34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Improvement of CVD Risk Assessment Tools' performance through innovative Patients' Grouping Strategies

S. Paredes†, T. Rocha†, P. de Carvalho‡, J. Henriques‡, J. Morais, J. Ferreira§, M. Mendes§*

Abstract – There are available in the clinical community several practical risk tools to assess the risk of occurrence of a cardiovascular event. Although valuable, these tools typically present some lack of performance (low sensitivity/low specificity) when applied to a general (average) patient.

This flaw is addressed in this work through an innovative personalization strategy that is supported on the evidence that current risk assessment tools perform differently among different populations/groups of patients.

The proposed methodology is based on two main hypotheses: *i*) patients are grouped through a proper dimension reduction technique complemented with an unsupervised learning algorithm, *ii*) for each group the most suitable risk assessment tool can be selected improving the risk prediction performance. As a result, risk personalization is simply achieved by the identification of the group that patients belong to.

The validation of the strategy is carried out through the combination of three current risk assessment tools (GRACE, TIMI, PURSUIT) developed to predict the risk of an event in coronary artery disease patients. The combination of these tools is validated with a real patient testing dataset: Santa Cruz Hospital, Portugal, N=460 ACS-NSTEMI[1] patients.

Considering the obtained results with the available dataset it is possible to state that the main objective of this work was achieved.

I. Introduction

The cardiovascular disease[2] (CVD) disease is the world's largest killer, responsible for 17.1 million deaths per year [1]. The correct diagnosis and prognosis of CVD is essential to reduce these statistics. In this context, the assessment of the risk of an event's occurrence, i.e. the evaluation of the probability of occurrence of an event given the patient's past and current exposure to risk factors, is critical to improve prognosis [2].

Several risk tools[3] were developed to assess the probability of occurrence of a CVD event within a certain period of time. These tools are very useful although they present some important weaknesses: *i*) they ignore the information provided by other risk assessment tools that were previously developed, *ii*) each individual tool considers a reduced number of risk factors, *iii*) they have difficulty in coping with missing risk factors, *iv*) they do not allow the incorporation of additional clinical knowledge, *v*) some tools do not assure the clinical interpretability of the respective parameters. These problems have already been addressed in previous works of this research team [3][4][5].

The problem of lack of performance exhibited by those tools is the main focus of this work. In fact, current tools often present sensitivity/specificity[4] values that do not assure a proper classification of the patients' risk when applied to a particular population. A viable alternative is the development of a new tool, specific for the population under analysis. This work addresses the problem researching a different approach. In order to circumvent this lack of performance, an innovative methodology is proposed. It is supported on the evidence that current risk assessment tools perform differently among different populations/groups of patients, which originates two main hypotheses: *i*) it is possible to group patients through a proper dimension reduction strategy complemented by an unsupervised learning algorithm; *ii*) for each particular group it is possible to select the most appropriate current risk assessment tool, such that the CVD risk of a patient that belongs to a given group can be accurately estimated.

This approach was validated with current risk assessment tools specific for secondary prevention on coronary artery disease (CAD) patients. The GRACE, TIMI (no ST-elevation) and PURSUIT were the selected tools [6][7][8]. The validation phase was supported by a real patient testing dataset: Santa Cruz Hospital, Lisbon/Portugal, N=460 ACS-NSTEMI patients.

The paper is organized as follows: in section II an outline of the developed methodology is presented. In section III the results of the validation procedure with the Santa Cruz dataset are discussed. Section IV summarizes the main conclusions and the main research paths to be followed up in the near future.

This work was supported by HeartCycle EU project (FP7-216695) and CISUC (Center for Informatics and Systems of University of Coimbra).

† Instituto Politécnico de Coimbra, Departamento de Engenharia Informática e de Sistemas, Portugal, {sparedes@isec.pt, teresa@isec.pt}.
‡ CISUC, Departamento de Engenharia Informática, Universidade de Coimbra, Coimbra, Portugal, {carvalho@dei.uc.pt, jh@dei.uc.pt}.
*Serviço de Cardiologia, Hospital Santo André, EPE, Portugal, {joaomorais@hsaleiria.min-saude.pt}.
§Serviço de Cardiologia, Hospital Santa Cruz, Lisboa, Portugal, {jorge_ferreira@netcabo.pt, miguel.mendes.md@sapo.pt }.

[1] ACS-NSTEMI Acute Coronary Syndrome with non-ST segment elevation.

[2] Cardiovascular disease is caused by disorders of the heart and blood vessels, including coronary heart disease (heart attacks), cerebrovascular disease (stroke), raised blood pressure (hypertension), peripheral artery disease, rheumatic heart disease, congenital heart disease and heart failure.

[3] In order to clarify, risk assessment models that have been statistically validated and are available in literature are going to be designated through this work as **risk assessment tools**.

[4] $SE = TP/(TP+FN)$; $SP = TN/(TN+FP)$; TP: True Positive; TN: True Negative; FN: False Negative; FP: False Positive

II. Methodology

The proposed methodology (Figure 1) is composed of two main phases: *i*) Grouping of patients; *ii*) Selection of risk tools.

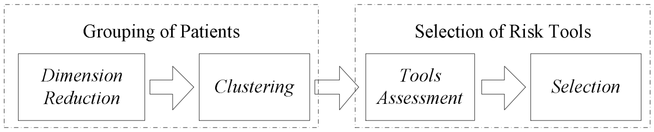


Figure 1 – Proposed Methodology.

As mentioned, the proposed personalization strategy relies on the creation of groups of patients. However, the heterogeneity of risk factors (quantitative/qualitative data, binary data) that usually characterize a specific patient, along with their high dimensionality (number of risk factors) constrain the derivation of those groups. Therefore, the reduction of dimensionality is implemented in order to facilitate/improve the clustering process. The second phase concerns the selection of the most suitable tool to classify patients from a given cluster.

A. *Grouping of Patients*

This phase involves two steps: *i*) dimension reduction; *ii*) clustering.

The dimension reduction, aiming for the creation of a low dimensional representation of a high dimensional data while preserving most of the intrinsic information[5] contained in the original data, can be very useful to facilitate the clustering process [9]. The second step consists of a clustering procedure, where groups of patients are created based on the information obtained through the dimension reduction procedure.

1) Dimension Reduction

The reduction of dimensionality can be formalized as: given a P dimensional data vector $\mathbf{x} = [x_1 ... x_P]^T$, a lower dimensional representation $\mathbf{y} = [y_1 ... y_Q]^T$ should be found with $Q \le P$, such that it captures the content in the original data according to some criterion [11].

There are two major categories of dimension reduction methods: *i*) linear methods, where each one of the components of $\mathbf{y}$ is a linear combination of the original variables, such that $\boldsymbol{Y}_{Q\times N} = \boldsymbol{W}_{Q\times P}\boldsymbol{X}_{P\times N}$ where P is the dimension of original data, Q denotes the dimension of the lower dimensional representation and N is the number of instances. Among the linear methods it is possible to identify the Principal Component Analysis (PCA), Independent Component Analysis [9][10]; *ii*) non-linear techniques, where it is not possible to determine a linear transformation weight matrix $\mathbf{W}$. Maaten [10] presents a very comprehensive overview on non-linear methods that are grouped in three main categories: *i*) Global techniques (isomap, neural networks); *ii*) Local techniques (Laplacian eigenmaps); *iii*) Global alignment of linear models (manifold charting).

However, in this work a different approach is followed. The reduction of dimensionality process is supported on the individual risk assessment tools (non-linear mapping). In effect, this approach seems very appropriate in this particular problem as these tools were developed to classify patients that are characterized by a set of heterogeneous risk factors. Additionally, this non-linear mapping allows the uniformization of each patient's data.

Thus, all instances $\mathbf{x}_i = [x_1^i ... x_P^i]^T \in \mathbf{X}_{P\times N}$, that correspond to the N patients are mapped into $\mathbf{y}_i \in \mathbf{Y}_{Q\times \mathrm{N}}$, $i = 1,...,N$ where y_q^i denotes the output of tool q to classify the patient i (e.g. $\mathbf{y}_i = [y_R^i\ y_P^i\ y_T^i\ ]$[6]). All the y_q^i should be normalized into the interval $[0,1]$

2) Clustering

This phase is responsible for the creation of the patient groups. Basically, using the proposed approach, patients are grouped based on the outputs of the risk tools instead on the initial risk factors. Let $\mathbf{Y}_{Q\times N}$ represent a set of N patients, the goal is to apply a clustering algorithm to $\mathbf{Y}$ in order to create K disjoint groups (clusters) $G = \{G_1,...,G_K\}$ of patients with similar characteristics.

The clustering process should assume that the dimension of the clusters must be defined considering the concept that supports the methodology, i.e. if the cluster is too big it may not provide a differentiation among the performance of the several risk assessment tools otherwise if the cluster is too small it will be impossible to apply the concept of patient grouping. Among the several clustering algorithms available, the subtractive clustering was selected [12][13].

B. *Selection of Risk Tools*

The performance of the several individual tools is assessed within each group of patients (created in the previous phase). This allows that each cluster be assigned the tool that presents the best performance. The final classification of a particular patient that belongs to a given cluster corresponds to the classification of the individual tool that has the best performance with patients from that cluster.

1) Tools Assessment

Each one of the considered individual risk assessment tools is tested within each cluster. Assuming that a risk assessment tool q considers J risk factors (subset of the P risk factors), an instance $\mathbf{x}_i^q$ (containing the J risk factors (values) of patient i) is applied to the q tool in order to obtain the respective $y_q^i \in \mathbf{y}_i$. Each y_q^i is normalized to the interval $[0,1]$ and converted to a risk class c_q^i according to

[5] The intrinsic dimensionality of data is the minimum number of parameters needed to account for the observed properties of data [10].

[6] In this work the models to combine are g**R**ace, **P**ursuit and **T**imi.

the original specifications of each tool [6][7][8]. Then for each patient i of each cluster G_k, $k = 1,...,K$ the output (class) of each tool q is compared with the real data (occurrence of an event) within a given period of time. This assessment allows computing the sensitivity and specificity of the risk prediction achieved by each tool.

2) Selection

The final classification of a patient i is based on the selection of the most suitable risk tool for its classification (Figure 2).

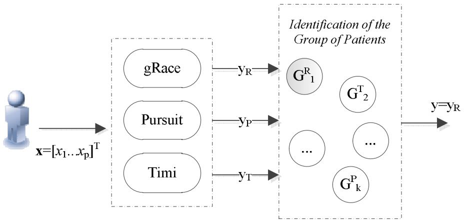

Identification of the Group of Patients
gRace
Pursuit
Timi
$\mathbf{x}=[x_1...x_p]^T$
y_R
y_P
y_T
G^R_1
G^T_2
...
...
G^P_k
$y=y_R$

Figure 2 – Classification[7]

The classification process may be depicted as follows: *i*) the different risk assessment tools assess the risk of a new patient i based on $\mathbf{x}_i$ in order to obtain $\mathbf{y}_i$; *ii*) the cluster G_k that the patient i belongs to is identified based on $\mathbf{y}_i$; *iii*) the best tool q to classify patients from G_k is selected. The final classification is provided by that tool.

The criteria to select the best tool q to classify patients from a cluster G_k can be given by:

```
IF $SE_q = 100\%$ THEN tool $q$ classifies cluster $G_k$
ELSE IF $SP_q = 100\%$ THEN tool $q$ classifies $G_k$
ELSE $G_k$ is classified by the best tool in the global dataset
END IF
```

C. Validation

The classification process (Figure 2) is implemented for the N patients that integrate the testing dataset in order to validate the model. The assessment of the classification performance is done through the comparison of obtained results with the real data. In this phase, Bootstrapping validation is adopted to reinforce the reliability of the obtained results.

III. Results

A. Testing Datasets

1) Santa Cruz Hospital Dataset

This dataset contains data from N=460 consecutive patients that were admitted in the Santa Cruz Hospital, Lisbon, with ACS-NSTEMI between March 1999 and July 2001. Table I presents the main clinical characteristics of such patients [14]. Continuous variables with a normal distribution are expressed as mean value and standard deviation. Discrete variables are presented as frequencies and percent values.

TABLE I
Clinical characteristics of patients that integrate the dataset

Model	Event
Age (years)	63.4 ± 10.8
Sex (Male/Female)	361 (78.5%) / 99 (21.5%)
Risk Factors:	
Diabetes (0/1)	352 (76.5%) / 108 (23.5%)
Hypercholesterolemia (0/1)	180 (39.1%) / 280 (60.9%)
Hypertension (0/1)	176 (38.3%) / 284 (61.7%)
Smoking (0/1)	362 (78.7 %) / 98 (21.3%)
Previous History / Known CAD	
Myocardial Infarction (0/1)	249 (54.0%) / 211 (46.0%)
Myocardial Revascularization (0/1)	239 (51.9%) / 221 (48.1%)
PTCA	146 (31.7%)
CABG	103 (22.4%)
Sbp (mmHg)	142.4 ± 26.9
Hr (bpm)	75.3 ± 18.1
Creatinine (mg/dl)	1.37 ± 1.26
Enrolment [0 UA, 1 MI]	180 (39.1 %) / 280 (60.9%)
Killip 1/2/3/4	395 (85.9%) / 31 (6.8%) / 33 (7.3 %) / 0%
CCS [0 I/II; 1 CSS III/IV]	110 (24.0%) / 350 (76.0%)
ST Segment Deviation (0/1)	216 (47.0%) / 244 (53.0%)
Signs of Heart Failure (0/1)	395 (85.9%) / 65 (14.1%)
Tn I > 0.1 ng/ml (0/1)	313 (68.0%) / 147 (32.0%)
Cardiac Arrest Admission (0/1)	460 (100%) / 0%
Aspirin (0/1)	184 (40.0%) / 276 (60.0%)
Angina (0/1)	19 (4.0%) / 441 (96.0%)

The event rate of combined endpoint (death/myocardial infarction) is 7.2% (33 events).

B. Individual Risk Assessment Tools

Table II presents the selected individual risk assessment tools to predict death/MI for CAD patients within a short period.

TABLE II
Short-TERM Risk Assessment Models

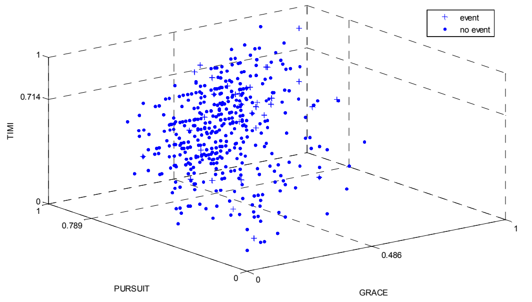
Model	Event	Time	Prev. Type	Risk Factors
GRACE [6]	D MI	6 m	S	Age, SBP, CAA HR, Cr, STD, ECM, CHF
PURSUIT [7]	D MI	30 d	S	Age, Sex, SBP, CCS, HR, STD, ERL, HF
TIMI [8]	D MI/UR	14 d	S	Age, STD, ECM, KCAD, AS, AG, RF

D: Death; **MI**: Myocardial Infarction; **UR**: Urgent revascularization; **m**: months; **d**: days; **S**: Secondary Prevention; **Cr**-Creatinine, **HR** – Heart Rate, **CAA** – Cardiac Arrest at Admission, **CHF** – Congestive Heart Failure, **STD** - ST Segment. Depression, **ECE** - Elevated Cardiac Markers, **KCAD**- Known CAD, **ERL** – Enrolment (MI/UA), **HF** – Heart Failure, **CCS** – Angina classification, **AS** - Use of aspirin in the previous 7 days, **AG** - 2 or more angina events in past 24 hrs, **RF** - 3 or more cardiac risk factors.

C. Dimensionality Reduction

The dataset after the dimensionality reduction from the original $P = 16$ risk factors to $Q = 3$ outputs of the risk tools is presented in Figure 3.

[7] *G_k^q denotes that tool q has the best performance on cluster G_k*



$$\mathbf{y}_i = [y_R^i \; y_P^i \; y_T^i \;]; \; c_R^i = \begin{cases} 0; \; y_R^i \le 0.486 \\ 1; \; y_R^i > 0.486 \end{cases} ; c_P^i = \begin{cases} 0; \; y_P^i \le 0.789 \\ 1; \; y_P^i > 0.789 \end{cases} ; c_T^i = \begin{cases} 0; \; y_T^i \le 0.714 \\ 1; \; y_T^i > 0.714 \end{cases}$$

Figure 3 – Dimensionality Reduction

D. Groups of Patients

Subtractive clustering was applied based on $\mathbf{Y}_{3\times 460}$.

TABLE IV
PERFORMANCE OF SELECTED INDIVIDUAL RISK ASSESSMENT TOOLS

C	GRACE		PURSUIT		TIMI		P	E
	SE	SP	SE	SP	SE	SP		
1	0	100	0	96.7	0	100	31	1
2	0	100	0	100	0	100	34	1
3	80	26.9	40	46.2	100	0	31	5
4	100	25	25	25	0	100	24	4
5	0	100	0	100	0	100	20	0
6	0	100	0	100	0	95	20	0
7	100	95.8	0	100	100	0	25	1
8	0	90.5	0	76.2	0	100	21	0
9	0	100	0	84.2	0	89.5	19	0
10	0	84.6	100	23.1	0	100	14	1
11	100	0	100	5.6	100	0	21	3
12	0	15	100	35	100	0	22	2
13	0	100	0	100	0	100	14	2
14	0	100	0	31.3	0	0	16	0
15	0	64.3	0	100	0	100	15	1
16	0	100	0	100	0	100	26	1
17	0	100	0	93.8	0	100	17	1
18	100	0	66.7	20	0	100	21	6
19	0	75	0	100	0	100	12	0
20	0	90.9	0	100	0	100	12	1
21	0	100	0	100	0	100	12	0
22	100	0	50	11.1	0	100	11	2
23	0	100	0	100	0	71.4	22	1

***C**: Clusters; **SE**: Sensitivity (%); **SP**: Specificity (%), **P**: Patients; **E**: Events*

The performance of each tool was assessed in each cluster (Table IV) according to the procedure detailed in Section II.B.1.

E. Validation

As referred the Bootstrapping validation ($N_B = 1000$ samples) was applied to the original dataset with the aim of reinforcing the obtained results (Table V):

TABLE V
PERFORMANCES COMPARISON – **SANTA CRUZ, (DEATH/MI)**

	%	GRACE	PURSUIT	TIMI	Groups
Boot. samples n=1000	SE	60.8 *(60.2; 61.3)*	42.4 *(41.9;43.1)*	33.5 *(33.0; 34.0)*	72.9 *(72.6; 73.5)*
	SP	74.9 *(74.8; 75.1)*	74.2 *(74.1;74.3)*	73.6 *(73.5; 73.7)*	74.9 *(74.8; 75.1)*

It is possible to conclude that the proposed combination of risk assessment tools achieves a higher sensitivity than all the individual tools (the best individual sensitivity is 60.8% while the sensitivity for the proposed strategy is 72.9%). The specificity values are equivalent among the several models (the best individual specificity 74.9% equals the value obtained through the proposed strategy). Statistical significance tests (Student's t-test) confirmed this conclusion.

IV. CONCLUSIONS

This work addressed the problem of lack of performance exhibited by CVD risk assessment tools, when applied to a particular patient. The proposed personalization approach focused the proper selection of these tools, based on the evidence that their performance differs among different groups of patients. Therefore, the creation of groups of patients where it is possible to identify a tool that assures a good performance was the main issue of this methodology.
The obtained results confirm that is possible to achieve higher sensitivity values without reducing the specificity values. These results are very promising, suggesting the potential of this approach to enhance the performance of current risk assessment tools in a clinical practice context. Its application to other populations will be the next step in future work, which will give additional significance to the developed strategy.

V.REFERENCES

[1] World Health Organization, "Cardiovascular Diseases (CVDs)", fact sheet n°317.: http://www.who.int/mediacentre (December 2010)

[2] Graham, I. et. al., "Guidelines on preventing cardiovascular disease in clinical practice: executive summary", European Heart Journal, Vol.28, 2375 – 2414, 2007.

[3] S. Paredes, T. Rocha, P. Carvalho, J. Henriques, M. Harris, J. Morais, "Long Term Cardiovascular Risk Models' Combination", Computer Methods and Programs in Biomedicine Journal, 2011.

[4] S. Paredes, T. Rocha, P. Carvalho, J. Henriques, M. Harris, J. Morais, "Cardiovascular Risk and Status Assessment", 32th Annual International IEEE EMBS Conference, Argentina, 2010.

[5] S. Paredes, T. Rocha, P. de Carvalho, J. Henriques, D. Rasteiro, J. Morais, J. Ferreira, M. Mendes, "Fusion of Risk Assessment Models with application to Coronary Artery Disease Patients ", 33th Annual International IEEE EMBS Conference, USA, 2011.

[6] Tang. E, et. al., "Global Registry of Acute Coronary Events(GRACE) hospital discharge risk scores accurately predicts long term mortality post-acute coronary syndrome", AHJ, Vol. 154, pp. 29-35 2007.

[7] Antman, E. et. al., "The TIMI risk score for Unstable Angina / Non-St Elevation MI – A method for Prognostication and Therapeutic Decision Making", JAMA ,Vol. 284, pp. 835-842, 2000.

[8] Boersma E., K. Pieper, E. Steyerberg, "Predictors of outcome in patients with acute coronary syndromes without persistent ST-segment elevation. Results from an international trial of 9461 patients", Circulation 101;2557–2567, 2000.

[9] Sugiyama et al, "Semi-supervised local Fisher discriminant analysis for dimensionality", Machine Learning Vol.78; n°1,2; pp.35-61, 2010.

[10] Maaten L. et al, "Dimensionality Reduction: A Comparative Review", Tilburg University Technical Report, TiCC-TR 2009-005, 2009.

[11] Fodor I., "A Survey on Reduction Techniques", Lawrence Livermore National Laboratory; Technical Report UCRL-ID-148494, 2002.

[12] Han J., Kamber M. and Pei J. "Data Mining: Concepts and Techniques", 3rd edition, Morgan Kaufmann, 2011.

[13] Hammouda, K., "A Comparative Study of Data Clustering Techniques, SYDE 625: Tools of intelligent systems design", Course Project, University of Waterloo, 2000.

[14] Gonçalves P. et al. "TIMI, PURSUIT and GRACE risk scores: sustained prognostic value and interaction with revascularization in NSTE-ACS", European Heart Journal, Vol. 26, pp. 865-872, 2005.