Long Term Cardiovascular Risk Models' Combination - A new approach

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

Abstract - This work addresses two major drawbacks of the current cardiovascular risk score systems: reduced number of risk factors considered by each individual tool and the inability of these tools to deal with incomplete information. To achieve this goal a two phase strategy was followed. In the first phase, a common representation procedure was considered, based on a Naïve-Bayes classifier methodology. Conditional probabilities parameters were initially evaluated through a frequency estimation method and after that optimized using a Genetic Algorithm approach. In a second phase, a combination scheme was proposed exploiting the particular features of Bayes probabilistic reasoning.

This strategy was applied to describe and combine SCORE, ASSIGN and Framingham models. Validation results were obtained based on individual models, assuming their statistical correctness. The achieved results are very promising, showing the potential of the strategy to accomplish the desired goals.

I. INTRODUCTION

In the context of cardiovascular diseases, risk assessment tools are of fundamental importance. In fact, they have a significant impact on the management of an individual patient, mainly supporting professionals in the stratification of patient's risk and in personal care plan definition. In this perspective, risk assessment helps professionals to adapt the personal care plan according to a given specific risk-reduction effort. Additionally, they are valuable tools to reduce lack or over treatment situations as well as tailoring the frequency of clinical follow-up visits [8][2].

The cardiovascular risk, i.e., the probability of occurrence of a cardiovascular event within a certain period of time, is commonly estimated based on risk score models. According to the period of time it is possible to identify two main categories of cardiac risk assessment tools: long term (years) and short term tools (months). Long term tools are widely available, while only a few studies have been conducted considering a short term period (months) [6][14]. Regarding long term risk assessment tools, numerous cardiovascular disease/coronary artery disease are available in literature: Framingham study [6], SCORE [4], Qrisk [5], ASSIGN [19], PROCAM [1], UKPDS [12], Joint British charts [3], New Zealand[10], Sheffield [11]. These risk score systems differ when considering input risk factors, disease (artery disease, heart failure, etc), events prediction (death, myocardial infarction, etc), prevention type (primary/secondary) and patients' specific condition (for example diabetics).

Although useful, these tools present some limitations. In fact, individually, they include relatively few risk factors and they cannot deal with incomplete information (missing risk factor) [2]. Additionally, they are not able to capture the dynamics of the risk evolution, they do not allow the incorporation of clinical knowledge and they are not appropriate to model a specific patient. The main goal of the present work is to develop a methodology that is able to create an adjustable model that can incorporate a higher number of risk factors and cope with incomplete information. This approach intends to take into account available information, try to profit from that by developing a strategy to combine that knowledge, rather than derive a new model.

In literature, several different ways to combine models are referred, basically organized according two main categories: models' output combination (static/dynamic voting, static/dynamic selection …) and models' parameters fusion [17]. The strategy followed in this work is included in the last category, and considers the combination of individual Naïve-Bayes models. In fact, Naïve-Bayesian models can handle with incomplete data sets, show causal relationships and facilitate the use of prior knowledge, which make them appropriate to model the individual risk scores [9,11].

Following a Bayesian modeling approach, a common representation of individual models is, in a first stage, carried out. Then, based on this common description, individual models are combined. As a result, it is possible to consider/integrate distinct inputs from individual tools in a global model and, consequently, to consider a larger number of risk factors which, with a single risk tool, would not be possible. Additionally, since Bayesian models are based on conditional probabilities, they provide an appropriate approach to deal with uncertainty. As a result, it is possible to handle directly with incomplete information (missing risk factors). Bayes model parameters learning involve conditional probabilities estimation. These values were, in a first step, evaluated through a frequency estimation method based on inputs and outputs discretization of individual models. Later, a Genetic algorithm (GA) approach was applied to optimize initial conditional probabilities.

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, Rua Pedro Nunes, 3030-199 Coimbra. {sparedes@isec.pt, teresa@isec.pt}

‡ CISUC, Departamento de Engenharia Informática, Universidade de Coimbra, Pólo II, 3030-290 Coimbra {carvalho@dei.uc.pt, jh@dei.uc.pt}

* Philips Research Europe, Aachen, Germany {matthew.harris@philips.com}

**Serviço de Cardiologia, Hospital Santo André, EPE Rua das Olhalvas, 2410-197 Leiria {joaomorais@hsaleiria.min-saude.pt}

The validation of the proposed combination strategy represents a major challenge, since there is no available global dataset. In order to circumvent this problem, validation was done based on the individual models, which have already been statistically validated. The paper is organized as follows: in section 2 an outline of the methodology is presented. It incorporates the individual models common description and their combination scheme. In section 3 some validation results are presented and, finally, in section 4, some conclusions are drawn.

II. METHODOLOGY

A. Common representation

The first phase addresses a common representation of the individual models. Bayesian networks were employed to model individual behaviors since they are suitable for these particular conditions. In fact, Bayes networks are simple, efficient and present a predictive performance competitive with other classifiers [16]. Furthermore, they can deal with incomplete information, a key aspect in the present work. Figure 1 shows a Naïve-Bayes structure, a particular case of Bayes network models. This structure assumes a very simple and particular configuration, composed of only one output (*C*) and several inputs (*Xi*).

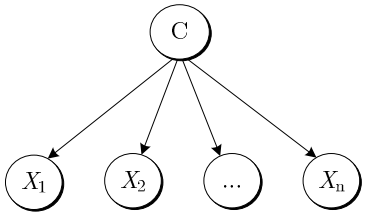

Figure 1 – Naïve-Bayes Structure.

Considering *x* as the value taken by variable *X* and *c* being the class label (mutually exclusive), Naïve-Bayes classifies the instance (*e*) in the class C_l that maximizes the conditional probability $P(C_l|e)$, according to equation (1), where α is a normalization constant.

$$P(C \mid X_1,...,X_n) = \alpha P(C)\prod_{i=1}^{n} P(X_i \mid C) \quad (1)$$

This specific structure assumes that this calculation is only valid if all the inputs X_i are conditionally independent, given the value of class *C* [7][13]. Considering individual cardiovascular risk models, risk factors (inputs) are usually unrelated. This fact validates the application of Naïve-Bayes structure in the present work.

B. Conditional probabilities estimation

Conditional probabilities (CP) of each input *Xi*, given the class *C*, are commonly evaluated from available datasets. When considering individual risk models, the CP parameter learning process was accomplished based on data generated from individual models' equations, accessible in literature [4][6][19]. Using the generated data, CP were directly built based on a frequency estimation method, equation (2).

$$P(X_i = x_j | C = c_k) = \frac{X_i = x_j \wedge C = c_k}{C = c_k} \quad (2)$$

It is important to emphasize that inputs considered in current cardiovascular risk assessment tools are from different natures (continuous/discrete). Consequently, the first operation to be performed was the inputs' discretization. There are several methods to perform that operation [20]. Here the discretization was performed in order to verify the intervals with clinical significance.

C. Combination of models

Combination of models aims to create a global system that can integrate information from different individual models. In this scheme several individual models *Mi* are considered, each one characterized by a specific conditional probability table (CPT), $P(Xi_j|Ci)$, and the individual risk model output, $P(Ci)$. Moreover, some risk factors (model inputs) may be considered by more than one model, while other inputs belong just to a particular model. Given these conditions, models' combination, *i.e*, global risk distribution $P(C)$, is obtained based on individual risk model outputs $P(Ci)$. In order to perform this combination some conditions have to be verified:

i) Individual models have the same number of output levels (classes). This ensures that models share the same risk assessment goal.

ii) Shared variables' CPT $P(Xi_j|Ci)$ present approximately the same *values*. This means that individual models classify the same information (common variables) in an similar way.

The value of $P(C)$ is determined based on a simple frequency calculation, considering the entire single models' outputs. Conditional probability tables of the global risk model, given by $P(X_i|C)$, can also be defined based on a frequency estimation method (1). Nonetheless, for each variable it is necessary to consider the correct dataset. In this manner, a CPT calculation for a variable that is used by more than one model must consider the dataset of those models. On the contrary, if a variable only belongs to one model the CPT table must match the respective individual CPT.

This approach builds a global model that has the same inference method as the individual models (2), and presents a behavior that can reproduce the performance of each one of the base models.

Since Bayes models are based on probabilistic reasoning, this inference mechanism is able to deal with lack of input information (missing risk factor). Additionally, it is possible to make use of all input variables as a whole, which is not possible when individual models are considered.

D. Validation and optimization strategy

As there is no available global data set, it is not possible to adequately validate the proposed strategy. However, it is important to stress that the validity of individual models is guaranteed, since they have been statistically validated (in literature). Currently, it is assumed that the validity of the

global model can be confirmed if it presents the same behavior of each individual model, when only the respective variables are considered.

To assure that global model reproduces individual behaviors in a very precise way Genetic Algorithms were applied to optimize the parameters of the global conditional probability table P(Xi|C), initially estimated based on a frequency estimation method.

III. Results

A. Bayes representation of individual risk tools

ASSIGN, Framingham and SCORE [4][6][19], were the selected models to validate the present methodology. SCORE is a well accepted tool and is applied in clinical practice in some European countries. It calculates a 10 year CV absolute risk for fatal events. Thus, it was included in the model but it will not be combined with the other models since it only considers fatal events.

ASSIGN and Framingham estimate a 10 year CV absolute risk events (non-fatal, death). These two models were combined according to the methodology explained in section II. These three models, characterized in table I, were implemented following the Naïve-Bayes classifier approach.

TABLE I
INDIVIDUAL MODELS CHARACTERIZATION

Models	Risk Factors	Output levels
SCORE (fatal events)	age, sex, tch, sbp, smok, rg	Low/Intermediate; High; Very High
ASSIGN (events)	age,sex, tch, hdl, sbp, diab, famh, cpd, sim	Low/Intermediate; High
Framingham (events)	age, sex, tch, hdl, sbp, smok, diab, bptr	

Inputs:
tch – Total Cholesterol; sbp –Systolic blood pressure; rg – European Region [low risk/high risk]; hdl - High-density lipoproteins; smok – smoking; diab – Diabetes; famh – Family history; cpd – Cigarettes per day; sim – Social deprivation index, bptr – Blood pressure treatment

Outputs:
SCORE categories [0 5 15 100]; Assign/Framingham [0 20 100]

1) Data set

Assuming the same approach proposed by Twardy *et. al.* [15,16] the continuous variables were normally distributed, as given in table II, taking into account the respective mean (μ) and standard deviation (σ).

TABLE II
CONTINUOUS VARIABLES GAUSSIAN DISTRIBUTIONS

Var.	μ	σ
Age	48.5	10.8
Sbp	129.7	17.6
Tch	212.5	39.3
Hdl	44.9	12.2

The values for the discrete variables were generated from models presented in [4][6][19], through a random process. Based on risk factors' values and on the equations/charts/scores available in literature (statistically validated models), the corresponding class was calculated for each instance. Using this approach two data sets were created $(x_{1i},...,x_{ni},c_i)$ for all $1 \le i \le N$: training set N=10000; testing set N=1000.

2) Input discretization

Continuous variables were discretized. As referred in section II, discretization levels were defined according to their clinical relevance, as shown in the table III.

TABLE III
CONTINUOUS VARIABLES' DISCRETIZATION

Var.	Range
age	[0 35 40 45 50 55 60 100]
sbp	[0 120 130 140 160 250]
tch	[0 100 190 250 400]
hdl	[0 35 45 55 400]
cpd	[0 1 15 25 40 60]
sim	[1 15 30 45 60 87]

The remaining inputs are discrete:

TABLE IV
DISCRETE VARIABLES

Var.	Categories
Sex	0/1 : Female /Male
Smok	0/1 : No/Yes
Rg	0/1 : Low risk / High Risk
Diab	0/1 : No/Yes
Bptr	0/1 : No/Yes
Famh	0/1 : No/Yes

3) Conditional Probability Tables (CPT)

Conditional probability tables were obtained based on a frequency estimate method (3). Therefore, for each input, a CPT was created. The number of lines and columns corresponds, respectively, to the number of categories and the number of class labels. Table V presents CPT for HDL - High-density lipoproteins input.

TABLE V
EXAMPLE: INPUT "HDL" CPT

	Low/Intermediate	High
0-35	0.1623	0.3002
35-45	0.2854	0.3115
45-55	0.3150	0.2480
> 55	0.2373	0.1403

4) Models' Performance and optimization

The individual models' performance was validated, based on the capacity to predict the correct class, in comparison with the original models. These results are presented in table VI. Despite the good performance of individual models, a genetic algorithm (GA) approach was, in a second step, applied, to optimize conditional table parameters. Table VI presents initial and optimized performance of individual models.

TABLE VI
INDIVIDUAL MODELS' PERFORMANCE

Models	Before GA %	After GA%
SCORE	91.3	99.6
ASSIGN	90.4	99.24
Framingham	89.6	99.03

B. Models' Combination

Individual models (ASSIGN and Framingham) were combined following the approach described in section II. Therefore, the global model has two outputs: one to calculate the 10 year absolute risk of fatal events (SCORE) and another to assess 10 year absolute risk of fatal/non-fatal events. Table VII shows the performance of the global model, before and after GA optimization process.

TABLE VII
GLOBAL MODELS' PERFORMANCE

Models	Before GA %	After GA%
ASSIGN	88.5	98.93
Framingham	89.0	98.88

These values were obtained taking into consideration the risk factors that belong to each model separately. The capacity to isolate individual models' behavior is a direct consequence of the global model's ability to deal with lack of input risk factors, since it is a straightforward operation to disable the influence of a specific variable. The respective CPT must be set to one in order to disable the influence of a particular variable (risk factor). Then results were compared to the original models.

Thus, it is possible to confirm that the global model behaves like individual models when the respective risk factors are the only information available. This aspect assures that the global model has the right structure and learns the correct parameters' values (CPT definition). Then, as highlighted in section II, the global model was validated exclusively based on statistically validated models.

C. Implementation

All the models as well as the remaining functions were implemented with *Matlab*. A graphical interface was also developed to ease the global model's performance evaluation.

IV. CONCLUSIONS

This work has proposed a strategy to overcome two major drawbacks of the current cardiovascular risk tools: reduced number of risk factors considered by each individual tool and the inability of these tools to deal with incomplete information. Based on a Naïve-Bayes classifier methodology a common representation scheme was implemented for the individual cardiovascular risk tools. Then, using this common description, a combination scheme was proposed exploiting the particular features of probabilistic reasoning. The validity of the proposed strategy was assessed considering the description of three individual models SCORE, ASSIGN and Framingham and the combination of the last two (ASSIGN and Framingham).

Ongoing research is mainly directed to improve the conditional probability tables of the global model and to increase the global model's risk discrimination (higher number of output classes). This last issue depends directly on the availability of suitable validated individual models (clinically relevant).

REFERENCES

[1] Assmann, G., H. Cullen, H. Schulte, "Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the Prospective Cardiovascular Münster (PROCAM)", Circulation, 105, 310—315, 2002.

[2] Bertrand, M. et al. "Management of acute coronary syndromes in patients presenting without persistent ST-segment elevation", European Heart Journal, Vol. 23, 1809–1840, 2002.

[3] British Cardiac Society; "Joint British recommendations on prevention of coronary heart disease in clinical practice". BMJ, Volume 320, 2000.

[4] Conroy R., Pyorala K., Fitzgerald A., et al., "Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project", European Heart Journal Vol. 24, 987-1003, 2003.

[5] Cox, J., Coupland C., Vinogradova Y., Robson J., May M., Brindle P., "Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study", BMJ, 2007.

[6] D'Agostino, R., Vasan R., Pencina M., Wolf A., Cobain M., Massaro J. K. "General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study", American Heart Association, 2008.

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

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

[9] Heckerman, D., Geiger, D. & Chickering, D. "Learning Bayesian networks: The combination of knowledge and statistical data", Machine Learning, 20, 197-243, 1995.

[10] Jackson R., "Updated New Zealand cardiovascular disease risk-benefit prediction guide". BMJ 320: 709-710, 2002.

[11] Quaglini, S., et. al. ; "Cardiovascular risk calculators: understanding differences and realising economic implications"; International Journal of Medical Informatics, Volume 74 , Issue 2 - 4, 191-199, 2005.

[12] Stevens, R., et. al. "The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes", Clinical Science 101: 671–679, 2001.

[13] Su J., Zhang H., "Full Bayesian Networks Classifiers", Proceedings of Machine Learning 2006.

[14] Tang. E, et. al., "Global Registry of Acute Coronary Events(GRACE) hospital discharge risk scores accurately predicts long term mortality post acute coronary syndrome", American Heart Journal, January, 2007.

[15] Twardy C., Nicholson A., Korb K., McNeil J., "Data Mining cardiovascular Bayesian networks", 2003 http://www.datamining.monash.edu.au/bnep

[16] Twardy C., Nicholson A., Korb K., McNeil J., "Knowledge engineering cardiovascular Bayesian networks from the literature" 2005.

[17] Tsymbal, A, et al. "Ensemble feature selection with the simple Bayesian classification", Information Fusion, 87-100, 2003. Available: http://www.datamining.monash.edu.au/bnepi.

[18] Wallis, E., L. Ramsay, I. Haq, P. Ghahramani, P. Jackson, K. Rowland, W. Yeo, "Coronary and cardiovascular risk estimation for primary prevention: validation of a new Sheffield table in the 1995 Scottish health survey population". BMJ 320: 671-676, 2000.

[19] Woodward, M., "Adding social deprivation and family history to cardiovascular risk assessment – The ASSIGN score from the Scottish Heart Health Extended Cohort, 2006.
Available: http://heart.bmj.com/cgi/rapidpdf/hrt.2006.108167v1

[20] Yang Y., Webb G. "A Comparative Study of Discretization Methods for Naive-Bayes Classifiers", Pacific Rim Knowledge Acquisition Workshop (PKAW), 159-173, Tokyo, 2002.