Proceedings of the 3rd World Congress on Electrical Engineering and Computer System and Science (EECSS'16)
Budapest, Hungary – August 16 – 17, 2016
Paper No. ICBES 120
DOI: 10.11159/icbes16.120

Recent Clinical Datasets in Supporting the Clinical Decision: A Portuguese Case Study

Simão Paredes[1], Teresa Rocha[1], Jorge Henriques[2], Paulo Carvalho[2], Diana Mendes[2,] João Morais[3]
[1]Instituto Politécnico de Coimbra, Instituto Superior de Engenharia de Coimbra
CISUC, Universidade de Coimbra
Coimbra, Portugal
sparedes@isec.pt; teresa@isec.pt
[2]Departamento de Engenharia Informática,
CISUC, Universidade de Coimbra
Coimbra, Portugal
jh@dei.uc.pt, carvalho@dei.uc.pt; diana.sxm@gmail.com
[3]Centro Hospitalar de Leiria
joaomorais@chleiria.min-saude.pt

Abstract - The evaluation of a patient's condition is a challenging task that physicians have to deal with in their daily clinical practice, as there are some specific conditions where the diagnosis is not straightforward. Therefore, clinical guidelines frequently recommend the use of models that were developed with the objective of aiding the clinical decision. However, these models have some significant flaws, namely under specific conditions can present a lack of performance. Moreover, large datasets that resulted from patients data gathered directly in the hospital or through telemonitoring systems are available. These datasets may comprise very useful information in order to complement the current clinical knowledge on a specific disease/condition.

The proposed approach addresses this issue, through three different perspectives: i) improving the representation of current clinical knowledge (model enhancement); ii) knowledge discovery strategies, able to extract useful new knowledge from existent clinical datasets; iii) flexible combination schemes that allow the combination of the new knowledge directly extracted from the datasets with the current clinical models.

This work is being developed in the context of cardiovascular disease (CVD), namely in the identification of the CVD risk of each patient as the correct stratification of patients may significantly contribute to the optimization of the health care strategies. A dataset made available by the Portuguese Society of Cardiology (PSC) comprising 11112 patients with acute coronary syndrome gathered between 1st of October 2010 and 6th of November 2014 is applied to support the present work.
Some preliminary results were achieved, showing the potential of the proposed strategy to aid the clinical decision.

Keywords: Clinical Decision-Support Systems; Knowledge discovery, Cardiovascular Risk Assessment.

1. Introduction

The evaluation of a patient's condition is a challenging task that physicians have to deal in their daily clinical practice, as there are some specific conditions where the diagnosis is not straightforward. The growing development of clinical decision support systems (CDS) that provide physicians with clinical knowledge and patient specific information intends to help physicians with this difficult task [1].

In fact, clinical guidelines often recommend the use of models that were developed with the objective of aiding the clinical decision [2]. Moreover, the information of each patient can be gathered (in hospitals or using telemonitoring systems) and stored, creating large and very complete datasets that may contain very useful information. This is relevant in several diseases, such as cardiovascular disease (CVDs), diabetes, Chronic Obstructive Pulmonary Disease (COPD), etc.

Cardiovascular disease (CVD) which includes coronary heart disease (e.g. myocardial infarction), cerebrovascular disease (stroke), heart failure, hypertension, is the world's primary cause of death. According to the World Health Organization (WHO) estimates, the number of people who die from CVD will reach 23.3 million by 2030 [3], which demonstrates the potential relevance of models that can aid the clinical decision.

In the context of CVD, there are several models that allow physicians to evaluate the probability of an individual developing an event based on a set of risk factors [4]. These risk assessment models can be divided in two main categories: long term (years) specific for primary prevention; short term (months) specific for secondary prevention [5, 6, 7]. Moreover, they can be systematized according to the predicted events (hard/soft endpoints); disease (coronary artery disease, heart failure, etc.); risk factors considered in the model (age, gender, etc.) and patient's conditions (ambulatory patients, hospitalized patients, etc.).

These risk assessment tools can be valuable elements to aid physicians in the adaptation of the patient's personal care plan however they present important weaknesses, namely: *i)* under specific conditions (e.g. different populations) they may present some lack of performance; *ii*) incapacity to incorporate new risk factors; *iii*) difficulty to cope with missing risk factors; and *iv*) eventual inability to assure the clinical interpretability of the model [8].

Furthermore, data mining methods can be applied to large clinical data sets in order to discover important correlations between clinical data and patient outcomes.

In this context, this work addresses the support to the clinical decision, through three different perspectives: *i)* improving the representation of current clinical knowledge (model enhancement); *ii)* knowledge discovery strategies, able to extract useful new knowledge from existent clinical datasets; *iii)* flexible combination schemes that allow the combination of the new knowledge directly extracted from the datasets with the current clinical models.

A real patient dataset provided by the Portuguese Society of Cardiology is applied to validate this work. This data set designated by National Registry on Acute Coronary Syndromes, is the largest Portuguese data set specific for acute coronary syndrome patients. It comprises of 11112 patients and corresponds to patients gathered between 2010 and 2014 which also represents an important advantage as it can reflect the most recent evolution in the therapeutic approaches.

The paper is organized as follows: section 2 provides an overview of the developed methodologies, Section 3 presents the main achieved results. Some final considerations are drawn in section 4.

2. Proposed Approach

As mentioned, the proposed approach relies on three different aspects as presented in Fig. 1: *i)* improving the representation of current clinical knowledge (A); ii) knowledge discovery from existent clinical datasets (B); *iii)* combination scheme that allows the combination of the new knowledge with the current clinical models (C).

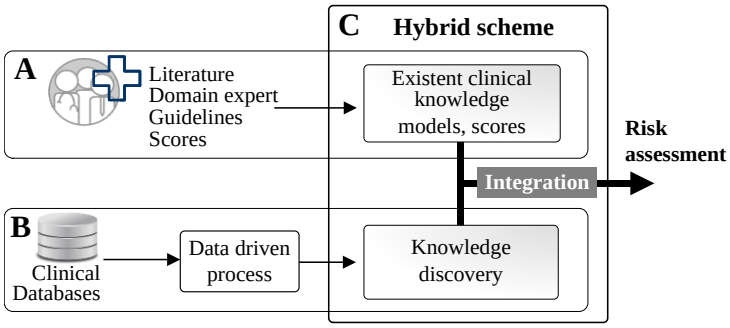

Fig. 1: Proposed Approach.

2.1. Representation of Current Clinical Knowledge (A)

This strategy relies on two main steps: i) representation of a score (currently applied in the clinical practice) based on a Decision Tree (DT); *ii)* adjustment/optimization of the DT thresholds, while keeping the structure unchanged (rules and risk factors. i.e., inputs).

The former is based on the representation of an individual model (usually represented as a score, *e.g*. GRACE score [5]) based on a decision tree (DT). This representation seems very suitable, as DT are based on rules that highlight the relations among the variables creating dependencies among them.

The latter intends to adjust the DT thresholds in order to improve the predictive performance of the model. Here, it is important to highlight that the optimization is restricted to the DT thresholds while the structure of the DT remains unaffected. This optimization was performed assuming two different methods: *i)* an innovative approach based on

membership functions; *ii)* genetic algorithms (used to compare their results with the results obtained with the membership functions).

Approach based on membership Functions

The algorithm was developed taking into account two main functions: the error function (EF) and the membership function (MF). EF corresponds to the module of the difference between the real output (Y_{real}) and the estimated output ($Y_{estimated}$), which indicates if the classification (risk class) is correct or not correct. The MF computes the degree of membership of each patient in relation to each rule. As each rule is composed by a set of thresholds (Ti), the MF is computed in each of them, being the final value equal to the respective average.

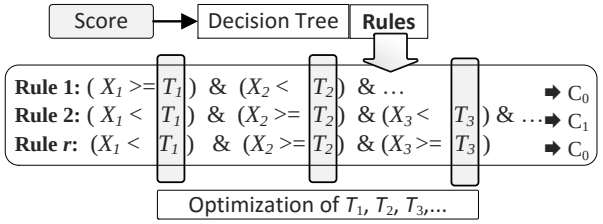

Fig. 2: New approach (Each Rule: Xi inputs, Ti thresholds, Ci the class).

The membership function (MF) takes into account if the variables are binary (0 - doesn't belong to the rule; 1 – belongs to the rule) or continuous (calculated based on a modified sigmoid function).

All the rules followed by the patients lead to an output. If the output is different from the real output, an optimization of the Ti is performed in order to reduce the EF.

This specific algorithm is in an ongoing development process. As presented in the following section, some of the preliminary results are very promising.

Genetic Algorithm Operation

The genetic algorithms (GA) are one of the most used methods in optimization problems [9]. In this work the initial population (P0) is composed by the initial values of thresholds of each node of the tree and the fitness function is defined through the geometric mean ($F = 1 - \sqrt{(SE \times SP)}$), where SE is sensitivity and SP specificity.

2.2 Knowledge Discovery from Clinical Datasets (B)

The availability of recent clinical datasets allows the extraction of knowledge from the data through the discovery of significant relationships between clinical data and patient outcomes.

Here the proposed methodology, can be systematized in four main steps: *i)* dealing with missing information (missing risk factors), where some techniques (e.g. mean substitution, multiple imputation, etc.) were applied; *ii)* dealing with imbalanced datasets due to the reduced event rate. In fact, an imbalanced dataset originates additional challenges in a classification problem as the results are biased to the majority class; iii) Dimensionality reduction strategy in order to avoid the problems created by a high dimensional space. A suitable feature selection process may allow the elimination of irrelevant information, minimize the over-fitting situations, improve performance as well as the model interpretability; iv) test of different classifiers such as Decision Trees (DT); Random Forest (RF) or k-nearest neighbour (kNN).

As presented in the results section, the obtained results are also very promising. This approach is detailed in a recent paper of this research team [10].

2.3 Combination Scheme (C)

This combination scheme aims to integrate the information from modules A and B, so it must be flexible to incorporate this diversity of information. The development of this scheme is an ongoing process that intends to improve a fusion approach that was previously developed by this research team and is detailed in [8].

This strategy aimed to combine CVD risk assessment tools and it was based on two main hypotheses: *i)* it is possible to create a common representation to the individual CVD risk assessment models; *ii)* it is possible to combine in a common framework the resulting individual models.

The first step of this methodology was to represent the selected CVD risk assessment tools using a Naïve Bayes classifier as it presents some characteristics that are particularly suitable for this representation.

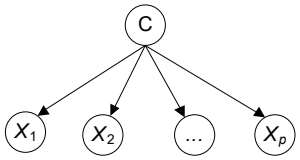

Fig. 3: Naïve Bayes Structure.

Where X is an observation (e.g., a set of risk factors), being $X_i\ i = 1, \dots, p$ the i[th] risk factor, and C a hypothesis (e.g., CVD risk level). It relies on the Bayes rule:

$$P(C|X) = \frac{P(X|C).P(C)}{P(X)} \quad (1)$$

The term $P(C|X)$ denotes a posterior probability, i.e., the probability of the hypothesis C after having seen the observation X. $P(C)$ is the prior belief, the probability of the hypothesis before seeing any observation (prevalence of the CVD risk level). $P(X|C)$ is a likelihood, the probability of the observation if the hypothesis is true (sensitivity of the clinical exam).

The goal is to represent the behaviour of a CVD risk assessment model, so the new model must learn the parameters $P(X|C), P(C), P(X)$ that allow the determination of $P(C|X)$. The parameters of an individual model are learned based on a training dataset that is applied to the correspondent CVD risk assessment tool.

The second step of the proposed methodology is the combination of individual Bayesian models, where a global model is directly created from the fusion of the individual models exploring the particular features of Bayesian inference mechanism.

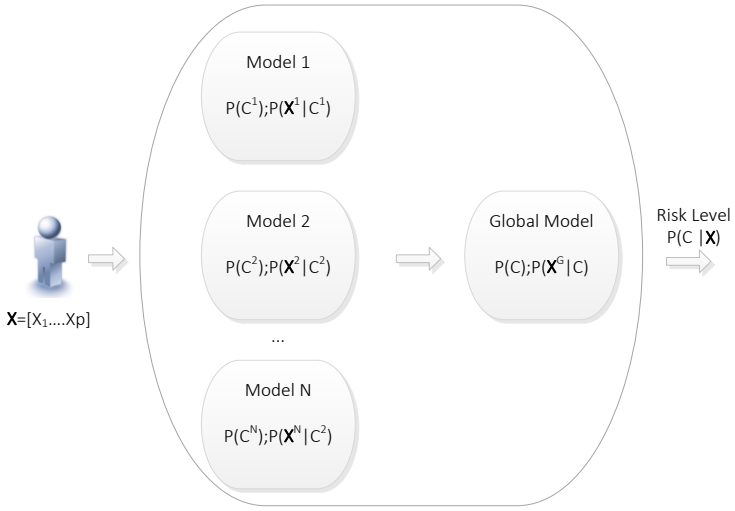

Fig. 4: Combination Scheme.

Each model i is characterized by the respective prior probability of output class $P(C^i)$ and its conditional probability table composed of $P(\boldsymbol{X}^i|C^i)$ where $\boldsymbol{X}^i$is the set of risk factors (inputs) considered by the model i.

The combination scheme implements the direct combination of the individual models' parameters, where $P(C); P(\boldsymbol{X}^G|C)$ are obtained based on the several $P(C^i); P(\boldsymbol{X}^i|C)$, through a weighted average combination scheme. It is important to emphasize, that a new model, based on the prevalence of a specific risk factor and on the risk associated with each one of its categories, can be directly created by the physician and easily incorporated in the combination scheme. Afterwards, an optimization procedure, based on genetic algorithm operation, is applied to the parameters of the global model, with the goal of improving its predictive performance (maximize simultaneously the sensitivity and specificity). A more detailed explanation can be found on [8].

3. Results

3.1 Portuguese Society of Cardiology Dataset

A dataset from the PSC was considered for the validation procedure of the developed algorithms. The information was collected in the context of the second stage of the national registers of acute coronary syndrome, gathered between 1st of October 2010 and 6th of November 2014. The dataset comprises 11112 patients of which 394 patients died, which represents an event rate of only (3.5%). Thus, the dataset is severely imbalanced which imposed some strategies (e.g. under and over-sampling) in order to overcome this additional difficulty.

The variables can be grouped in several categories: i) demographics; ii) admission diagnosis; iii) clinical history; iv) symptoms of acute episodes; v) ECG characterization; vi) medication; vii) biomarkers. Table 1 presents the some of the most relevant characteristics of the considered dataset.

Table 1: PSC data set characterization.

Model	Events
Gender (Male/Female)	71.96%/28.04%
Smoking (1/0/undefined)	27.61%/71.66%/0.73%
Pain on admission > 20 minutes(1/0/undefined)	76.20%/8.79%/15.01%
Cardiac arrest(1/0/undefined)	1.91%/96.09%/2%
Atrial fibrillation(1/0/undefined)	5.30%/94.46%/0.24%
Elevation of ST segment(1/0/undefined)	40.86%/53.76%/5.38%
Unstable angina (1/0/undefined)	7.60%/92.4%/0%
Acute myocardial infarction without ST – segment elevation (1/0/undefined)	47.62%/52.38%/0%
Left ventricular function (1/2/3/4/undefined)	57.22%/ 18.76%/ 12.60%/ 4.53%/6.89%
Killip class(1/2/3/4/undefined)	83.80%/ 9.77%/ 4.08%/ 1.91% /0.44%

3.2 Representation of Current Clinical Knowledge

A ten-fold cross validation was applied in order to validate this approach. The performance of the original score (GRACE model) was compared with the optimized Decision Tree (threshold optimization through membership functions and genetic algorithm operation). The main results are presented on
Table 2.

Table 2: Comparison of the Models' Performance.

Model	SE (%)	SP (%)	G_{mean} (%)
Score (GRACE)	91.2	50.7	68.0
Decision Tree MF	80.5	74.2	77.3
Decision Tree GA	81.4	74.6	77.9

Based on these preliminary results, it is possible to conclude that the DT can improve the performance of the original score, namely to improve the specificity as well as the geometric mean that provides information about the balance of the classifier. Although, in both cases (membership function and genetic algorithms) the sensitivity decreased when compared with GRACE.

3.3 Knowledge Discovery from Clinical Datasets

This algorithm was applied to PSC data set considering in a first phase only binary variables (the categorical features (e.g. Killip class) were transformed in binary variables). Several data space dimensions were assessed. Table 3 presents the

initial data space where 78 binary variables were considered (continuous variables were excluded) as well as a reduced data space with only 6 variables (smoking, cardiac arrest (on admission), complete right bundle branch block, left ventricular ejection fraction (normal); left ventricular ejection fraction (very depressed); Killip class = 1).

Table 3: Comparison of the classifiers performance with different data space dimensions.

Classifier	Model	SE (%)	SP (%)	G_{mean} (%)
Decision Tree (DT)	Subset 1 78 variables	74,5%	75,5%	75,0%
Random Forest (RF)		81,80%	79,90%	80,8%
kNN		68,3%	82,9%	75,2%
Decision Tree (DT)	Subset 2 6 variables	81,9%	77%	79,2%
Random Forest (RF)		80,00%	77%	78,5%
kNN		78,0%	77,3%	77,7%

The first conclusion that it is possible to derive from these results is that the subset created (after the feature selection process) present similar performances, when compared with the initial set of features, which shows that it is possible to obtain a significant predictive with a reduced data space (6 features) even when only binary variables are considered.
In relation to the classifiers performance, it is possible to conclude that the DT presents the best performance when compared with RF and kNN.

4. Conclusions

This work addresses clinical decision support according to three different perspectives: *i)* improving the representation of current clinical knowledge; *ii)* knowledge discovery; *iii)* combination scheme that allows the combination of the new knowledge with the existing clinical models. All of these approaches rely on the existence of large and recent clinical datasets that can boost the development of those algorithms.

This paper presented the particular case of the cardiovascular disease where a large dataset comprising of 11112 Portuguese patients was applied. The preliminary results are very encouraging which suggest that this integrated strategy can be potentially very useful in aiding the clinical decision in the daily clinical practice.

This work is in an ongoing process, namely to improve the developed algorithms and the respective validation. Recently, this research team obtained a large dataset from Brazilian patients that is being explored.

Acknowledgements

The authors would like to thank the PSC for their collaboration, in particular in providing the clinical dataset.

References

[1] Eichner D. et al., "Challenges and Barriers to Clinical Decision Support (CDS) Design and Implementation Experienced in the Agency for Healthcare Research and Quality CDS Demonstrations." (Prepared for the AHRQ National Resource Center for Health Information Technology under Contract No. 290-04-0016.) AHRQ Publication No. 10-0064-EF. Rockville, MD: Agency for Healthcare Research and Quality, 2010.[2] Perk J. et al., "European guidelines on cardiovascular disease prevention in clinical practice" (version2012). *European Heart Journal* vol. 33, no.13, pp. 1635-701, 2012

[3] World Health Organization. (2013).Cardiovascular Diseases (CVDs), Fact sheet n°317. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs317/en/index.html. [Accessed in December 2015]

[4] Ricotta J et. al., "Cardiovascular disease management: the need for better diagnostics." *Medical & Biological Engineering & Computing*, vol. 46, pp. 1059-1068, 2008.

[5] Tang, E. W. 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*, vol. 153, no. 1, pp. 30-35, 2007.

[6] Antman, E. et al.. "The TIMI risk score for Unstable Angina / Non-St Elevation MI – A method for Prognostication and Therapeutic Decision Making." *Journal of American Medical Association- JAMA*, 284, 835-842, 2000.

[7] 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, American Heart Association - AHA*, vol. 101, pp. 2557-2657, 2000.

[8] Paredes, S. et al. , "Integration of Different Risk Assessment Tools to Improve Stratification of Patients with Coronary Artery Disease" *Computer in Medical & Biological Engineering & Computing*, (MBEC) doi: 10.1007/s11517-015-1342-3, 2015.

[9] Eiben, A. *Introduction to Evolutionary Computing*. ISBN: 978-3540401841, Springer, 2003.

[10] Mendes D. et. al., "Assessment of Cardiovascular Risk based on a Data-driven Knowledge Discovery Approach" in *37th Annual International IEEE EMBS Conference*, Italy, 2015.