35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Cardiovascular Disease Risk Assessment Innovative approaches developed in *HeartCycle* project

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

Abstract –Two innovative CVD event risk assessment strategies were developed in the scope of HeartCycle project: *i)* combination of individual risk assessment tools; *ii*) personalization of risk assessment based on grouping of patients.

These approaches aimed to defeat some of the major limitations of the tools currently applied in the daily clinical practice, namely to: *i)* improve the risk prediction performance when comparing it to the one achieved by the individual current risk assessment tools; *ii)* consider the available knowledge provided by other risk assessment tools; *iii)* cope with missing risk factors; *iv)* incorporate additional clinical knowledge.

Two different real patients' datasets were applied to validate the developed strategies: *i)* Santa Cruz Hospital, Portugal, N=460 ACS-NSTEMI[1] patients; *ii)* Leiria Pombal Hospital Centre, Portugal, N=99 ACS-NSTEMI.

Based on the gathered results, we propose a new strategy in order to improve patients' stratification.

I. Introduction

Cardiovascular disease[2] (CVD) remains the leading cause of premature death worldwide. However, prevention can be very effective since more than 50% of the reduction seen in Coronary Heart Disease (CHD) mortality is due to changes in modifiable risk factors [1]. Actually, 77% of the disease burden in Europe is accounted for disorders related to lifestyle, while 80% of CHD could be prevented by maintaining healthy lifestyles [2].

In this context the current health care paradigm must be changed. The health system has to change from reactive care towards preventive care and simultaneously transfer the care from the hospital to the patient's home. Health telemonitoring systems are very important as they allow the remote monitoring of patients who are in different locations away from the health care provider. Clinical data (weight, blood pressure, electrocardiogram, etc.) can be collected, processed or sent to the care provider. As a result of the data processing, feedback can be provided directly to the patient as well as to the care provider. This remote monitoring is more challenging to the care provider, as the reliability/quality of the clinical decision must be guaranteed in order to optimize the patient's care plan.

HeartCycle project is one of these systems and it intends to provide a closed-loop disease management solution for CHD and HF patients [3]. The CVD risk assessment, i.e. the evaluation of the probability of occurrence of an event[3] given the patient's past and current exposure to risk factors, is critical to achieve that goal [4]. The risk assessment is in fact a key factor to help the clinical decision as well as to motivate the patient increasing the treatment compliance with the corresponding health benefits (patient seen as a co-producer of health).

Few topics have received as much attention in the cardiovascular research area over the last years as risk prediction [5]. Several risk assessment tools[4] were developed to assess the probability of occurrence of a CVD event within a certain period of time. Although useful, they present some important weaknesses as they: *i*) may present some lack of performance; *ii)* ignore the information provided by other risk assessment tools that were previously developed; *iii*) consider (each individual tool) a limited number of risk factors; *iv*) have difficulty in coping with missing risk factors; *v*) do not allow the incorporation of additional clinical knowledge; *vi*) do not assure the clinical interpretability of the respective parameters; *vii*) impose a selection of a standard tool to be applied in the clinical practice.

The identified weaknesses were addressed through the development of two different methodologies: *i*) combination of individual risk assessment tools; *ii*) personalization based on grouping of patients.

These approaches were applied to current risk assessment tools specific for secondary prevention CHD patients, where GRACE, TIMI-NSTEMI and PURSUIT were the selected tools [6][7][8]. The validation phase was supported by two real patient testing datasets: *i)* Santa Cruz Hospital, Lisbon/Portugal, N=460 ACS-NSTEMI patients; *ii)* Leiria Pombal Hospital Centre, Portugal, N=99 ACS-NSTEMI.

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

This work was partially financed by iCIS (CENTRO-07 -ST24-FEDER-002003), 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 Death, myocardial infarction, hospitalization, etc.

4 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.

II. METHODOLOGY

The proposed methodologies are presented in Figure 1.

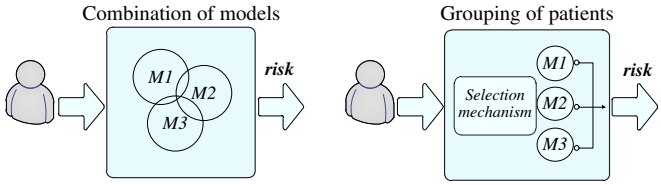

Figure 1 – Proposed Methodologies.

The combination methodology creates a flexible framework that is able to combine a set of distinct current risk assessment tools. The methodology is based on two main hypotheses: *i*) it is possible to implement a common representation (naïve Bayes classifier) of the individual risk assessment tools; *ii*) it is possible to combine individual models exploiting the particular features of Bayesian probabilistic reasoning.

The second methodology, personalization based on grouping of patients, is proposed as an approach to enhance the performance of the risk prediction when compared to the one obtained with current risk assessment tools. It is based on two 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.

A. Combination of Individual Risk Assessment Tools

The implementation of this approach is composed of two main phases: 1) common representation (naïve Bayes classifier) of individual tools; 2) a combination scheme that exploits the probabilistic nature of naïve-Bayes inference mechanism.

1) Common Representation of Individual Tools

Current individual risk score tools are diversely represented (equations/scores/charts) which hinder their combination. To defeat this obstacle, all the individual risk score tools were represented as naïve-Bayes classifiers. This classifier was selected since it is efficient, simple and can deal with lack of input information (missing risk factors) [9]. Its inference mechanism is given by:

$$P(C \mid \mathbf{x}) = P(C \mid X_1,...,X_p) = \alpha\ P(C) \prod_{i=1}^{p} P(X_i \mid C) \qquad (1)$$

where $\mathbf{x} = \{X_1,...,X_p\}$ is a set of observations (clinical examination, laboratory measurements, etc.) and C a hypothesis (e.g. risk level is "High"). The term $P(C \mid \mathbf{x})$ is the probability that the hypothesis is correct after observations have occurred (e.g., the probability that risk is "High" given the results of a clinical examination, measurements, etc.). $P(C)$ is the probability that the hypothesis is correct before seeing any observation (in this example, the prevalence of the risk level). $P(X \mid C)$ is a likelihood expressing the probability of the observation X being made if the hypothesis is correct (equivalent to the sensitivity of the clinical examination). α is a normalization constant.

This inference mechanism (naïve Bayes) assumes that observations (attributes) are conditionally independent, given the value of hypothesis C. However, even if this condition is not verified, naïve Bayes often presents a good performance [9].

Conditional probability tables $P(X \mid C)$ of each individual tool were derived based on equations/scores available in literature [6][7][8]. The training dataset must contain all the risk factors that belong to the different individual tools. As a result, conditional probability tables for the p attributes were constructed based on (2).

$$P(X_i = x_j \mid C = c_k) = \frac{\sum_{1}^{m} (X_i = x_j \wedge C = c_k)}{\sum_{1}^{m} (C = c_k)} \quad i = 1,....,n \qquad (2)$$

It is assumed that class C has several categories (mutually exclusive), where variable c_k denotes the k class label of variable C. Furthermore, it is assumed that variable x_j denotes a particular value of the attribute X_i and m is the total number of training instances.

2) Individual Models Parameters' Weighted Average

The combination strategy implemented through (3), must be able to assign different weights for the individual models according to their performance in a specific dataset.

$$\begin{aligned} P(C) &= \sum_{j=1}^{l} P(C_j) \times \frac{w_j}{\Gamma} \quad where \quad \Gamma = \sum_{j=1}^{l} w_j \\ P(X_i \mid C) &= \sum_{j=1}^{b} P(X_i^j \mid C_j) \times \frac{w_j}{\vartheta} \quad where \quad \vartheta = \sum_{j=1}^{b} w_j \end{aligned} \qquad (3)$$

The value l is the number of individual models, b is the number of individual models that contain the attribute X_i, C_j denotes each individual model, w_j is the weight of model j. The combination scheme includes an optimization based on genetic algorithms. It intends to adjust the models' parameters that result from the combination strategy in order to improve the performance of the global model [10].

B. Personalization Based on Grouping of Patients

The proposed personalization strategy relies on the creation of groups of patients and on the selection of the proper risk assessment tool.

1) Grouping of Patients

This phase involves two steps: *i*) dimension reduction; *ii*) clustering. The dimension reduction process is supported on the individual risk assessment tools (non-linear mapping). This approach seems very appropriate 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 $\mathbf{y}_i \in \mathbf{Y}_{Q \times N}, i = 1,...,N$ where y_q^i denotes the output of tool q to classify the patient i. This

dimension reduction can be very useful to facilitate the clustering process [11].

Clustering, applied through subtractive clustering [12], is responsible for the creation of the patient groups. Patients are grouped, based on the outputs of the risk tools ($\mathbf{Y}$), in order to create K disjoint groups (clusters) of patients with similar characteristics. The clustering process should assume that 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.

2) 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 to 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. A more detailed explanation of this methodology can be found in [13].

Figure 2 represents the classification process, where G_k^q denotes the tool q with the best performance on cluster G_k .

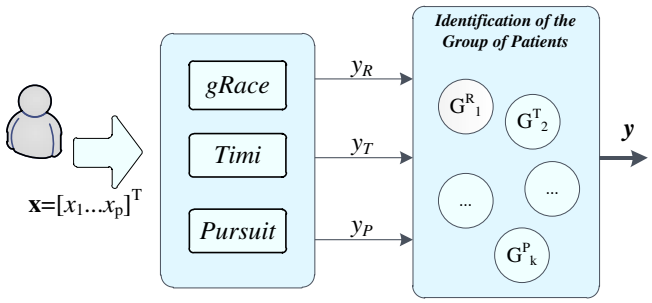

Figure 2 – Classification.

III. RESULTS

A. Combination of Individual Risk Assessment Tools

1) Testing datasets

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

The Leiria-Pombal Hospital Centre comprises N=99 ACS-NSTEMI patients admitted during 2007. There were 5 events of the observed endpoint (30days / death), which originated an endpoint rate of 5.1%.

The training data set was created a $\mathbf{x}^i = [x_1^i ... x_p^i]$ for all i; $1 \le i \le N$: with $N = 1000$, based on the approach proposed in [14].

2) Individual Risk Assessment Tools

This combination methodology was applied to three individual risk assessment tools (GRACE, TIMI, PURSUIT) developed to predict death/MI for CHD patients within a short period [6][7][8].

3) Individual Models' Combination

The Bayesian global model was derived according to the methodology explained in II. The global voting model was implemented considering the votes (0/1) of the three individual models.

TABLE I

PERFORMANCES COMPARISON – SANTA CRUZ, (DEATH/MI)

	%	GRACE	PURSUIT	TIMI	ByG	Vot
Original	SE	60.6	42.4	33.3	60.6	48.5
	SP	74.9	74.2	73.5	67.0	75.6
	Gmean	67.3	56.0	49.4	63.4	60.6
	AUC	0.675	0.575	0.525	0.635	0.625
Boot Samples n=1000	SE	60.8 (60.2; 61.3)	42.4 (41.9;43.1)	33.5 (33.0; 34.0)	60.6 (60.1;61.3)	48.6 (48.0;49.2)
	SP	74.9 (74.8; 75.1)	74.2 (74.1;74.3)	73.6 (73.5; 73.7)	67.0 (66.9;67.2)	75.6 (75.5;75.8)
	Gmean	67.3 (67.0; 67.6)	55.8 (55.5;56.2)	49.3 (48.9; 49.7)	63.6 (63.3;63.9)	60.3 (60.0;60.7)

SE: Sensitivity; SP: Specificity; D: Death; MI: Myocardial Infarction; *(;)=95% Confidence Interval;* ByG – Bayesian Global Model, Vot - Voting

4) Optimization

The methodology can be adjusted to a specific population. Table II presents the optimization results, obtained through a genetic algorithm approach.

TABLE II

PERFORMANCES COMPARISON

		Santa Cruz 30 days/D/MI		Santa Cruz 30 days/D		Santo André 30 days/D	
		ByG	ByG AO	ByG	ByG AO	ByG	ByG AO
Original	SE	60.6	72.7	61.5	76.9	80.0	80.0
	SP	67.0	69.1	65.7	70.7	67.0	82.9
	G_{mean}	63.4	70.9	63.5	73.7	73.2	81.5
	AUC	0.635	0.7	0.625	0.725	0.725	0.8
Boot Samples n=1000	SE	60.6 (60.1;61.3)	72.9 (72.4;73.4)	61.6 (60.7;62.5)	77.3 (76.5;78.0)	80.3 (78.9;81.5)	79.8 (78.6;81.0)
	SP	67.0 (66.9;67.2)	69.1 (69.0;69.2)	65.8 (65.6;65.9)	70.6 (70.5,70.8)	66.8 (66.4;67.2)	83.8 (83.3;84.2)
	G_{mean}	63.6 (63.3;63.9)	70.9 (70.6;71.1)	63.1 (62.7;63.6)	73.6 (73.3;74.0)	72.3 (71.5;73.1)	80.9 (80.0;81.6)

ByG – Bayesian Global Model; **ByG AO** – Bayesian Global Model After Optimization.

The Bayesian global model presents a better performance than the other models. It is also possible to conclude that genetic algorithms' optimization improved the performance of the Bayesian global model.

The ability of the different classifiers to deal with missing risk factors was also assessed through the comparison of the Bayesian approach (before and after the optimization procedure) with the voting model. It was possible to conclude that in the majority of the test cases the global Bayesian

model after optimization presented the best performance (highest sensitivity/highest specificity).

B. Personalization Based on Grouping of Patients

This methodology was applied to the Santa Cruz hospital dataset (combined endpoint, D/MI). GRACE, TIMI and PURSUIT were the selected individual risk assessment tools to validate this second approach.

As referred, the first step was the dimensionality reduction from the original $P = 16$ risk factors to $Q = 3$ outputs of the risk tools. Through subtractive clustering, 23 clusters were obtained based on $\mathbf{Y}_{3\times460}$. The performance of each tool was assessed in each cluster.

Table III presents the main validation results where, similarly to the previous validation procedure, Bootstrapping validation ($N_B = 1000$ samples) was applied to the original dataset with the aim of reinforcing the obtained results:

TABLE III
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 achieved 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%).

IV. FUTURE WORK AND CONCLUSIONS

Figure 3 presents the future developments of this wok:

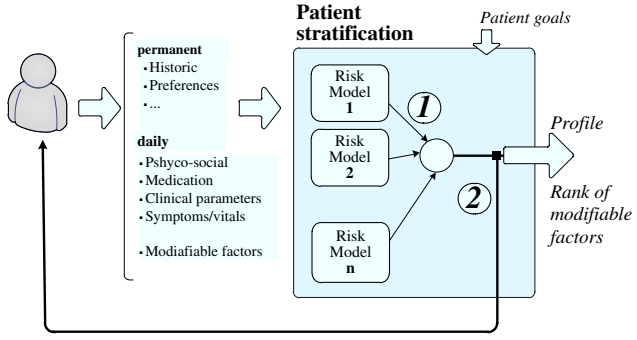

Figure 3 – Future Developments

The main focus should be the enhancement of the patient stratification, as this is the key aspect to improve the patient's care plan.

Patient stratification should be supported based on two main aspects: *i)* Risk level assessment; *ii)* Ranking of modifiable risk factors.

The former should consider the combination of the well-known and validated risk assessment tools (available in the medical community) into a global multi-model approach. This framework should be complemented with a personalization approach, through group personalization which is directly related with the two developed methodologies in this work.

The latter, is concerned with modifiable factors' prediction strategies that must be merged with the outcomes of the risk level assessment module. Here, optimization techniques must be applied to improve patients' stratification. The aim is to forecast what intervention is most likely to work best for each patient. In particular, the procedure is carried out according to the identification of the relevant modifiable risk factors in order to achieve the patient's goal (usually low risk profile).

To conclude, it is possible to affirm that the initial goals of this work were achieved while the obtained results are very promising. However some future work should be pursued in order to improve patient stratification and consequently the respective care plan.

V. REFERENCES

[1] Perk J. *et al.*, "European Guidelines on cardiovascular disease prevention in clinical practice"; European Heart Journal, Vol. 33, pp. 1635–1701, 2012.

[2] Boye N. *et. al*., "PREVE White Paper – ICT Research Directions in Disease Prevention", FP7 – 248197, 2010.

[3] Reiter, N. *et. al.*, "HeartCycle: Compliance and Effectiveness in HF and CAD Closed-Loop Management", Proceedings of the 31st Annual International Conference of the IEEE EMBS, 2009.

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

[5] Lloyd-Jones, D., "Cardiovascular Risk Prediction: Basic Concepts, Current Status and Future Directions", Circulation, AHA, Vol.121, pp. 1768 – 1777, 2010.

[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, *et. al.*, "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] Friedman N., Geiger D., Goldszmidt M., "Bayesian network classifiers", Machine Learning, Vol.29, 131-163, 1997.

[10] S. Paredes, *et. al.*, "Fusion of Risk Assessment Models with application to Coronary Artery Disease Patients ", 33th Annual International IEEE EMBS Conference, USA, 2011.

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

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

[13] S. Paredes *et. al.*, "Improvement of CVD Risk Assessment Tools' performance through innovative Patients' Grouping Strategies", 34th Annual International IEEE EMBS Conference, USA, 2012.

[14] Twardy C. *et. al.*., McNeil J., "Data Mining cardiovascular Bayesian networks, School of Computer Science Software Engineering, Monash Univ., Melbourne, Tec. report, 2004.