Cardiovascular Attributable Risk and Risk Factors Evaluations as a Matter of Statistics and Data Mining Confluences

Dan-Andrei SITAR-TAUT[1], Adela SITAR-TAUT[2]
[1]Department of Business Information Systems, Faculty of Economics and Business Administration, Babeş-Bolyai University, Cluj-Napoca, Romania
[2]Department of Cardiology, Clinical Rehabilitation Hospital, University of Medicine and Pharmacy, Cluj-Napoca, Romania
dan.sitar@econ.ubbcluj.ro, adela.sitar@umfcluj.ro

Cardiovascular diseases represent a severe threat for humanity, being the first cause of death and hospitalization in both genders. An impressive number of studies have been developed in order to identify a set of factors causing this kind of illness, but only few of them were able to pay significant resources in analyzing large population samples (tens of thousands) and for longer periods of time (decades). This paper's objective is to continue the previous researches of the eProCord project and to validate with concrete data the theoretical model developed for the attributable risk (AR). It will consider the same risk factors for myocardial infarction identified by INTERHEART study and the same work hypothesis. We will also evaluate if a certain value of the AR is also confirmed by the invoked disease of the patient. Using statistical and data mining tools we will investigate the prediction potential of the chosen factors and the opportunity to extend them in order to capture any cardiovascular disease. The empirical tests rely for now on a sample of 236 patients.
Keywords: *Cardiovascular Disease, Myocardial Infarction, Attributable Risk, Roc, Data Mining, Classification*

1 Introduction

Cardiovascular diseases represent, for the moment, the number one cause in terms of morbidity and mortality, in both genders [1], [9]. CVD is ubiquitous. The situation is not different for Romania, WHO 2009 statistics showing that cardiovascular mortality is increasing and exceeds other causes (like cancer and injuries) [18].

Coronary artery disease (CAD) prevention has moved beyond the secondary prevention of CAD events, first place being taken by early identification and treatment of individuals thought to be at risk [13]. Unfortunately nonfatal myocardial infarction (MI) occurs without prior recognized symptoms in approximately one quarter of the cases [13]. Global risk assessment and preventive measures are now recommended as standard practice in cardiovascular disease prevention, even in high risk asymptomatic individuals [13], [15].

Risk assessment can be performed using one of the several risk tools - based upon multivariate risk prediction equations derived from large prospective cohort studies or randomized trials -, that combine values for different risk factors into a global risk estimate. [15]. The Framingham Heart Study established the independent impact of cigarette smoking, high blood pressure, high total cholesterol and LDL cholesterol, low HDL cholesterol, diabetes, male sex and advancing age on the development of CVD [9]. Other tools such as the Prospective Cardiovascular Munster Heart Study (PROCAM), the Systematic Coronary Risk Evaluation system (SCORE), United Kingdom Prospective Diabetes Study (UKPDS) tool for diabetics, and the Reynolds Risk Score have been developed [1]. But, most of them even validated in many different populations and ethnic groups and appropriately recalibrated [1] are difficult to be applied in distinct ethnic populations, and recalibration is often challenging and its applicability may be limited [1]. In the same time, the advent of other markers of CV disease, including high sensitivity C-reactive protein (HS-CRP), homocysteine, adhesion molecules, lipoprotein (a), can add

prognostic value to standard risk formulae, can make recalibration process easier, their incorporation into present clinical practice being challenging.

Endothelial dysfunction is frequently discussed as a potential major mechanistic contributor to atherothrombosis. Noninvasive techniques for assessing vascular wall status or cardiovascular function are useful in some of these individuals because they will enable a more accurate assessment of risk and thereby result in the risk status of the patient being raised to "high" [13].

Receiver Operating Characteristic (ROC) curves illustrates the ability of a diagnostic test to discriminate, in this case between "myocardial infarction" and "non myocardial infarction" patients. They plot the relationship between sensitivity and one minus specificity at a range of cut off test values.

Data mining (DM), a step of the Knowledge Discovery from Databases (KDD) process [7] is designed to find unpredictable patterns, associations or relationships between data by using various analytical techniques [4] and mainly applied on large datasets. DM can be also considered a nontrivial extraction of implicit, previously unknown, and potentially useful information from data [3], [11], [14].

Classification is one of the fundamental techniques in data mining relying on unsupervised learning. In this type of problem, the goal is to learn a classifier from a given set of instances with class values to assign correctly a class value to a test instance[6], [16]. Classification may embrace different learning types, based on decision tables and trees, neural networks, instance-based one, but this aspect does not represent the current paper's goal. In a certain case, a classifier performance may be given by the higher number of correctly classified instances from the total instances number, but this status has not be limited just to this aspect. There are other criteria, like precision, recall, different types of errors, Cohen's coefficient, etc. that can characterize a classifier from different angles of view. We will consider just the C4.5 algorithm (J58 in Weka) [16] the most popular classifier, and probably the best performing one [5]. It is also used in various researches in medicine related fields [2] [10].

In a classification problem, not all attributes contribute in the same manner to the classification's success. We have the same in medicine, where a potential factor – smoking, obesity, diabetes, age, etc. – can produce different impacts in the onset of cardiovascular diseases. There are many criteria able to rank the attributes. One is the information gain. The entropy measures the amount of information. Information gain represents the difference between the entropy before and after testing the attribute value [8]. The importance of risk factor identification can be applicable in disease prevention, but the figures obtained or the position in a ranking must not be generalized in disease evaluation.

Methods: The following tools were used: MS Access for data input, processing, and query; MS Excel for data pre-processing and export; MedCalc 10.3.0.0 and SPSS 17.0 (Demo Version) – for statistical processing, ROC creation and for identifying the cut-off values; Weka for additional pre-processing, classification, and attribute evaluation purposes.

Table 1. Abbreviations list

Risk Factor Abbreviation	***Description***
Curr_Smoking	Current smoking status
Diabetes	History of diabetes mellitus
HBV	High Blood Pressure presence
AB_Ob	Abdominal obesity presence
NO_Fruits	NO fruit consumption
NO_Exercises	NO physical exercises
NO_Alc_Small	NO alcohol use in small quantities
AR	Attributable risk
MI	Myocardial infarction presence
CVD	Cardiovascular disease presence

2 Statistical Analysis

INTERHEART was a large, standardized, international, case-control study, designed in order to assess the importance of risk factors for coronary heart disease worldwide [17]. The study included about 15,000 cases (patients with MI history) compared with a similar number of controls (without MI) from 52 countries, being investigated 27,098 subjects. The INTERHEART STUDY has shown that nine easily measured risk factors (smoking, lipids, hypertension, diabetes, abdominal obesity, inappropriate diet, physical inactivity, alcohol consumption, and psychosocial factors) were associated with more than 90% increase in the risk of an acute myocardial infarction in this large global case-control study, results being consistent across all the geographic regions and ethnic groups of the world [17].

eProCord is a Romanian research program that investigates the cardiovascular diseases and intends to quantify their associated risks by considering not only the classical risk factors, but also some of new ones. Researches take place in an interdisciplinary environment. Further details can be found in [12].

Table 2. Risk factors ampleness. Comparison between studies

			Percentages (%)							
Research Project	N		1 Curr Smoking	2 Diabetes	3 HBP	4 AB OB	5 NO Fruits	6 NO Exercises	7 NO Alc Small	MI
IHTERHEART*	27.098	N	64.8	87.4	70.2	60.7	38.8	16.6	24.2	54
		Y	35.2	12.6	29.8	39.3	61.2	83.4	75.8	46
eProCord	236	N	80.9	75.8	29.7	11	71.2	30.5	53.8	93.6
		Y	19.1	24.2	70.3	89	28.8	69.5	46.2	6.4
* The percentages are calculated without considering the individual loss factors, by reporting the declared figures to the whole number of valid cases										

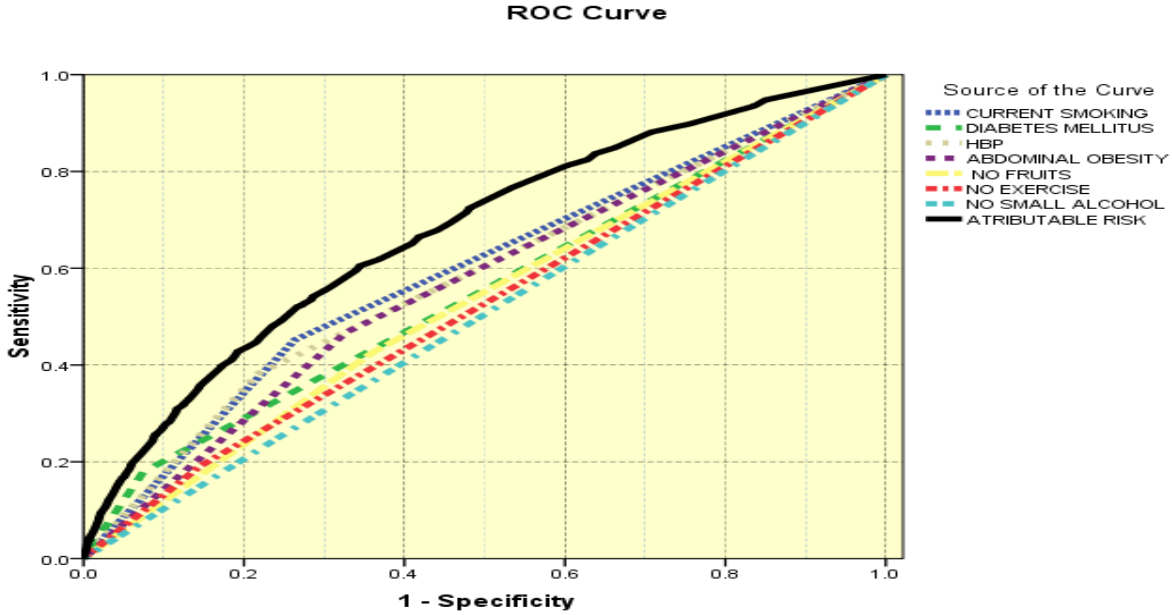

Fig. 1. Comparative theoretical AUCs[1]

[1] Revised from [12] by including 2 additional risk factors. This is the theoretical approach on 30,000 subjects in INTERHEART study conditions.

The two studies are comparable just in the mentioned risk factors consideration, but not also in their proportions. As Table 2 depicted, the INTERHEART study consists in evaluation of a large number of subjects having in general more cardiovascular risks, which are concretized also in a higher MI presence. Only 6.4% of eProCord patients are diagnosed with this disease. Only the factors 3, 4, and 5 surpass the figures from the other project.

The previous researches gave promising assertions regarding de AR for MI presence in patients. The Area under the Curve (AUC) for AR – as we theoretically computed on a pseudorandom INTERHEART look like population, as a product of each AR MI-related factor – was the best in comparison with any other individual risk factor. It is presented in Figure 1.

According to real data obtained in our investigations during eProCord program, the obtained results are synthesized in Figure 2 and Table 3 – eProCord AUCs.

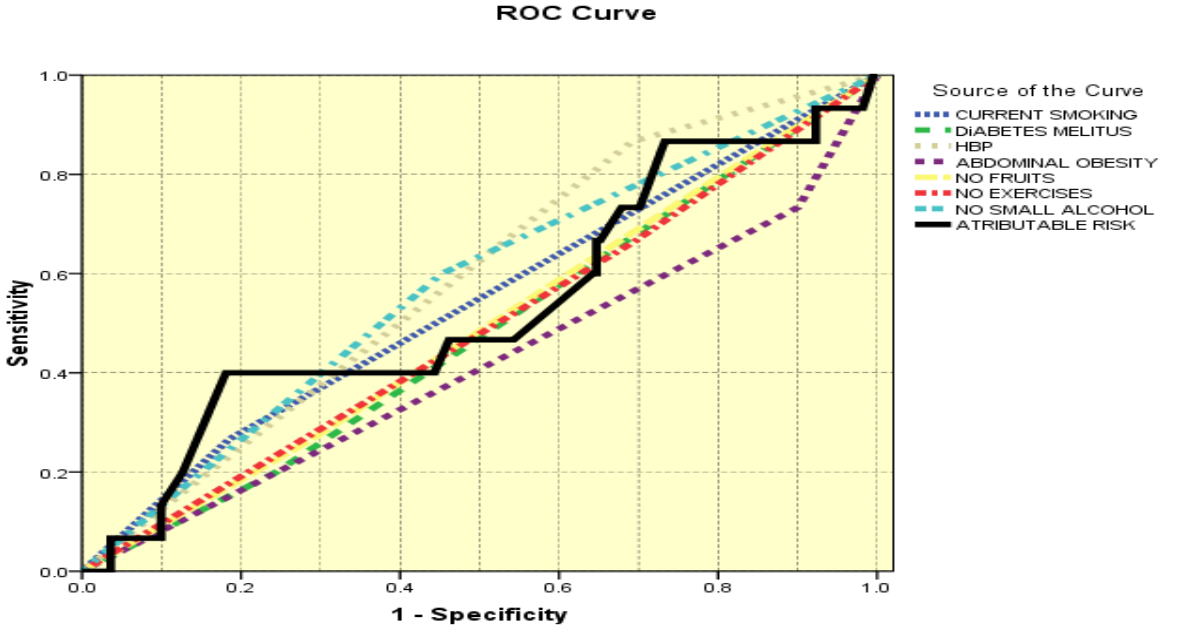

Fig. 2. eProCord ROC curves in 2010

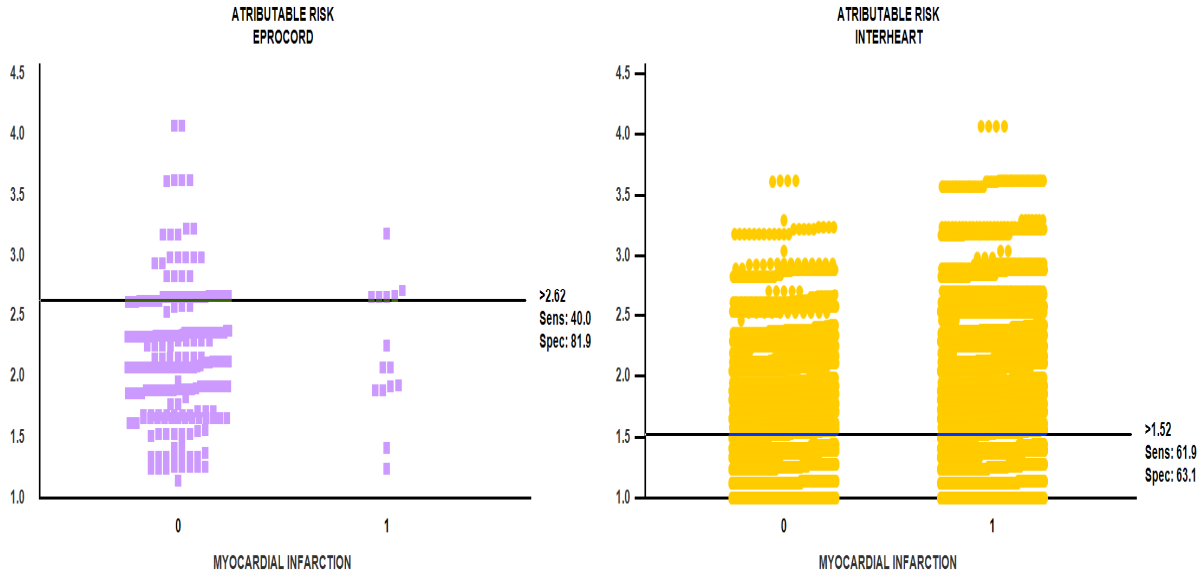

Fig. 3. Comparative AR and cut-off values eProCord vs. INTERHEART

A cut-off value of 2.6159 provides the best combination possible between sensibility and specificity, 40% and 81.9%.
A distribution of the AR values for subjects with MI absence (coded with 0) or presence (coded with 1) in the two studies is revealed in Figure 3. The cut-off values are also emphasized. The AUC value of AR is the eighth one, being the last one from all considered variables.

Table 3. eProCord AUCs

Rank#		AUC	STANDARD ERROR	95% CONFIDENCE INTERVAL
1	ABDOMINAL OBESITY	0.584	0.0721	0.518 to 0.647
2	DIABETES MELLITUS	0.522	0.0761	0.456 to 0.587
3	HYPERTENION	0.587	0.0795	0.521 to 0.651
4	NO EXERCISE	0.515	0.0765	0.449 to 0.580
5	CURRENT SMOKING	0.541	0.0787	0.475 to 0.605
6	NO FRUITS	0.511	0.0766	0.446 to 0.577
7	NO SMALL ALCOHOL	0.574	0.0794	0.508 to 0.638
8	ATTRIBUTABLE RISK	0.533	0.0784	0.467 to 0.598

As it can be noticed from the reported data, the AR cannot predict as good as we hoped the presence of MI. The eProCord's AUC was lower than the theoretical one computed according to INTERHEART data. This indicates that cardiovascular risk factors and the risk associated with them represent an uncompleted solved equation, new risk factors requiring to be evaluated. In the same time, the cut-off value was greater than the predicted one (2.6159 > 1.52 [12]), determining a lower sensibility, but a better specificity. Thus, the healthy subjects are identified very accurately from the point of view of MI. If we established a cut-off value equal to 1.52 (as in the theoretical model) the sensibility is 86.7%, but with a very low specificity (9%).
It also has to be mentioned the fact that our model was probably not completely developed. In addition, we have to emphasize the fact that the INTERHEART study was made using some anthropometric measurements possible not fully applicable in our region. On the other hand, there are studies considering the same risk factors but with minor discrepancies in their interpretation (e.g. various stages of obesity, smoking status).
No viable regression models could be revealed.

3 Data Mining Analysis

We will use the data mining tools in order to confirm or infirm some of the results obtained, or to discover new ones.
Using the attribute evaluation by considering the information gain as criterion, the ranking from Table 4 is obtained.
Table 4 indicates that the two studies did not give the same importance to the risk factors in MI cases identification. INTERHEART considers the current smoking being the most important, since eProCord puts more value in abdominal obesity. The less important risk factor for MI is considered the diabetes versus the lack of fruits consumption in eProCord. On the other hand, the information gain values are very small, meaning the current ranking is not very solid and a higher number of instances may alter the existing order.

Table 4. Attribute ranking comparison

Risk Factor	INTERHEART	eProCord	
	Ranking	Ranking	InfoGain
Curr_Smoking	1	4	0.001689
AB_Ob	2	1	0.009409
NO_Exercises	3	6	0.000181
HBP	4	2	0.007208
NO_Alc_Small	5	3	0.003754
NO_Fruits	6	7	0.000112
Diabetes	7	5	0.000479

We tested if the AR value confirms the presence or absence of the MI, by J48 classifier. The results indicate that (probably) a lower value of AR indicates the absence of MI in 221 cases (93.6441 %), but none of the rest (15, 6.3559 %) is captured as having MI. Thus, even if the classifier is able to identify all non-MI patients, it cannot predict any ill patient. This is what was statistically demonstrated by higher specificities. There was no evidence about the possibility to identify this illness.

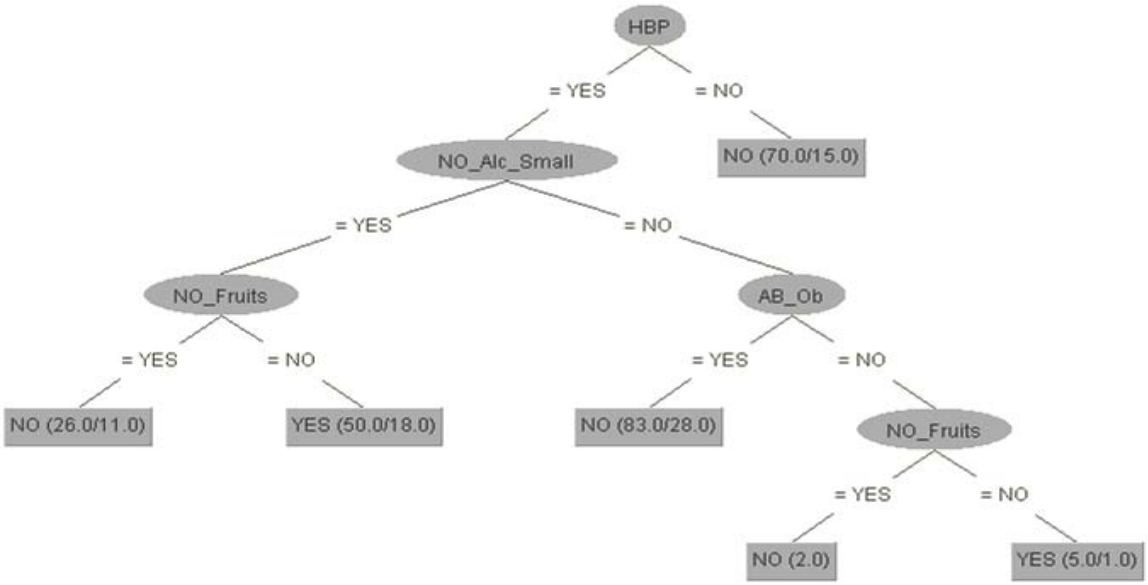

Fig. 4. CVD classification by risk factors

Cohen's k coefficient and classification errors are very bad. But eProCord study is not strictly oriented on MI; also it is not exclusively limited to these factors and their interpretation. What if this AR is able to predict any CV disease? The results are similar, because now all 146 (61.8644 %) without CVD are identified, but none having this is captured (90, 38.1356 %).

The theoretical AR score developed according to INTERHEART's conditions was not validated for now by statistical and data mining tools. In the following section we will test if the risk factors are able to determine more accurate if a patient does or does not have MI, independently by the AR. The results: 221 (93.6441%) correctly classified instances and 15 (6.3559 %) are absolutely identical with those obtained between AR and MI. This means that in MI risk evaluation, the AR score that we calculated can replace successfully the presence of all 6 risk factors, but unfortunately it is not able to identify correctly the presence of MI (just its absence). Is the same hypothesis true for any of the cardiovascular diseases? The resulted decision tree is depicted in Figure 4.

Using the same classifier, the elected set of risk factors correctly classifies 150 instances (150 subjects, 63.5593 %), and incorrectly 86 patients (36.4407 %). For the first time our classifier identifies 29 subjects having at least one CVD, but 63.5593% is a moderate accuracy. The attributes have the following ranking HBP, Diabetes, NO_Alc_Small, NO_Fruits, Curr_Smoking, Ab_Ob, and NO_Exercises.

No regression models can be also revealed using Weka.

4 Conclusions

The classical factors are not enough to capture accurately the cardiovascular disease in general, or one of its particular types. On the other hand, even if the main risk factors are still clear, there are still controversies regarding some methodological or additional aspects characterizing them, like "current",

"old", "absolute", "in the last five years", "less", "minor", "important", their presence in different stages, or regional applicability. New and more precise factors must be taken into account. Statistical and data mining tools may be convergent or complementary. The results provided are more effective in large datasets. The small number of subjects and mainly a smaller number for those having MI cannot contribute to solid conclusions. Comorbidities must be also considered. Our project intends further investigations by extending the number of subjects, variables – i.e. considering the endothelial dysfunction in its various aspects – combining the biometrical measurements with the patients' life style and general or more related analysis or explorations.

Acknowledgement
This paper was supported by Research Project No. 947, ID_2246/2009 Code, and part of PN II Program financed by the Romanian Ministry of Education, Research and Innovation – The National University Research Council.

References

[1] J.A. Batsis and F. Lopez-Jimenez, "Cardiovascular risk assessment-from individual risk prediction to estimation of global risk and change in risk in the population," *BMC Med*, vol. 8, pp. 29-34, 2010.

[2] G. Dolce, M. Quintieri, S. Serra, V. Lagani and L. Pignolo, "Clinical signs and early prognosis in vegetative state: A decisional tree, data-mining study," *Brain Injury*, vol. 22(7), pp. 617-623, 2008.

[3] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17, pp. 37-54, 1996.

[4] W.J. Frawley, P.G. Shapiro, and C.J. Matheus, "Knowledge discovery in databases-an overview," *AI Magazine*, vol. 13, pp. 57-70, 1992.

[5] R. Hewett, J. Leuchner, S.D. Mooney and T.E. Klein, "Analysis of mutations in the colia1 gene with second-order rule induction," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 17(5), pp. 721-740, 2003.

[6] L. Jiang, C. Li and Z. Cai, "Decision tree with better class probability estimation," *International Journal of Pattern Recognition & Artificial Intelligence*, vol. 23(4), pp. 745-763, 2009.

[7] I.N. Lee, S.C. Liao and M. Embrechts. "Data mining techniques applied to medical information," *Med inform*, vol. 25(2), pp. 81-102, 2000.

[8] S. Ohta, R. Kurebayashi and K. Kobayashi, "Minimizing False Positives of a Decision Tree Classifier for Intrusion Detection on the Internet," *Journal of Network & Systems Management*, vol. 16(4), pp. 399-419, 2008.

[9] L. Pilote, K. Dasgupta, V. Guru, K. Humphries, J. McGrath et al., "A comprehensive view of sexspecific issues related to cardiovascular disease," *CMAJ,* vol. 176(6), pp. 1-44, 2007.

[10] V. Podgorelec, P. Kokol, B. Stiglic and I. Rozman, "Decision Trees: An Overview and Their Use in Medicine," *Journal of Medical Systems,* vol. 26(5), pp. 445-462, 2002.

[11] R. Sabzevari and G.A. Montazer, "An Intelligent Data Mining Approach Using Neuro-Rough Hybridization to Discover Hidden Knowledge from Information Systems," *Journal of Information Science and Engineering*, vol. 24, pp. 1111-1126, 2008.

[12] D.A. Sitar-Tăut, A.V. Sitar-Tăut and L. Mocean, "Research about Implementing E-Procord - New Medical and Modeling Approaches in IT&C Age Applied on Cardiovascular Profile Evaluation at Molecular Level," *JAQM*, vol. 4(2), pp. 175-189, 2009.

[13] C. Sidney and M.D. Smith, "Current and future directions of cardiovascular risk prediction," *American Journal of Cardiology*, vol. 97(2), pp. 28-32, 2006.

[14] S. Ting, C. Shum, S. Kwok, A. Tsang and W. Lee, "Data Mining in Biomedicine: Current Applications and Further Directions for Research," *Journal of*

Software Engineering & Applications, vol. 2(3), pp. 150-159, 2009.

[15] M.D. Whitfield, M. Gillett, M. Holmes and E. Ogden, "Predicting the impact of population level risk reduction in cardiovascular disease and stroke on acute hospital admission rates over a 5 year period-a pilot study," *Public Health*, vol. 120(12), pp. 1140-1148, 2006.

[16] I.H. Witten and E. Frank*, Data Mining: Practical Machine learning tools and techniques, 2nd Edition*. San Francisco: Morgan Kaufmann, 2005.

[17] S. Yusuf, S. Hawken, S. Ôunpuu, T. Dans, A. Avezum, F. Lanas, M. McQueen, A. Budaj, P. Pais, J. Varigos and L. Lisheng, "On behalf of the INTERHEART Study Investigators Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study," *Lancet*, vol. 364, pp. 937-952, 2004.

[18] World Health Organization, *World Health Statistics 2009*. France: World Health Organization, 2009.

Dan-Andrei SITAR-TĂUT has a Bachelor's degree in Business Information Systems from Faculty of Economics and Business Administration, Babeş-Bolyai University of Cluj-Napoca and a Master's degree in Informatics Strategies Applied in Economy and Business from the same educational institution. He also holds a PhD diploma in Cybernetics and Economic Statistics. He is the author of 2 books and more than 35 papers in the field of Databases, ERP, Data mining, and Web related fields.

Adela-Viviana SITAR-TĂUT has a Bachelor's degree in Business Information Systems from Faculty of Economics and Business Administration, Babeş-Bolyai University of Cluj-Napoca and in General Medicine from "Iuliu Haţieganu" Medicine and Pharmacy University from the same city. Currently she is an internal medicine specialist physician and has a PhD in Medicine. She participated and published papers to many national and international events on Medicine, Business Information Systems, and Bioinformatics related topics.