ Electronic Department, Favaloro University, Solis 453 (C1078AAI), Buenos Aires, Argentina.

E-mail: cbucci@rec.utn.edu.ar

E-mail: wlegnani@rec.utn.edu.ar

E-mail: armen@ieee.org

Abstract. The scope of this work is to show another way to grouping population with clinical variables measured in health centres and to assign a cardiovascular risk indicator. To do this, two different datasets were used, one coming from France and another coming from Uruguay. The well proved Framingham index was used to validate the results. The preliminary results are very auspicious to encourage the research and get deeper knowledge of the cardiovascular risk indicators.

1. Introduction

The cardiovascular system plays a role in maintaining homeostasis - that is, a stable environment - inside the body. This system can perform, or indicate to other systems that carry out short-term rapid adjustments in response to the demands requested by the body due to the development of various activities and changing external conditions. For example, when the blood supply increases, the flow must be reduced to other organs, or heart rate should be increased. Throughout these adaptations, blood pressure must remain constant to maintain vital functions of all body tissues. To perform these adjustments, the cardiovascular system communicates with other organs through a complex network of mechanisms for monitoring and alarms. This sends signals about their condition and, in turn, receives messages that control its operation (Zaret et al., 1992).

Consequently, it is very interesting the development of models and / or signal processing tools for predicting cardiovascular risk, both in order to understand the possible mechanisms that increase it, to be able to early intervention through prevention campaigns or perform therapeutic treatments.



2. Cardiovascular System Description

2.1. State of art

The cardiovascular system is divided into the systemic circulation provided by the left ventricle and the pulmonary circulation provided by the right ventricle. Each movement can be conveniently divided into four parts: the heart pump, arteries, veins and microcirculation. While each part separately analyzes always remember that it is a closed system with each of the parties interacting in varying degrees with each of the other system. For example, the same amount of blood must flow through the left and right side of the heart. There may be temporary changes that lead to the redistribution of blood in the circulatory system, but these differences cannot be sustained for long and in that sense we must quickly restore homeostasis. The heart is a unique organ that is divided into the left and right sides. The arteries are the major blood vessels that carry blood from the heart to the microcirculation of the tissue which must be balanced and then the veins carry blood back to the heart. In mechanical terms the great arteries and veins are different from other vessels because of its size and inertial characteristics rather than viscous effects in blood flow between them (Formaggia et al., 2009).

2.2. Cardiovascular Risk

Cardiovascular diseases come up suddenly affecting the heart (heart attack) or brain (stroke) at a given time.

Everyday life coupled with poor eating habits and sedentary lifestyle leads these diseases. Its effects are manifested in the arteries of the heart, brain, kidneys and lower limbs.

The main factors that influence these diseases are cholesterol, hypertension, smoking, diabetes, stress, physical inactivity, obesity, heredity and heart rate. It should also be noted that the existence of several factors increases the risk of cardiovascular disease. Most of them are very serious and can leave sequels or even lead to death. Often it generates irreversible consequences.

Therefore early detection of patterns to determine a potential risk is vital to diagnose a preventive treatment.

2.3. Framingham Method

The *Framingham Heart Disease Epidemiology Study* is designed to measure certain selected constitutional factors and certain of the conditioning factors in a large number of "normal" persons selected at random and to record the time during which these selected factors act and interact before clinical cardiovascular disease result (Gordon and Kannel, 1968).

3. Cluster Analysis Modelling Based On Real Datasets

3.1. Cluster Analysis

The Cluster Analysis (CA) is related to the identification of groups of data sets. The main objective of these techniques is the division or partition of a set of observations into a number of subgroups or clusters which is called "cluster" so that all observations within a subgroup are similar to each other and differentiate into the fullest extent of the observations belonging to other clusters, while observations from different groups need not be similar to each other (Timm, 2002). Consequently, the objective of this methodology is the identification of hidden structures that are hidden in the data sets (Everitt, et al., 2011) and (Fielding, 2007). The non-hierarchical clustering procedures usually follow the following steps (Everitt, 1993), (Decker and Lenz, 2007) and (Abonyi and Feil, 2007).

1. Select the k (n -dimensional) centroids of the clusters or seeds.
2. Allocate each observation to the nearest centroid using some measure of distance, (Usually Euclidean).
3. Relocate each observation to one of the k clusters based on a preset criterion.
4. To close the process if you cannot perform any relocation of the observations in any cluster or reallocation satisfies the convergence criterion given, otherwise return to step 2.

The non-hierarchical clustering method begins by selecting k centroids or seeds. These k seeds can be the first k observations that may be taken at some level separation is otherwise identify them at random, other algorithms begin with random seed and then through some very well defined algorithm relocated (Timm, 2002). Once the seeds are selected, each of the observations is evaluated to assign it to one of the clusters defined by them.

In step 2 the seed may or may not be updated. In this way you can use two tests to determine whether to perform the replacement. An observation can replace one or a few seeds if the distance between seeds is less than the distance between an observation and the next nearest seed. The displaced seed becomes an observation for recalculation of the centroids. If an observation fails this test may pose a second instance of evaluation (Timm, 2002).

The observation replaces the nearest seed if Euclidean distance there from all next seed is greater than the shortest distance from the nearest seed remaining seeds.

In the first instance all observations are associated with k clusters. Then this process is repeated until the changes in the seeds of small clusters are based on some predetermined convergence criterion.

3.2. Description of Real Datasets

In this paper were used two distinct datasets coming from different countries, one of them from France (dataset A) and another from Uruguay (dataset B). Of course people included in dataset A has different lifestyle, feeding and genetic characteristics to people included in dataset B, the use of these different sources of information was used in order to compare results of classifying cardiovascular risk. The dataset A is compound by a matrix with 618 patients (rows) and 10 measured clinic variables (columns) and the dataset B consist of 123 patients (rows) and 11 measured clinic variables (columns).

3.3. Modelling Validation

One of the more extended methods for validating the number of clusters for aggregating the values which comes from the dataset during a non-hierarchical clusterization process is to compute the silhouette index or silhouette coefficient.

This index is used with a plot that means of assessing the quality of a cluster solution, enabling the investigator to identify “poorly” classified objects and so distinguishing clear-cut clusters from weak ones. Silhouette plots for cluster solutions obtained from different choices for the number of groups can be compared, and the number of groups chosen so that the quality of the cluster solution is maximized (Everitt, et al., 2011).

To find detailed information about the silhouette concept the reader can see (Fielding, 2007, Everitt, et al., 2011, Kogan, 2007 and Kaufman and Rousseeuw, 1990).

3.4. Framingham Index

The aim of the *Framingham logistic model*, deals with on common factors identification or characteristic that contribute to cardiovascular disease, through monitoring of development

(evolution), over a period of time which a large group of participants has not yet developed symptoms of cardiovascular illness, or suffered heart attacks or strokes (Armentano 2011).

The variables involved are *Sex* (dichotomous variable), *Age* in years, *Serum Cholesterol LDL* (mg / dl), the *fraction of cholesterol linked to high density lipoprotein HDL*, *Systolic Pressure*, *Diastolic Pressure*, the presence of *Diabetes* (dichotomous variable: Yes, No) and *Smoking* (dichotomous variable: Yes, No) (Armentano 2011):

4. Design and Implementation

4.1. Patients Dataset A

The dataset A is coming from different patients taking at the Pole Cardiovasculaire Hopital European Georges Pompidou, Paris, France, at which has measured the clinical variables described in the Table 1. The size matrix size was described previously.

Table 1. Original dataset matrix 618 rows x 10 columns.

Variable	Age ^a	Sex ^b	SBP ^c	DBP ^d	LDL ^e	HDL ^f	Smoke ^g	Diab. ^h	Fram. ⁱ	SCa ^j
Patient #1	40	1	116	70	93	46	1	0	0,029596604	0
Patient #2	49	1	156	94	155	25	0	0	0,145509165	0
Patient #3	44	1	132	88	200	51	1	0	0,156247042	0
.....
Patient #618	46	1	136	92	200	30	0	0	0,212639732	1

^a Age in years, ^b Sex (dichotomous variable), ^c SBP Systolic Pressure, ^d DBP Diastolic Pressure, ^e LDL Serum Cholesterol, ^f HDL the fraction of cholesterol linked to high density lipoprotein, ^g Smoke (dichotomous variable: Yes, No), ^h Diab. the presence of Diabetes (dichotomous variable: Yes, No), ⁱ Fram. calculated value of Framingham Index, ^j SCa measured value of Score Calcic

The first step is the matrix preparation for the experiment in which case the binary data such as “yes” or “no” in case of measures variables of “smoke”, “diabetic” and “sex” was eliminated.

In order to prepare the data matrix to the clusterization algorithm is to apart the column containing the Score Calcic index to be used as comparison variables of the results. With these considerations a 618 rows by 5columns data matrix was used during the clustering process as is showed in Table 2.

Table 2. Study dataset matrix 618 x 5

Variable	Age	SBP	DBP	LDL	HDL
Patient #1	40	116	70	93	46
Patient #2	49	156	94	155	25
Patient #3	44	132	88	200	51
.....
Patient #618	46	136	92	200	30

Once obtained the final matrix it is proceeded to get the correct number of clusters that the system needs to make a correct separation of population. To carry out with this job the silhouette coefficient explained before was applied, and the results are showed in Table 3.

Table 3. Silhouette coefficient.

Cluster #	Silh. Coeff.
2 Clusters	0.4946
3 Clusters	0.4515
4 Clusters	0.4180
5 Clusters	0.3899

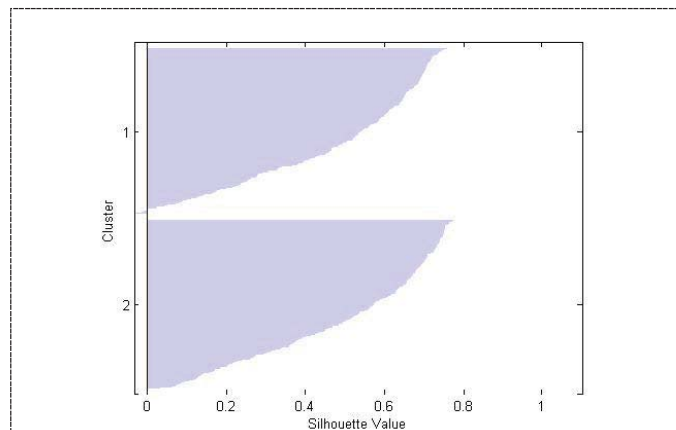


Figure 1. Silhouette coefficient for 2 clusters, for dataset A

As can see in Table 3, the correct numbers of clusters are 2 because it is the highest indicator of the four coefficients. As much as it sticks to 1 more reliable will be the result selected (J Kogan, 2007). It is possible to show a graphical representation (Everitt, et al., 2011) of this result as can see in Figure 1.

The subdivision of the dataset A into two clusters was performed by the technique named k-means cluster analysis.

After running the clustering process it is obtained 2 groups of patients grouped in cluster #1 and cluster #2. A quick verification of the correct grouping is showed making a scatter plot representation of the results (Figure 2).

The result shows that the best cut is represented by the LDL and SBP due to both groups of patients are well separated without mixing patients of each cluster. The rest of scatter plots with different variables do not generate two well separated clusters.

Table 4. Arithmetic mean for each cluster

Cluster #	Age	SBP	DBP	LDL	HDL
Cluster #1	48	141	90	159	48
Cluster #2	47	133	85	216	47

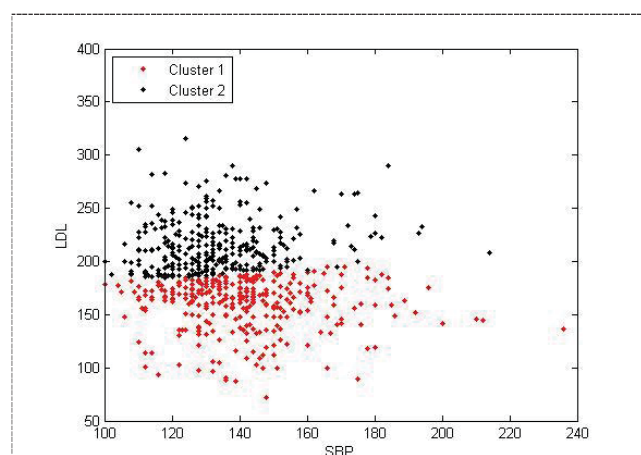


Figure 2. Matrix 618 x 5 (scatter plot LDL and SBP), for dataset A

Now it is easy to determinate graphically which patient in which clusters are by grouping patients in cluster #1 in red colour on the bottom area and cluster #2 in black colour on top area. Considering a boundary cutting line value of 190.50 on the LDL axis is possible to define that 73 % of patients in cluster #2 (246 patients) are located above this cutting and 27 % (66 patients) are below it. It is considered that 100 % (372 patients) of cluster #1 are also below of the cutting line.

The calculated “arithmetic mean” for each variable included in cluster #1 (306 patients) and #2 (312 patients) is showed in Table 4

It shows that the “arithmetic mean” factors are in general in the same level, the biggest difference between cluster #1 and #2 is in the variable LDL.

An additional validation will be done by implementing of Welch's T-Test. This procedure is used only when it can be assumed that the two population variances are different (the sample sizes may or may not be the same) and therefore must be estimated separately (Welch, 1947).

The output of the test implemented has two possible results: 1 or 0. If 1, the population A and B are both different means which imply that A and B are well separated, then it is possible for instance to use “Age” in order to separate the population in two groups, or could be interpreted as the age is a factor to classify cardiovascular risk. Otherwise, when the result is 0 two groups aren't well separated, and the means are equal at a certain statistical level of significance.

The Table 5 shows the result T-Test with 2 different quantity of population for comparing cluster #2 (312 patients) and cluster #1 (306 patients). In this case the population could be well separated implementing Age, SBP, DBP, LDL, Fram.

Table 5. T-Test result 618 x 5 (Adding Fram. comparison)

Variable	Age	SBP	DBP	LDL	HDL	Fram.
T-Test	1	1	1	1	0	1

Now it will be necessary to insert the measured SCa column and then repeating again the same procedure as follows:

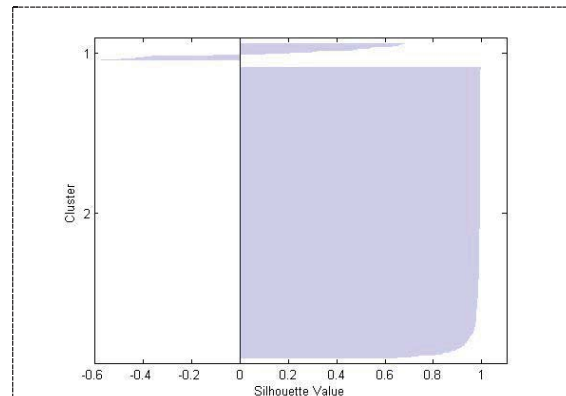
Table 6. Study dataset matrix 618 x 6

Variable	Age	SBP	DBP	LDL	HDL	SCa
Patient #1	40	116	70	93	46	0
Patient #2	49	156	94	155	25	0
Patient #3	44	132	88	200	51	0
.....
Patient #618	46	136	92	200	30	1

The output of the silhouette coefficient is showed in Table 7 and Figure 3 where the highest indicator of silhouette coefficient is represented for the model of 2 clusters:

Table 7. Silhouette coefficient

Cluster #	Silh. Coeff.
2 Clusters	0.9283
3 Clusters	0.8302
4 Clusters	0.3700
5 Clusters	0.4252

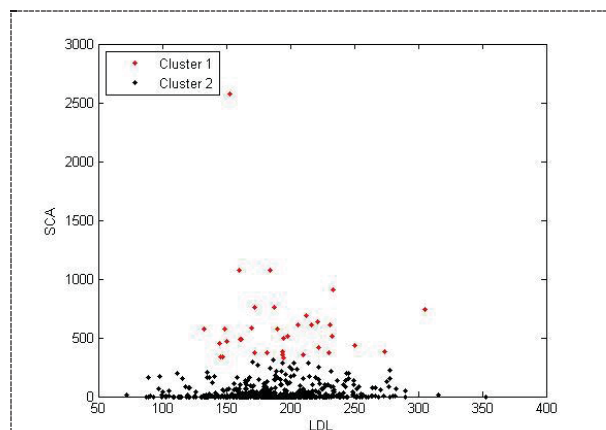
**Figure 3.** Silhouette coefficient for 2 clusters.

Two groups of patients will be determined for cluster #1 and cluster #2 will be determined by using k-means. Scatter plot will show graphically the best cut for instance of variables LDL and SCa.

Assuming a boundary cutting line value of 328 on the SCa axis (LDL 195.075) is possible to define two groups of patients well separated in cluster #1 and #2 following Figure 4. A total of 35 patients grouped 100 % in cluster #1 are located above this cutting line with the higher values of SCa. The rest of 583 patients are 100 % all grouped in cluster #2.

Table 8. Arithmetic mean for each cluster

Cluster #	Age	SBP	DBP	LDL	HDL	SCa
Cluster #1	56	146	89	193	48	606
Cluster #2	47	136	87	187	48	29

**Figure 4.** Matrix 618 x 6 (scatter plot LDL and SCa).

The calculated “Arithmetic mean” for each variable included in cluster #1 and #2 is showed in Table 8

It shows that the “arithmetic mean” factors are in general in the same level, the biggest difference between cluster #1 and #2 is in the variable SCa.

The Table 9 shows the result T-Test with 2 different quantity of population for comparing cluster #1 (35 patients) and cluster #2 (583 patients). In this case the population could be well separated implementing Age, SBP, SCa, Fram.

Table 9. T-Test result 618 x 6 (Adding Fram. comparison)

Variable	Age	SBP	DBP	LDL	HDL	SCa	Fram.
T-Test	1	1	0	0	0	1	1

4.2. Patients Dataset B

The experiment for dataset B will be the same than dataset A by means of repeating all the same processes. The only difference is that patients considered come from Uruguay and the clinical variables measured are a little bit different.

The dataset B consists of real measured variables matrix in different patients analyzed at Republic University, Uruguay. The matrix size is about 123 patients located at the rows and 11 columns representing the clinic variables according to Table 10.

Table 10. Original dataset matrix 123 rows x 11 columns.

Variable	Age ^a	SBPp ^b	SBPa ^c	DBP ^d	LDL ^e	HDL ^f	BMI ^g	PWV ^h	CIMTr ⁱ	CIMTI ^j	AAP ^k
Patient #1	52	129	121	79	259	47	26.3	16.03	0.957	0.976	13
Patient #2	53	111	106	74	253	47	28.7	9.54	0.600	0.593	10
Patient #3	42	162	156	104	252	36	23.3	8.63	0.572	0.553	32
.....
Patient #123	60	140	128	72	54	70	23.1	11.97	0.864	0.811	15

^a Age in years, ^b SBP Systolic Pressure (Peripheral), ^c SBP Systolic Pressure (Aortic), ^d DBP Diastolic Pressure, ^e LDL Serum Cholesterol, ^f HDL the fraction of cholesterol linked to high density lipoprotein,

^g BMI body mass index, ^h PWV pulse wave velocity, ⁱ CIMTr carotid intima-media thickness (right),

^j CIMTI carotid intima-media thickness (left), ^k AAP Aortic Augmented Pressure

In preparation for the experiment it will be also necessary to eliminate some measured variables such as SBP(A), BMI, PWV, CIMT(R), CIMT(L) and AAP in order to obtain the same 5 variables in the columns, then finally getting Table 11 as a matrix of 123 x 5 variables.

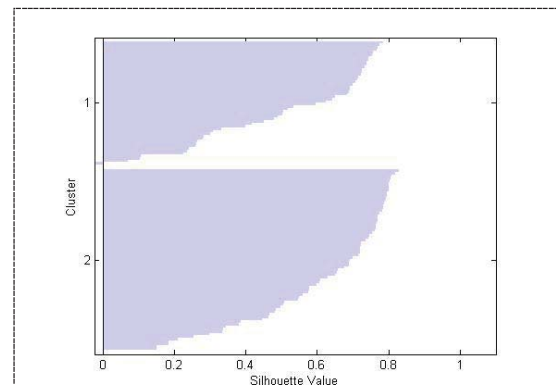
Table 11. Study dataset matrix 618 x 5

Variable	Age	SBP	DBP	LDL	HDL
Patient #1	52	129	79	259	47
Patient #2	53	111	74	253	47
Patient #3	42	162	104	252	36
.....
Patient #123	60	140	72	54	70

The output of the Silhouette coefficient is showed in Table 12 and Figure 5 where the highest indicator of Silhouette coefficient is represented for the model of 2 clusters:

Table 12. Silhouette coefficient

luster #	Silh. Coeff.
2 Clusters	0.5789
3 Clusters	0.4985
4 Clusters	0.4206
5 Clusters	0.4226

**Figure 5.** Silhouette coefficient for 2 clusters.

Two groups of patients will be determined for cluster #1 and cluster #2 by using k-means and with the best cut of scatter plot for instance of variables LDL and SBP.

Following the Figure 6, the boundary cutting line value will be 159 on the LDL axis which defines two groups of patients separated in cluster #1 and #2. In cluster #2 are included a total of 49 patients (98 %) located above this cutting and 1 patient (2 %) below it. The rest of 73 patients grouped 100 % in cluster #1 are located below cutting line.

Table 13. Arithmetic mean for each cluster

Cluster #	Age	SBP	DBP	LDL	HDL
Cluster #1	50	126	79	124	70
Cluster #2	51	122	78	192	53

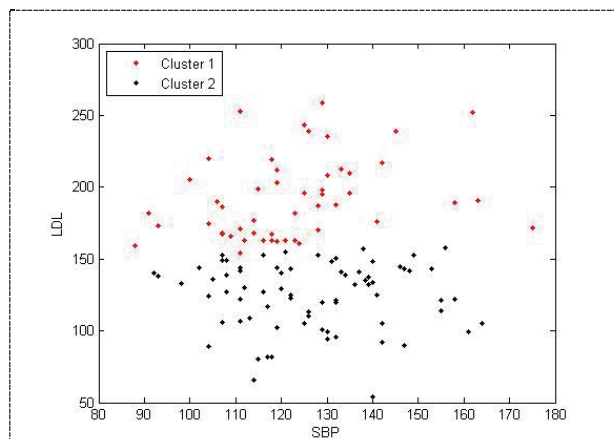
**Figure 6.** Matrix 123 x 5 (scatter plot LDL and SBP).

Table 13 shows that the “arithmetic mean” factors are in general in the same level, the biggest difference between cluster #1 and #2 is in the variable LDL.

The Table 14 shows the result T-Test with 2 different quantity of population for comparing cluster #2 (50 patients) and cluster #1 (73 patients). In this case the population could be well separated implementing LDL, HDL.

Table 14. T-Test result 123 x 5 (Adding PWV comparison)

Variable	Age	SBP	DBP	LDL	HDL	PWV
T-Test	0	0	0	1	1	0

There are no changes (less than 1%) in the results when is introduced PWV in order to implement a running clustering algorithm for new Dataset matrix of 123 patients x 6 variables.

5. Summary Report

In the present work regarding two independent datasets has been clearly exposed that the clustering algorithm can be used to classify standard clinical measurements in order to obtain a cardiovascular risk pattern. In this way the cluster with this proposal appears a cluster containing patients with the low cardiovascular risk another cluster characterized by high cardiovascular risk. To validate the results the Framingham index was applied. The number aggregate (clusters) was justified by means of the silhouette coefficient.

The SCa clinical value has shown a good discriminator between clusters as was seen in the cluster dataset A. In another way if SCa is not available as dataset B the values of LDL was used as proxy discriminator to differentiate the two clusters obtained for classification of clinical cases.

6. References

- [1] The Framingham Study, An Epidemiological Investigation of Cardiovascular Disease – Section 1, T Gordon and W B Kannel, 1968
- [2] Cluster Analysis, 5th Edition, King's College London, UK, Wiley, B S Everitt, S Landau, M Leese, D Stahl, 2011
- [3] Cluster and Classification Techniques for the Biosciences, Cambridge University Press, A Fielding, 2007.
- [4] Finding Groups in Data - An Introduction to Cluster Analysis, Vrije Universiteit Brussel, L Kaufman and P Rousseeuw, 1990.
- [5] Introduction to Clustering Large and High-Dimensional Data, Cambridge University Press, J Kogan, 2007.
- [6] The generalization of "Student's" problem when several different population variances are involved", B L Welch, 1947
- [7] The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test, G D Ruxton, 2006
- [8] Performance of five two-sample location tests for skewed distributions with unequal variances. M W Fagerland, L Sandvik, 2009
- [9] A note on preliminary tests of equality of variances. British Journal of Mathematical and Statistical Psychology, D W Zimmerman, 2004.
- [10] T-Tests, non-parametric tests, and large studies—a paradox of statistical practice. BioMed Central Medical Research Methodology, M W Fagerland, 2012.
- [11] Integrated e-Health Approach Based on Vascular Ultrasound and Pulse Wave Analysis for Asymptomatic Atherosclerosis Detection and Cardiovascular Risk Stratification in the Community, D B Santana, Y A Z'ocalo and R L Armentano, 2011.
- [12] Cardiovascular Mathematics, Modeling and simulation of the circulatory system, Springer-Verlag Italia, L Formaggia, A Quarteroni, A Veneziani, 2009
- [13] Heart Book, University of School Medicine, Yale, B L Zaret, M Moser, L S Cohen 1992