Title: Cardiovascular risk analysis by means of pulse morphology andclustering methodologies

Author: Vânia G. Almeida J. Borba H. Catarina Pereira Tânia Pereira Carlos Correia Mariano Pêgo João Cardoso

PII: S0169-2607(14)00244-2
DOI: http://dx.doi.org/doi:10.1016/j.cmpb.2014.06.010
Reference: COMM 3816

To appear in: *Computer Methods and Programs in Biomedicine*

Received date: 16-9-2013
Revised date: 12-6-2014
Accepted date: 17-6-2014

Please cite this article as: Vânia G. Almeida, J. Borba, H. Catarina Pereira, Tânia Pereira, Carlos Correia, Mariano Pêgo, João Cardoso, Cardiovascular risk analysis by means of pulse morphology andclustering methodologies, *Computer Methods and Programs in Biomedicine* (2014), http://dx.doi.org/10.1016/j.cmpb.2014.06.010

Highlights

- We developed a clustering methodology to deal with Arterial Pressure Waveform (APW) parameters.
- A new non-invasive device is being evaluated.
- We collect a database of 116 subjects.
- This study compares different methods used in the APW analysis toward cardiovascular diseases.
- It is possible to assess potential cardiovascular risk in healthy subjects using cluster analysis.

Cardiovascular risk analysis by means of pulse morphology and clustering methodologies

Vânia G. Almeida[1], J. Borba[1], H. Catarina Pereira[1,2], Tânia Pereira[1], Carlos Correia[1], Mariano Pêgo[3] and João Cardoso[1]

[1] *Physics Department, Electronics and Instrumentation Group, University of Coimbra*

[2] *Intelligent Sensing Anywhere*

[3] *Cardiology Department, Hospital and University Coimbra Center*

Abstract

The purpose of this study was the development of a clustering methodology to deal with Arterial Pressure Waveform (APW) parameters to be used in the cardiovascular risk assessment. One hundred sixteen subjects were monitored and divided in two groups. The first one (23 hypertensive subjects) was analysed using APW and biochemical parameters, while the remaining 93 healthy subjects were only evaluated through APW parameters. The Expectation Maximization (EM) and k-means algorithms were used in the cluster analysis, and the risk scores (the Framingham Risk Score (FRS), the Systematic COronary Risk Evaluation Project (SCORE), the Assessing cardiovascular risk using Scottish Intercollegiate Guidelines Network guidelines (ASSIGN) and the PROspective Cardiovascular Münster (PROCAM)), commonly used in clinical practice were selected to the cluster risk validation. The result from the clustering risk analysis showed a very significant correlation with ASSIGN ($r = 0.582,\ p < 0.01$) and a significant correlation with FRS ($r = 0.458,\ p < 0.05$). The results from the comparison of both groups also allowed to identify the cluster with higher cardiovascular risk in the healthy group. These results give new insights to explore this methodology in future scoring trials.

Email address: vaniagalmeida@lei.fis.uc.pt (Vânia G. Almeida[1])

Keywords: Arterial Stiffness, Pulse Wave Analysis, Risk Scores, Clustering Analysis

1. Introduction

The atherosclerotic Cardiovascular Disease (CVD) is the most common cause of death worldwide, resulting from the combination of several risk factors [1]. The international guidelines [2, 3] consider that individuals with established CVD should be the first priority for preventive measures application. The concern in changing the current healthcare paradigm, from reactive towards preventive care, aims at identify individuals for risk in early stages of disease development, and then, direct more efforts and attention to the risk factors modification [4, 5]. Fortunately, this is an emergent tendency that can be addressed using the traditional risk scores, but also using innovative predictive algorithms.

During the last years many risk estimation systems have been developed in order to assist clinicians in the risk assessment, and in the individual chances prediction, for CVD development. The major challenges of these tools are the capabilities to: (1) identify high risk individuals, (2) weight the individual effects of all risk factors, (3) stratify or organize who needs lifestyle advice or medical therapy, and finally (4) avoid overmedicalization of individuals at low risk [6]. Taking this challenges into account several risk factors were identified, by their association with an increased risk for CVD development. CVD risk assessment tools differ from each other on the selected risk factors, the disease for what they were designed (Coronary Heart Disease (CHD), heart failure, *etc.*), the selected event type, the considered period of time (long or short term) and the cohort location. The most popular are: Framingham Risk Score (FRS), PROspective CArdiovascular Münster (PROCAM), ASsessing cardiovascular risk using Scottish Intercollegiate Guidelines Network (ASSIGN) and Systematic COronary Risk Evaluation project (SCORE).

These tools are important to help physicians in their daily practice. However,

its application in different populations remains a topic concerning attention. The research needs to be directed at refining the accuracy of prediction models and, most importantly, examining ways of turning them into effective clinical tools. Several risk prediction models for cardiovascular disease are available today and their head to head comparison and application in different populations would benefit from standardized reporting and formal, consistent statistical comparisons. The work presented by [7] reinforces this statement. The limitations of the comparison of different methods is associated to missing information, which makes difficult to reach robust conclusions about the best model or the ranking of models' performance. And, additionally most studies did not statistically compare the models that were examined. The inclusion of standardized reporting of discrimination, calibration, and reclassification metrics with formal statistical comparisons would contribute to the successful application of different risk scores in distinct populations.

The trends for the risk overestimation in low-risk populations and underestimation in high-risk groups have been successfully demonstrated by Cooney, *et al.* [6]. It is known that an examination of 5 % SCORE can equate to a $10-25$ % FRS risk, depending on which of the several FRS functions is selected [3]. Haq, *et al.* [8] studied several methods for risk estimation (FRS, PROCAM, Dundee, and British regional heart-BRHS) and the results demonstrated a close agreement between all these, regarding average risk and showed moderate agreement for estimation among individuals. Finally, it was also concluded that FRS function is acceptably accurate in northern European populations.

The arterial stiffness measurement currently assumes an increasing role in clinical assessment due to its predictive value in cardiovascular events in patients with various risk levels, such as it was demonstrated by several studies [9, 10, 11, 12]. There are several advantages of using non-invasive methods over invasive measurements, *e.g.*, the potential use in follow-up strategies in populations without symptomatic CVD, such as

children or young adults. Furthermore, non-invasive tools can be essential to the CVD assessment in addition to the established risk factors in populations at high risk aiming the prevention of coronary vascular diseases. Inferences about CVD progressive development can be assessed by the analysis of the mechanical properties of arteries through a variety of indices based on the Pulse Wave Analysis (PWA) [13, 14]. The analysis is based on the identification of the key features in the arterial pressure wave profile, such as Systolic Wave Transit Time (SWTT), Reflected Wave Transit Time (RWTT) and dicrotic notch (evaluated by Left Ventricular Ejection Time (LVET)), and can include time or amplitude considerations, as well as variability based parameters [15]. The wave reflections are often addressed, in terms of the Augmentation Index (AIx), which expresses the ratio of the "augmented pressure" assigned to the reflected wave towards each overall pulse.

Data mining techniques have attracted a great deal of attention due to their ability to extract implicit and potentially useful information from large volumes of data [16]. Their feasible implementation in Computer-Aided Diagnosis (CAD) methodologies has given new insights in the development of innovative and effective decision support systems for CVD premature risk assessment [15, 17, 18, 19]. An interesting approach is the exploration of different classifiers, as it was proposed by Jovic *et al.* [20]. The electrocardiogram (ECG) classification problem was addressed using a combination of several features in the analysis of the Heart Rate Variability (HRV). Other approach presented by Tsipouras *et al.* [18] was based on the development of a fuzzy rule-based decision support system for CAD diagnosis. On the other hand, multi-classifiers should perform better in some situations, overcoming errors from single classifier analysis [21]. The incorporating of the prediction outcome of each one of the individual classifier was suggested, as a way to reduce the classification errors [22].

On other hand, clustering analysis is another important branch of unsupervised learning that allows the arrangement of objects into groups (*i.e.*, the clusters), wherein

the objects in the same cluster are more similar (in one or more characteristics), than those in different clusters [23]. There is a wide variety of clustering methodologies available in literature, essentially organized in three general classes [24]. The three types include parametric model-based, hierarchical and partitioning algorithms. Shah *et al.* [25] has proved the usefulness and feasibility of using clustering risk factors in the detection of CVD in youth, by the comparison with the Patholobiological Determinants of Atherosclerosis in Youth (PDAY) risk score. Other studies have also referred the role of clustering methodologies for CVD assessment, such as the work developed by Haseena *et al.* [26], where a fuzzy C-mean clustered probabilistic neural network for ECG beats discrimination was described. Clustering methodologies were also successful applied in other medical fields, such as in the identification of patterns in blood glucose measurements and regular insulin doses taken before meal time [27].

Our aim is the development of a clustering methodology to deal with Arterial Pressure Waveform (APW) based parameters to cardiovascular risk assessment. The evaluation was performed through the strength of the relationships with traditional risk scores. In the current paper, Section 2 details the subjects and methods used during data analysis, including a quick and up-to-date literature survey on attempts for risk scores and clustering methods used. The results are presented and discussed in Sections 3 and 4, respectively. Finally, in Section 5 some guidelines for further research are presented along with the main conclusions of the current work.

2. Methodology

2.1. Database

The data used in this study were obtained from 116 subjects divided in two groups, as depicted in Figure 1(a). Data were collected with approval by the ethical committees of the Coimbra Hospital and University Centre (CHUC), Portugal, with informed consent. Hypertension was diagnosed when Systolic Blood Pressure (SBP) $\geq$ 140

mmHg and/or Diastolic Blood Pressure (DBP) $\geq$ 90 mmHg, or if the patient was taking anti-hypertensive medication. Age, smoking habits, and familiar hypertensive history were recorded by structured questioning in accordance with the criteria used by each one of the scores. Current smoking was defined as having smoked the last cigarette less than 1 year before. Diabetes mellitus was considered for those subjects that presented a fasting blood sugar level $>$126 mg/dl, or current prescription of an oral hypoglycemic drug or insulin.

Two groups of subjects were analysed: Group C and Group H. The group H inclusion criteria included diagnosed hypertension by a clinician and for Group C young subjects without any known cardiovascular complication.

- Group H consists of 23 hypertensive subjects, 10 men and 13 women. Lipidic values were measured: the serum total cholesterol (Total-CH), the high density lipoprotein cholesterol (CH-HDL) and the triglyceride (TGL) levels. All subjects were tested at the same time of day to avoid any diurnal variations. Additionally, the APW parameters were also computed: SWTT, RWTT, SWA, LVET and AIx.

- Group C consists of 93 young and healthy subjects between $18-30$ years. The APW parameters were computed: SWTT, RWTT, SWA, LVET and AIx.

2.2. PWA Measurements

APWs were recorded at a sampling rate of 1 kHz with a single non-invasive PZ probe developed in a previous work [28], and shown in Figure 1(a) (main text). The probe is held in place by a neck collar specially developed for carotid measurements. The mechanical interface between the probe head and the sensing point also plays an important role on the data quality. The probe head is based on a mushroom-shaped PZ sensor that transmits the distension associated to the pressure wave in such a way that transversal and shear effects are suppressed and only radial applied forces are allowed. Accuracy tests were performed at a bench test, where 1.80% was the maximum Root

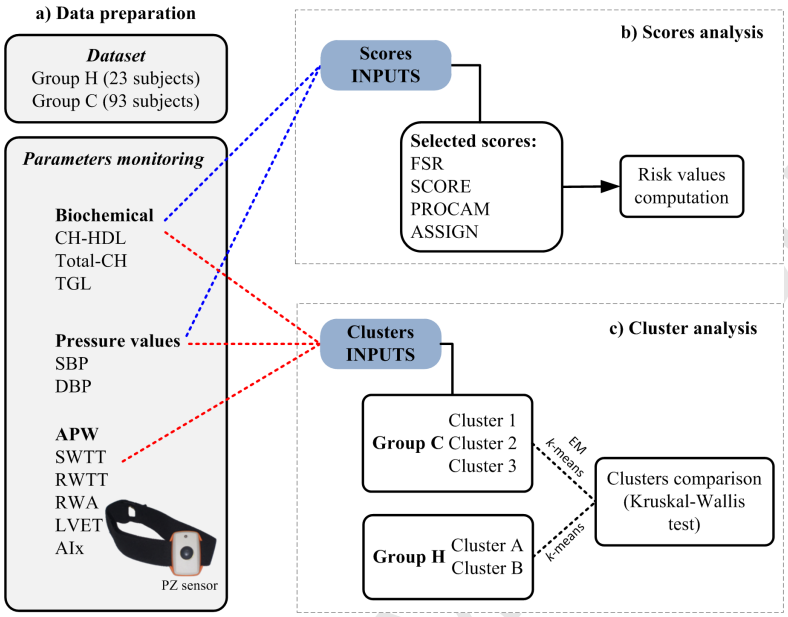

Figure 1: (a) Firstly, biochemical (CH-HDL - High density lipoprotein cholesterol and Total-CH - Cholesterol Total), pressure values (SBP - Systolic Blood Pressure and DBP - Diastolic Blood Pressure) and APW parameters (RWTT - Reflected Wave Transit Time, RWA - Reflection Wave Amplitude, SWTT - Systolic Wave Transit Time, LVET - Left Ventricular Ejection Time and AIx - Augmentation Index) were collected. (b) Four risk scores were studied to use as the reference along this work, namely: FRS - Framingham Risk Score, PROCAM - PROspective CArdiovascular Münster risk score, ASSIGN - ASsessing cardiovascular risk to Scottish Intercollegiate Guidelines Network and SCORE - Systematic COronary Risk Evaluation. (c) The Expectation Maximization (EM) and *k*-means were the selected clustering algorithms, and the Kruskal-Wallis Test was used to compare groups.

Mean Square Error (RMSE) introduced by the electronic circuits and by the mechanical interface.

Some data concerning validation was also published in [29] comprising three sets of recordings for the carotid pressure waveform at left and right carotid arteries, under standardized conditions, in 20 volunteers by three trained operators. Inter and intra-operator differences were calculated being good indicators, similar to other data reported in literature for commercial devices [30].

For each subject at least three consecutive measurements were performed. APW

parameters were tabulated for the set of pulses, considering each one of the subjects, after to the segmentation process. The PWA parameters, schematically represented in Figure 2, are related to amplitude and temporal characteristics of APW, namely:

- Reflected Wave Transit Time (RWTT) is determined by the time interval between the foot of the carotid pressure waveform to the first inflection point, which corresponds to the foot of the global reflected pressure wave.
- Reflection Wave Amplitude (RWA) is measured at inflection point in reference to the normalized amplitude.
- Systolic Wave Transit Time (SWTT) is defined as the interval between the waveform foot and the systolic peak.
- Left Ventricular Ejection Time (LVET) is measured by the time interval from the foot of the waveform to the dicrotic notch.
- Augmentation Index (AIx) is computed as the quotient between the reflected wave amplitude and the pulse pressure (PP), expressed as a percentage value. The equations used for its computation were defined by Murgo, *et. al.* (1980) [31], presented below. A type A waveform is defined when the systolic peak occurs in late systole after the inflection point and the type C when the systolic peak precedes the inflection point.

$$AIx = \pm \frac{SWA - RWA}{PP}$$

2.3. Risk Scores Selection

Four risk scores were considered along this work namely, FRS, PROCAM, ASSIGN and SCORE. The FRS was developed from a general population in Framingham, Massachusetts, USA, while the remaining scores derive from European

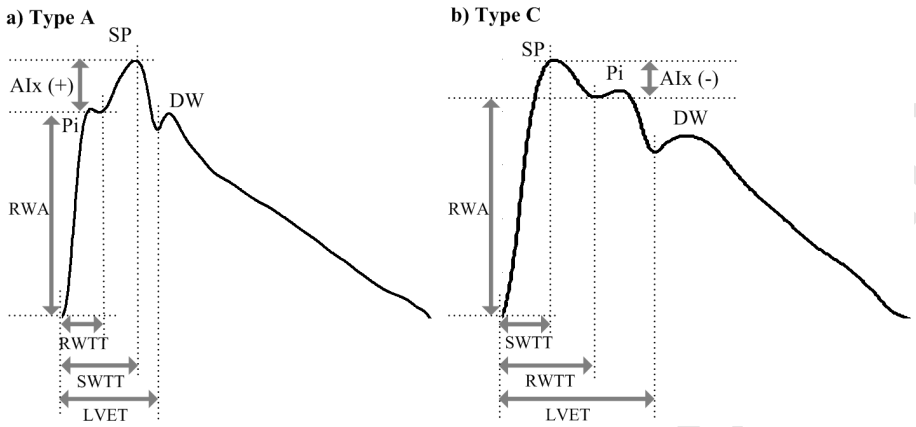

Figure 2: Arterial pressure waveforms measured non-invasively at carotid artery. In a) a type A contour is represented , where the peak systolic pressure (SP) occurs in late systole after an inflection point (Pi) that resulted from the reflection wave. In these conditions AIx is computed as a positive value. In b) a type C contour is represented. Notice that in a type C contour the SP precedes the Pi and AIx is computed as a negative value. The dicrotic wave is represented by DW.

studies. The largest is the SCORE, since it consists of $\approx$ 205 000 subjects from 12 cohort studies from European countries. The other two, ASSIGN and PROCAM, were developed in Scotland and Germany, respectively. In general, these tools have distinct characteristics associated, different risk factors, specific disease, event type, period of time and cohort locations. Table 1 summarizes the main characteristics for each one.

2.4. Statistical Analysis

The data were analysed using the SPSS 16 statistical package (SPSS Inc., Chicago, Illinois, USA). Descriptive statistics were conducted to describe the sample characteristics. The normality results were assessed by the Kolmogorov-Smirnov test. To determine the relationships among variables, the Spearman rank-order correlation test was used due to the non-normal nature of the distribution. Afterwards, also due to the non-normal distribution the Kruskal-Wallis test was conducted to cluster comparisons. A p value < 0.05 was considered statistically significant.

Table 1: Risk assessment tools (10 years term).

Model	Patients	Country of Origin	Risk factors
FRS	8491	USA	Age, Gender, Total-CH, CH-HDL, SBP, BPT, SMK
SCORE	205178	Finland, Russia, Norway, Denmark, UK (England), UK (Scotland), Sweden, Belgium, Germany, Italy, France and Spain	Age, Gender, Total-CH, CH-HDL, SBP, and SMK
ASSIGN	13297	Scotland	Age, Gender, FH, DB, SMK, SBP, Total-CH and CH-HDL
PROCAM	5389	Germany	Age, CH-LDL, SMK, CH-HDL, SBP, PE, DB, and TGL

Total-CH − Total Cholesterol, BPT − Blood Pressure Treatment, FH − Family History, TGL − Triglycerides, SMK − Smoking habits, DB − Diabetes, PE − Previous Event, SBP − Systolic Blood Pressure, CH-HDL − High-Density Lipoprotein, CH-LDL − Low Density Lipoprotein.

2.5. Clustering Analysis

During the cluster assignments, the pulses are independently grouped according to their cluster similarities. The adopted strategy for the determination of the ideal number of clusters consisted, firstly, in the selection of two clusters ($k = 2$) followed by a tentative to split each of these clusters. During the process, clusters can also be merged if they are sufficiently close or, if there is too many patterns and unusually large variance.

The maximum level of similarity for each subject occurs when all pulses fit in only one cluster. Expected Maximization (EM) and *k*-means clustering algorithms are the most studied. Weka 3.6.8 framework software was the selected tool to use during the cluster analysis (Figure 1(c)).

2.5.1. k-Means

The *k*-means algorithm is a type of partitional clustering that continuously iterates until a specific criterion function (usually the square error) converges. It acknowledges

the number of desired cluster inputs (k) and divides the set of objects (n) into k clusters. The result is a higher intra-cluster and lower inter-cluster similarity. The cluster similarity is measured considering the mean value of the objects contained in the cluster, which is, in fact, the cluster centroid. k-Means clustering is relatively scalable and efficient in processing large datasets. However, it cannot handle categorical attributes and it is not suitable for dealing with non-convex shapes. Also, it is quite sensitive to the presence of noise and outliers. It requires an efficient data pre-processing analysis before its application [23].

2.5.2. Expectation-Maximization

The Expectation-Maximization (EM) clustering algorithm is a complex probabilistic extension of the k-means method that primarily differs by the way how the initial groups are obtained. Instead of assigning each object to a cluster, with which it is most similar, EM assigns each object to a cluster according to a weight that represents the probability of membership. In this manner, there are no strict boundaries between clusters, and new means are determined based on weighted measures [23].

3. Results

3.1. Subject Characteristics

The clinical characteristics of our study population are shown in Table 2. The mean age of subjects in Group H is 58.10 ± 11.64 years. The blood pressure values in this group are elevated (170.98 ± 11.41 mmHg and 100.29 ± 9.05 mmHg for SBP and DBP, respectively) relatively to normal ranges, $130 - 139$ mmHg for SBP and $85 - 89$ mmHg for DBP [3]. TGL is also a marker of increased risk in this group due to be higher than the guideline recommendations (< 150 mg/dL) [3]. The same situation is verified for Total-CH values, which are also superior to the guideline recommendations (< 190 mg/dL). However, the CH-HDL is within the optimal limits, as it is higher than the minimum threshold of $40 - 45$ mg/dL. Finally, it is also possible

Table 2: Characteristics of subjects

Variable	Group H (mean±SD)	Group C (mean±SD)	Units
Number of subjects	23	93	–
Gender(m/f)	12/11	31/62	–
Age	58.10 ±11.64	21.19 ± 2.28	years
SBP	170.98 ±11.41	108.26 ± 11.88	mmHg
DBP	100.29 ±9.05	69.54 ± 7.64	mmHg
BMI	27.87 ±5.23	21.62 ± 2.63	kg/m^2
Total-CH	205.10 ±25.42	–	mg/dL
CH-HDL	65.70 ±28.80	–	mg/dL
TGL	155.25 ±28.71	–	mg/dL
SWTT	224.08 ±52.86	163.45±60.64	ms
RWTT	115.83 ±27.93	159.29 ±43.47	ms
RWA	0.72 ±0.14	0.87±0.09	a.u.[a]
AIx	25.69 ± 17.14	0.35±15.69	%
LVET	305.40 ±47.46	281.99 ±53.06	ms

SBP - Systolic Blood Pressure, DBP - Diastolic Blood Pressure, BMI - Body Mass Index, CH-HDL-High Density Lipoprotein Cholesterol, Total-CH - Cholesterol Total, TGL - Triglycerides, RWTT - Reflected Wave Transit Time, RWA-Reflection Wave Amplitude, SWTT - Systolic Wave Transit Time, LVET - Left Ventricular Ejection Time, AIx - Augmentation Index.

[a] Arbitrary amplitude units

to conclude that the mean Body Mass Index (BMI) in Group H is also superior to the ideal recommendations (< 25 kg/m^2) [3].

Regarding Group C, the mean age is 21.19±2.28 years, and the pressure values are within normal ranges. Focusing on the APW differences between Group H and Group C, SWTT occurs later in the first group, where the arrival time assumes the value of 224.08 ± 52.86 ms, while in Group C, this is at 163.45 ± 60.64 ms. LVET arrival time is quite similar for both groups: 305.40 ± 47.46 ms in Group H and 281.99 ± 53.06 ms in Group C. The RWTT occurs earlier in Group H $(115.83 \pm 27.93$ms) at lower amplitude (0.72 ± 0.14), in opposition to the values in Group C (159.29 ± 43.47 ms, at the amplitude of 0.87 ± 0.09). Group H is also characterized by higher positive AIx values (25.69 ± 17.14 %), in opposition to the Group C (0.35 ± 15.69 %).

3.2. Clustering Analysis

The cluster analysis was performed for Group C and Group H, independently. Different nomenclatures were adopted to help understanding the data at hand, numeric labels (1, 2, ...) for the Group C and alphabet labels (A, B, ...) for the Group H. After, the cluster analysis, the clusters were compared using the Kruskal-Wallis test.

3.2.1. Group C

The first approach consisted in the selection of the best clustering algorithm to deal with the set of features at test, as well as the ideal number of clusters to the group characterization. Figure 3 displays RWTT and SWTT plot for 2-clusters analysis using EM (a) and k-means (b) algorithms. Red and blue coloured points represent the pulse labels (blue = Cluster 1, red = Cluster 2). Categorical features (gender, smoker) were not considered during the analysis, since k-means analysis is not able to deal with this kind attributes.

The figure shows that the EM plot has an unsatisfactory division between the clusters, with some points from Cluster 2 identified as being in the Cluster 1 area. Visually, the results from the k-means clustering have a more efficient separation, as the dataset is partitioned in two homogeneous risk groups. For both EM and k-means, the Cluster 1 represents the pool of healthier subjects when compared to the Cluster 2, since it represents the cases where the reflection wave arrives after to the systolic peak. The analysis of Cluster 2 also indicates the presence of another sub-cluster. Taking this into account, the k-means method was used to explore the distribution of a third cluster, since it performed well, comparatively to the EM, in the two cluster distribution. The obtained results for a 3 cluster distribution are presented in Figure 4. The detailed information about clusters distribution is presented in Table 3. The Cluster 3 (green homogeneous zone) is mostly representative of Type C APW pulses, where $RWTT > SWTT$. The mean AIx value of the cluster centroid is negative (-11.2 %), thus being considered the lower risk cluster. Cluster 2 predominantly

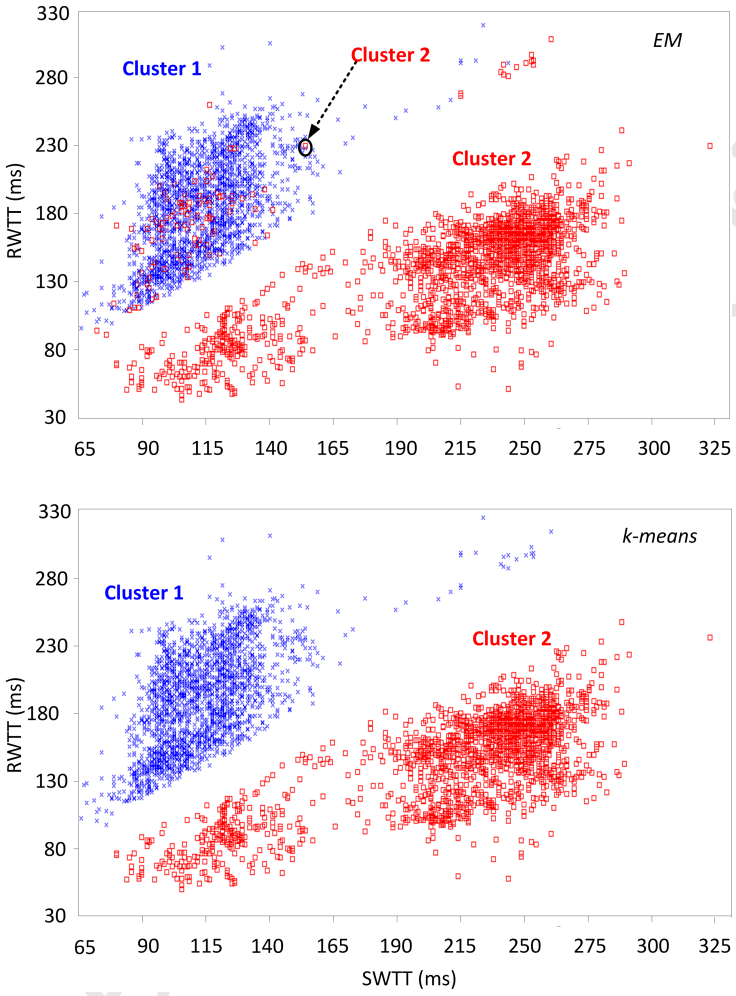

Figure 3: Clusters performance obtained for EM (top) and k-means (bottom) algorithms using two clusters, Cluster 1=blue, Cluster 2=red.

consists of Type B pulses, with $SWTT > RWTT$ and $AIx > 0$, being thus considered the intermediate risk group. Cluster 1 pulses represents the less homogeneous group, with some points also scattered across the Cluster 2 area. These pulses are mainly APW Type A pulses, with some punctual Type B pulses. This group evidences a higher cardiovascular risk comparatively to the Cluster 2 and Cluster 3.

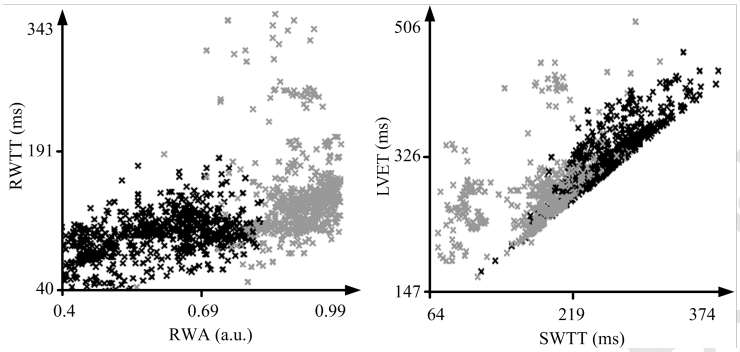

Figure 4: RWTT and SWTT plot after *k*-means clustering for using three clusters. Blue=Cluster 1, red=Cluster 2 and Green=Cluster 3.

3.2.2. Group H

The *k*-means was also used in the Group H analysis. Figure 5 depicts the 2-cluster distribution, where Cluster A and Cluster B were the labels adopted.

Table 3: Average values for the 3-clusters groups.

Attributes	**Cluster**		
	1	**2**	**3**
Pulses	458	1550	2463
Age (y)	21.6	21.0	21.7
Weight (kg)	63.0	55.4	63.2
Height (m)	1.7	1.6	1.7
BMI (kg/m^2)	21.3	20.8	21.8
SBP (mmHg)	109.9	106.1	108.6
DBP (mmHg)	69.6	70.3	68.9
HR (bpm)	72.8	67.5	72.9
SWTT (ms)	172.7	234.4	117.1
RWTT (ms)	103.0	143.2	179.9
LVET (ms)	240.3	306.4	274.4
SWA (a.u. [a])	1.0	1.0	1.0
RWA (a.u. [a])	0.8	0.9	0.9
DWA (a.u. [a])	0.8	0.8	0.7
AIx (%)	21.0	12.6	-11.2

[a] Arbitrary Amplitude Units

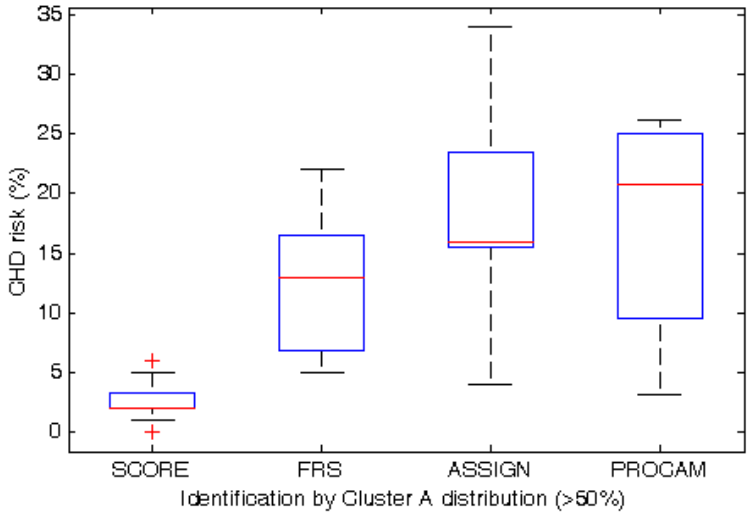

Figure 5: Scatter plots for: (a) RWTT and RWA, (b) LVET and SWTT, where the grey and black markers denote Cluster A and Cluster B, respectively.

The characteristics of each of the clusters are presented in Table 4. It can be verified that the clusters do not differ significantly according to the SBP and DBP values. The most significant parameters (except APW parameters) were the CH-HDL and the TGL

Table 4: Cluster distributions for subjects in Group H.

Attributes	**Cluster A**	**Cluster B**
SBP (mmHg)	169.68±10.38	172.56±13.59
DBP (mmHg)	101.22±7.03	102.00±7.14
Total-CH (mg/dL)	206.50±26.16	204.01±13.69
CH-HDL (mg/dL)	59.35±25.07	66.33±23.31
TGL(mg/dL)	160.41±44.95	142.94±40.96
RWTT (ms)	99.90±23.62	140.32±42.76
RWA (a.u. [a])	0.60±0.11	0.90±0.06
SWTT (ms)	257.34±35.76	185.34±40.70
LVET (ms)	326.24±39.43	272.16±44.17
AIx (%)	40.01±11.15	6.2±9.67
Clustered instances	723 (52 %)	664 (48 %)

[a] Arbitrary Amplitude Units

values. Cluster A is characterized by higher TGL values and lower CH-HDL levels, characteristics of subjects at risk. From the waveform analysis, lower RWTT, higher SWTT and, consequently, higher positive AIx values were obtained. The suitable number of clusters was considered to be two as may be confirmed by visual inspection.

During the evaluation, the pulse instances were independently grouped according to their cluster similarities, being the maximum level of similarity achieved, for each subject, when all the pulses fit in a single cluster. So, for each subject a value representing the cluster risk (as percentage) was computed and used in the correlation with other available parameters (including the studied scores). Since only two clusters are studied, at this point, the values of Cluster A (%) are symmetric to the values of Cluster B (%). The Cluster A (%) correlations are presented in Table 5.

The lipidic and the BP values present low significance with the Cluster A (%). On the other hand, significant correlation values were obtained between Cluster A (%)

Table 5: Spearman's correlation coefficients obtained for Cluster A (%).

Attributes	**Cluster A (%)**
SBP	-0.131
DBP	-0.080
Total-CH	0.011
CH-HDL	-0.177
TGL	0.017
AIx	0.921 **
SWTT	0.649 **
RWTT	-0.632 **
RWA	-0.933 **
LVET	0.574 **
FRS	0.458 *
SCORE	0.275
ASSIGN	0.582 **
PROCAM	0.391

Significant levels: ** $p < 0.01$, *$p < 0.05$

and PWA parameters. Additionally, it was obtained a significant correlation between Cluster A (%) and the risk scores. In this case was observed a very significant correlation with ASSIGN ($r = 0.582,\ p < 0.01$), a significant correlation with FRS ($r = 0.458,\ p < 0.05$) and, weaker correlations with SCORE ($r = 0.275,\ p = 0.241$) and PROCAM ($r = 0.391,\ p = 0.088$). These results are good indicators for the use of this methodology as a tool for the cardiovascular risk assessment.

Figure 6 shows the cluster A (%) distribution (considering a threshold of more than 50 %) for each of the scores. Targeting data by cluster A (threshold >50%), obtained a median risk by the SCORE function of 2.0% per year, FRS function of 13.0% per year, ASSIGN function of 16.0% per year and by the PROCAM function of 20.7% per year. For the FRS, ASSIGN and PROCAM scores, these values correspond to the borderline between high risk and an intermediate risk [32, 33]. However, for SCORE, the observed median value corresponds to low risk, being SCORE the less correlated

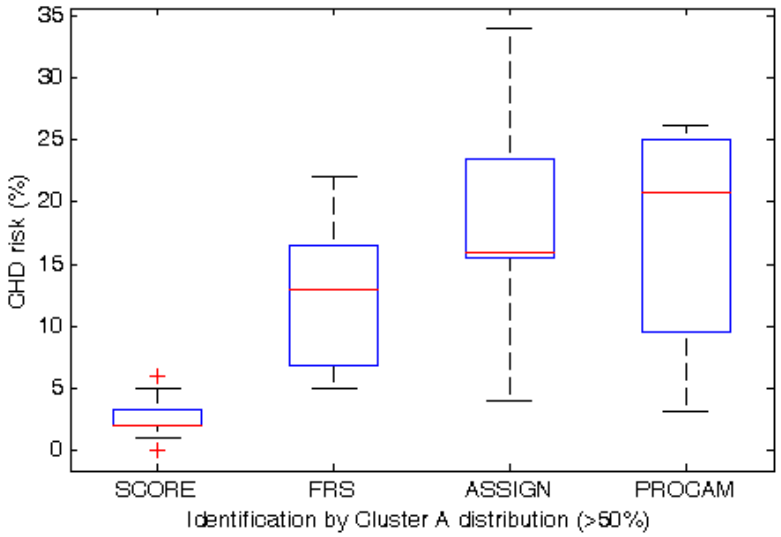

Figure 6: Boxplots of CHD risk distribution for SCORE, FRES, ASSIGN and PROCAM. The horizontal lines represent the medians, the boxes represent the interquartile ranges (50% of the distribution) and the whiskers represent the range of values obtained for subjects from GroupH.

function with the cluster results.

3.2.3. Comparison of Groups

The evaluation was performed by the comparison of the clusters in Group H (Cluster A and Cluster B) with clusters in Group C (Cluster 1, Cluster 2 and Cluster 3). The Cluster A is the cluster at higher risk in the hypertensive groups, as previously discussed. Since, there is no medical information about the risk associated to Group C (we have only informations concerning the APW parametrizations), each one of clusters from Group C was compared to the Group H clusters. The comparison was performed using the chi-squared (χ^2) value, obtained from the Kruskal-Wallis test, as shown in Table 6.

There are significant differences for the majority of clusters in Group H and Group C, as expected due to the distinct population characteristics of each group. However, some similarities were found between the Cluster B (the score at lower risk in Group H) and the Cluster 1 (belonging to the Group C), namely: for LVET ($\chi^2 = 4.79,\ p < 0.01$) and SWTT ($\chi^2 = 0.02,\ p < 0.01$) measures. From the analysis of Table 3, it is possible to conclude that this is the cluster at higher risk in Group C. However, it is not possible to conclude that the risk associated is effectively risk

Table 6: Comparison of clusters from Group H (Cluster A and Cluster B) and Group C (Cluster 1, Cluster 2 and Cluster 3) using Kruskal-Wallis test, measured through χ^2 value.

Group H	**Group C**	**Parameters**					
		RWTT	*RWA*	*SWTT*	*RWTT*	*LVET*	*AIx*
ClusterA	Cluster1	38.66 **	402.42**	526.65**	38.66**	369.23**	402.40 **
	Cluster2	1149.82**	1399.96 **	248.87 **	1149.82**	154.81**	1399.97**
	Cluster3	1454.05**	1545.50**	1619.16**	1454.05**	706.65**	1638.37**
ClusterB	Cluster1	536.38 **	458.25**	0.02	536.38**	4.79	538.34 **
	Cluster2	265.77**	3.16 **	892.34 **	265.77**	598.03**	49.97 **
	Cluster3	593.88**	4.06**	998.62**	593.88**	60.22**	1158.33**

Significant levels: ** $p < 0.01$

associated to the development of CVD. This conclusion is only possible from the comparison with the clusters of the hypertensive group (Group H). And, the similarities are evident for the cluster in analysis (Cluster 1) and the cluster at lower risk in the hypertensive group (Cluster B), leading to assume that the subjects belonging to the Cluster 1 need medical advice, and that they may be developing CVD. It would not be trustable, if the similarities occur for the group of subjects in advanced stage of disease (Cluster A). This conclusion supports that is possible to screen a healthy population (using only waveform parameters) concerning the cardiovascular risk using clustering methodologies.

4. Discussion

In this paper a new approach to morphological pulse analysis is presented, and an innovative methodology to cardiovascular risk assessment, taking as reference the CHD risk scores, is applied.

The information that is extracted from the clustering analysis can be crucial to fully understand of the data, mainly when there is no, or little, available information. It was verified, that Total-CH and TGL values are intrinsically related to the APW variables, and than an increase of these levels is associated to the RWTT decrease and SWTT increase. On the other hand, the higher CH-HDL values are associated to the increase of RWTT and decrease of SWTT values. Significant correlations for the cluster output with the ASSIGN ($r = 0.582,\ p < 0.01$) and with FRS ($r = 0.458,\ p < 0.05$) were also verified.

This method is particularly interesting, considering that this approach can avoid the requirements of the classification procedures which often require costly labelling of a large set of patterns, such as biochemical analysis used by traditional risk scores, as demonstrated on the cluster analysis of subjects in Group C. It was possible to identify the similarities with Group H clusters, using only morphological parameters. The

possibility to emulate the performance of the traditional risk scores by the use of only non-invasive parameters, which can be easily obtained using a PWA technique, is an efficient alternative approach to study large datasets. In this sense, clustering analysis is an easy method for large populations screening producing valuable knowledge for posterior prioritising people for pharmacological measures.

The healthcare public and private system faces the challenge of an increasingly aging population and the escalation of medical costs. Therefore, the development of techniques that allow for earlier identification of subjects at risk, by the incorporation of personalized, predictive and preventive methodologies, could bring an interesting impact comparatively to the traditional cardiovascular tools used for risk prediction. It is evident the foreseeable impact that an accurate, non-invasive and easy-to-use instrument for hemodynamic condition assessment could impart on the diagnosis and follow-up of the CVD. However, a similar methodology could be applied to other measurements (*e. g.*, invasive pressures, electrocardiogram) used in traditional clinical path of cardiovascular patients.

5. Conclusions

This paper demonstrates the utility of clustering techniques in risk scoring when applied to a medical dataset. When compared with the traditional risk scores, clustering methodologies showed good performance with significant correlation values. This is a simple, yet reliable tool, that can be used in scoring trials.

As this approach needs some developments, a computational tool to integrate these results with other machine learning techniques such as classification algorithms, is currently being developed [15]. A larger sample will also be studied in future trials for the stratification by medication use, age, diabetes duration, and/or gender. Given these considerations, this application can be an interesting methodology to be used in further clinical studies to pulse morphological analysis. It can also be potentially

useful in other medical applications, such as in the anaesthesiology room, where the number of parameters and devices are high, as a tool that incorporates several important information.

Acknowledgements

Authors acknowledge *Fundação para a Ciência e Tecnologia* for funding scholarship SFRH/BD/61356/2009 and PTDC/SAU-BEB/100650/2008, project developed under the initiative of QREN, funding by UE/FEDER, through *COMPETE-Programa Operacional Factores de Competitividade*. The authors also thanks to ISA-Intelligent Sensing Anywhere and Hospital and University Coimbra Center (*C.H.U.C.*) for the collaboration in clinical trials.

References

[1] W. H. Organization, The global burden of disease: 2004 update, World Health Organization.

[2] R. E. W. Kavey, S. R. Daniels, R. M. Lauer, D. L. A. Atkins, L. L. Hayman, K. Taubert, American heart association guidelines for primary prevention of atherosclerotic cardiovascular disease beginning in childhood, Circulation 107 (11) (2003) 1562–1566.

[3] J. Perk, G. De Backer, H. Gohlke, I. Graham, Z. Reiner, M. Verschuren, C. Albus, P. Benlian, G. Boysen, R. Cifkova, C. Deaton, S. Ebrahim, M. Fisher, G. Germano, R. Hobbs, A. Hoes, S. Karadeniz, A. Mezzani, E. Prescott, L. Ryden, M. Scherer, M. Syvanne, W. J. Scholte op Reimer, C. Vrints, D. Wood, J. L. Zamorano, F. Zannad, European guidelines on cardiovascular disease prevention in clinical practice (version 2012). The fifth joint task force of the european society of cardiology and other societies on cardiovascular disease

prevention in clinical practice (constituted by representatives of nine societies and by invited experts). developed with the special contribution of the european association for cardiovascular prevention & rehabilitation (eacpr), Eur Heart J 33 (13) (2012) 1635–701.

[4] A. Beswick, P. Brindle, Risk scoring in the assessment of cardiovascular risk, Curr Opin Lipidol 17 (2006) 375–386.

[5] M. D. Whitfield, M. Gillett, M. Holmes, E. Ogden, Predicting the impact of population level risk reduction in cardio-vascular disease and stroke on acute hospital admission rates over a 5 year period–a pilot study, Public Health 120 (12) (2006) 1140–8.

[6] M. T. Cooney, A. L. Dudina, I. M. Graham, Value and limitations of existing scores for the assessment of cardiovascular risk: a review for clinicians, J Am Coll Cardiol 54 (14) (2009) 1209–27.

[7] G. C. M. Siontis, I. Tzoulaki, K. C. Siontis, J. P. A. Ioannidis, Comparisons of established risk prediction models for cardiovascular disease: systematic review, BMJ 344. doi:10.1136/bmj.e3318.

[8] I. U. Haq, L. E. Ramsay, W. Yeo, P. R. Jackson, E. J. Wallis, Is the framingham risk function valid for northern european populations? a comparison of methods for estimating absolute coronary risk in high risk men, Heart Vessels 81 (1999) 40–46.

[9] G. M. London, A. P. Guerin, Influence of arterial pulse and reflected waves on blood pressure and cardiac function, Am Heart J 138 (3) (1999) 220–223.

[10] J. D. Cameron, B. P. McGrath, A. M. Dart, Use of radial artery applanation tonometry and a generalized transfer function to determine aortic pressure

augmentation in subjects with treated hypertension, JACC 32 (5) (1998) 1214–20.

[11] S. Laurent, J. Cockcroft, L. V. Bortel, P. Boutouyrie, C.Giannattasio, D. Hayoz, B. Pannier, C. Vlachopoulos, I. Wilkinson, H. StruijkerBoudier, Expert consensus document on arterial stiffness methodological issues and clinical applications, European Heart Journal 27 (21) (2006) 2588–2605.

[12] I. Ikonomidis, S. Tzortzis, T. Papaioannou, A. Protogerou, K. Stamatelopoulos, C. Papamichael, N. Zakopoulos, J. Lekakis, Incremental value of arterial wave reflections in the determination of left ventricular diastolic dysfunction in untreated patients with essential hypertension, J Hum Hypertens 22 (10) (2008) 687–98.

[13] A. P. Avolio, M. Butlin, A. Walsh, Arterial blood pressure measurement and pulse wave analysis–their role in enhancing cardiovascular assessment, Physiol Meas 31 (1) (2010) R1–47.

[14] B. Hametner, S. Wassertheurer, J. Kropf, C. Mayer, A. Holzinger, B. Eber, T. Weber, Wave reflection quantification based on pressure waveforms alone—methods, comparison, and clinical covariates, Computer Methods and Programs in Biomedicine 109 (3) (2013) 250 – 259.

[15] V. Almeida, J. Vieira, P. Santos, T. Pereira, H. Pereira, C. Correia, M. Pego, J. Cardoso, Machine learning techniques for arterial pressure waveform analysis, Journal of Personalized Medicine 3 (2) (2013) 82–101.

[16] I. Yoo, P. Alafaireet, M. Marinov, K. PenaHernandez, R. Gopidi, J. Chang, L. Hua, Data mining in healthcare and biomedicine: A survey of the literature, J. Med. Syst.

[17] K. B. Wagholikar, V. Sundararajan, A. W. Deshpande, Modeling paradigms for

medical diagnostic decision support: a survey and future directions, J Med Syst 36 (5) (2012) 3029–49.

[18] M. G. Tsipouras, T. P. Exarchos, D. I. Fotiadis, A. Kotsia, K. Vakalis, K. K. Naka, L. K. Michalis, Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling, IEEE Transactions On Information Technology In Biomedicine 12 (4) (2008) 447–458.

[19] S. Paredes, T. Rocha, P. de Carvalho, J. Henriques, M. Harris, J. Morais, Long term cardiovascular risk models' combination, Comput Methods Programs Biomed 101 (3) (2011) 231–42.

[20] A. Jovic, N. Bogunovic, Electrocardiogram analysis using a combination of statistical, geometric, and nonlinear heart rate variability features, Artificial Intelligence in Medicine 51 (2011) 175–186.

[21] L. I. Kuncheva, J. C. Bezdek, R. P. Duin, Decision templates for multiple classifier fusion: An experimental comparison, Pattern Recognition Letters 34 (2).

[22] D. Ruta, B. Gabrys, An overview of classifier fusion methods, Computing and Information Systems 7.

[23] J.Han, M. Kamber, Data mining. Concepts and techniques, Elsevier, San Francisco, CA, 2006.

[24] V. Melnykov, G. Shen, Clustering through empirical likelihood ratio, Computational Statistics Data Analysis.

[25] A. S. Shah, L. M. Dolan, Z. Gao, T. R. Kimball, E. M. Urbina, Clustering of risk factors: a simple method of detecting cardiovascular disease in youth, Pediatrics 127 (2) (2011) e312–8.

[26] H. H. Haseena, A. T. Mathew, J. K. Paul, Fuzzy clustered probabilistic and multi layered feed forward neural networks for electrocardiogram arrhythmia classification, J Med Syst 35 (2) (2011) 179–88.

[27] J. Namayanja, V. P. Janeja, An assessment of patient behavior over time–periods: A case study of managing type 2 diabetes through blood glucose readings and insulin doses, Journal of Medical Systems 36 (S1) (2012) 65–80.

[28] V. G. Almeida, H. C. Pereira, T. Pereira, E. Figueiras, E. Borges, J. M. R. Cardoso, C. Correia, Piezoelectric probe for pressure waveform estimation in flexible tubes and its applications to the cardiovascular system, Sensors and Actuators A 169 (2011) 217–226.

[29] V. Almeida, H. Pereira, T. Pereira, L. Ferreira, C. Correia, J. Cardoso, Assessment of the pulse wave variability for a new non-invasive device, in: Y.-T. Zhang (Ed.), The International Conference on Health Informatics, Vol. 42 of IFMBE Proceedings, Springer International Publishing, 2014, pp. 240–243.

[30] M. Frimodt-Moller, A. H. Nielsen, A. L. Kamper, S. Strandgaard, Reproducibility of pulse-wave analysis and pulse-wave velocity determination in chronic kidney disease, Nephrol Dial Transplant 23 (2) (2008) 594–600.

[31] J. Murgo, N. Westerhof, J. P. Giolma, S. Altobelli, Aortic input impedance in normal man: relationship to pressure wave forms, circulation, Circulation 62 (1980) 105–116.

[32] M. Versteylen, I. Joosen, L. Shaw, J. Narula, L. Hofstra, Comparison of framingham, procam, score, and diamond forrester to predict coronary atherosclerosis and cardiovascular events, Journal of Nuclear Cardiology 18 (5) (2011) 904–911.

[33] Assign score website, `http://assign-score.com/` (2014).