RESEARCH ABOUT IMPLEMENTING E-PROCORD – NEW MEDICAL AND MODELING APPROACHES IN IT&C AGE APPLIED ON CARDIOVASCULAR PROFILE EVALUATION AT MOLECULAR LEVEL[1]

Dan–Andrei SITAR–TAUT[2]
PhD, Business Information Systems Department,
Babes - Bolyai University, Cluj Napoca, Romania

E-mail: dan.sitar@econ.ubbcluj.ro

Loredana MOCEAN[3]
PhD, Business Information Systems Department,
Babes - Bolyai University, Cluj Napoca, Romania

E-mail: loredana_mocean@yahoo.com

Adela–Viviana SITAR–TAUT[4]
PhD Candidate, Cardiology-Rehabilitation Department,
University of Medicine and Pharmacy Iuliu Hatieganu, Cluj Napoca, Romania

E-mail: adela.sitar@umfcluj.ro

Abstract: *This paper offers a new homogeneous approach and intends to provide a starting model with implemented elements useful in medicine and national health policies proposals. The final model will be entitled e-ProCord and it will represent one of the aims for an important research project in the field of cardiovascular evaluation at molecular level.*

Key words: *information system; database; cardiovascular risk; molecular; data minig*

1. Introduction

The EU is spending 50 million Euro annually from 2003 to 2008 to improve the collection of data, the exchange of information and to offer more information about primary and secondary prevention. The European Commission's major public health goals are: contribute to reducing the incidence of major diseases in the EU; contribute to the development of more effective and efficient health systems; providing medical information and analysis to support these goals. [1][5]

The e-Health project is one of the main priorities of the European Union. E-Health means better health as well as better ways of preventing illness through information

technology and communications (ITC). The e-Health plan of action explains the use of these technologies in order to provide better and cheaper health services while reducing waiting times and human error throughout Europe. The European plan is to develop electronic systems to store data regarding the population's health, identifying patients and last but not least broadband Internet transmissions.

We consider that at least part of the European Commission priorities for public health can be covered by the proposed project, considering the fact that it is attempting the identifying of various risk factors, the way in which these interact for the favoring for the appearance of clinical manifestations belonging to cardiovascular diseases, the way in which they can be influenced. We intend to create and validate a new cardiovascular risk scoring including not only the classic risk factors, but also markers of the endothelial dysfunction, that can be use as platform to create new policies of prevention.

The modern technologies will be applied to allow – through the use of WEB instruments – the anytime, anywhere secure access according to user categories to the available information. We will decide for an open source platform using MySQL as database server, and PHP as scripting programming language. We will use these technologies because they are free of charge. The software applications are cross-platform, they can be used not only on Windows operating systems, but also on Unix or Linux based ones. The database will suffer a solid design, using normalization as modeling tool. Normalization requires splitting the existing relations into atomic ones, which are free of redundancy and updating anomalies. It is also an answer for data confidence and security. The initial launch of the web site has more like a documentary and promoting goal; then it will become a powerful tool that can be used not only for our team's purposes, but also to assist the academic and non-academic communities. Mainly, in medicine there are huge amounts of unstructured data. The way how data is handled will be homogenized via XML. Simultaneously, XML will permit massive data updates with almost no effort from user's side. The data mining tools will offer a different modeling approach.

Databases represent an ordinary reality of the modern economic life. We find them everywhere, within prehistoric or advanced stages, simple or complex, small or colossal, locally accessed or by means of network technologies, being present in all fields. They represent collections, or integrated systems, coherent and shared system files [2], whose foundations were laid decades ago.

The relational model emerged as a viable alternative to previous and problematic hierarchical and network data models. The key paper on this revolutionary model [3] was published in ***1970***, the work of the mathematician ***E. F. Codd***. Even now, a large part of his ideas are still applicable, while other was validated in time. The process of logical data modeling is somewhat controversial and still debated. One of the most important aspects – if not crucial ones – of the logical data modeling within relational databases is the **normalization** process. Rigorous or flexible, mathematically sound or flimsily applied, either theoretical or pragmatic, the normalization built up its reputation as both "angel and demon" of the relational database design. It is generally defined as a process of elimination of redundancies within the entities. The normalization process is not part of the relational theory, rather completes it. The literature mentions 10 normal forms (and a few other less familiar), in the following relationship: [4]

1NF ⊆ 2NF ⊆ 3NF ⊆ EKNF ⊆ BCNF ⊆ 4NF ⊆3,3NF ⊆ 5NF ⊆ 6NF ⊆DK/NF

To compromise orthogonality and its disadvantages due to physical implementation, specialists agree that 3NF or BCNF are sufficient enough. We subscribe to this statement.

The various medical analyses we want to evaluate for our patients requires also one of the three supertype-subtype approaches. We will choose that one fitting better also for future processing (e.g. statistical), requiring minimum structural and content changes.

2. Actual knowledge stage in IT

The **Web** is believed to be *the largest federative databases system* worldwide. Although the Internet has skyrocketed with no apparent connection with the databases technology, nowadays, no relationship between the Web and database technologies is almost unconceivable. The web applications have passed the initial development stage and have evolved to a maturity stage characterized by complex applications. One of the great challenges that appear in this stage is the *communication between the applications and different types of platforms*. In the latest years this communication has become a basic condition for developing future systems, imposed by the necessity of a global environment and correlative virtual space. In these circumstances the interoperability of a system represents an important challenge, concerning the development of web applications and their necessity of interaction and correlation with one another. *XML represents in this context a viable solution for developing some web applications that will interact with informational systems that use different technologies and run on different platforms.*

Thus, XML has imposed as a metalanguage capable of offering the possibility of communication between applications running on different platforms and of creating a virtual correlative environment. XML permits the separation of information from its presentation, reaching in this way to create a self-descriptive document and permits the data transfer among applications. The advantage of this parting is the possibility of using also in other devices that can have internet connections. XML is independent of the platform, extensible and Unicode compliant. The impact of XML technologies can be considered a real evolution in the information technology domain. [5] In general web applications are based on the n-tier architecture in which the layer applications communicate between them, every layer being specialized and serving a certain functionality in the application. The advantages offered by such an architecture, flexibility, distributed components have imposed the n-tier architecture in realizing web applications. To the advantages offered by such an architecture are being added flexibility, independence towards the platform offered by XML documents. So, XML documents are used for transport of information between the layer of an application and towards other web applications no matter of the technologies used or the platforms that they run on. In actual fact it results a flexible, extendible and reusable structure, in which conditions a very important aspect must be taken into consideration: the application and data model can change over time and the advantages of a document structure that adjusts to different changes allows obtaining some important advantages.[6] As a format of database, XML has some advantages: is **self descriptive**, is **portable** (Unicode) and can **describe data in a branched structure**. A **drawback** of this language is the **difficult access to data as a result of the data processing need**.

The facilities offered by XML are:

★data storage (XML documents)

★descriptive language (DTS, XML Schema, etc.)
★query language (Xquery, Xpath, XQL, XML-QL)
★programming interface (SAX, DOM).

However, it has to be taken into account the fact that a series of *advantages* like the storage efficiency, security, indexation, transitioning, data integrity or the multi-user access are imposing the relational databases as a storage base for information [7], the XML documents being used for assuring the transfer of information from the database to different modules of an application or even to different applications.

One of the most prevalent standards in the medical field is Health Level 7 (HL7), which has now reached the 3rd version. Developed into a public-private partnership, the standard enjoys a wide acceptance by the American companies, being also supported by affiliated organizations from over 26 countries, the European ones inclusively. [8] HL7 has imposed as a standard for the electronic support exchange of information in the health field, in the clinical domain, as well as in the administrative one. The major objective of the HL7 standard constitutes the easy exchange of messages between the applications that manage the medical data. HL7 methodology is based on accepting as an event inside the medical procedure determines an exchange of messages between two or more applications. For example, hospitalizing a patient determines collecting information about him/her and transmitting them to other systems. The standard describes a Reference Informational Model (RIM) – fundamental model from which all HL7 messages derive. In 2000 as a component of RIM, CDA (the architecture of medical documents) – has been approved – as a set of XML specifications for the exchange of structured medical documents. CDA HL7 version 3 allows through a common form the exchange of medical data related to patients between different subsystems of the same hospital or even of different hospitals. This HL7 standard aims to constitute the base for the universal registration of medical data. The project frame for development (HDF) HL7 version 3 represents an evolved process that tries to develop specifications that facilitate interoperability between health caring systems.

Data mining, also known as knowledge discovery in databases, is the most recent technology for analyzing data, along with OLAP and data warehouses. Data mining refers to solving problems through already existing data present in the databases.

Data mining represents an analytical process that explores a very high number of data in searching for some patterns or relationships between variables, then generalizes these results into a model, formula or branched decision tree, and finally verifies the correctness of the generated model through testing it on the existing data set or of a new one. Initially data mining was a statistical term meaning overusing of data for deducing inferential invalids [9]. Bonferroni's theory warns us that in the situation that there are many possible conclusions some of them may be real only for pure statistics motives, without a physical validity. That is why the necessity of creating a new automatic instrument appeared, that should transform (through a corresponding procedure) different data contained in extra large databases into information and knowledge useful in the observing process as well as in decision making. In the data mining process, the data is stored electronically and the search is made automatically or at least completed with the help of computer [10] Defining for a data mining process is the fact that allows the discovery of some knowledge without prior formulation of some hypothesis. *Through this it is not aimed the check-up, confirmation/information of hypothesis, but it is aimed to discover some unexpected, unintuitive knowledge that can contradict the intuitive perception, being completely unknown*

at the moment of the process realization. The key elements that make Data Mining a distinct form of software are: [11] **Automatic analyze ✹Large or complex data sets.** One of the attractions of data mining is that it makes possible the analysis of extra large data sets in a reasonable period of time. Data mining is also suitable for complex problems that imply relatively small groups of data but in which there are a lot of fields or variables to analyze. **The second choice is the one that characterizes our study ✹The discovery of significant patterns or trends that otherwise would pass unnoticed.**

The objective in data mining is to discover the relationships between data that can provide useful meanings. The data mining instruments can scan databases and can identify patterns, previously hidden in a single step.

A successful example [9] is ✹ Comparing human kinds accomplishing or not a certain condition has allowed the discovery of a multitude of genes that together determine a lot of diabetes cases. This way of extracting knowledge from data becomes important in the moment of building human kind.

The process of data knowledge acquisition comprises the following steps:

✹selecting and processing data;

✹transforming data;

✹performing some methods and techniques for extracting knowledge from data (patterns, models);

✹validation and interpretation of results (consolidating results and obtaining the so called knowledge).

The process of acquisition of knowledge is iterative because, during this process, the mentions steps are executed repeatedly, through recommencing of some of them. Although the method and techniques for extracting knowledge are applied automatically, the process of acquisition of knowledge from data requires human effort.

The IT&C related elements mentioned above will be applied in all the stages of the project. They will be the platform for the advanced medical data processing (the discovery of new relationships between factors, new methods in statistical processing, etc.) in order to measure the cardiovascular risk factors.

3. Actual knowledge stage in medicine

The data accumulated in last years' medical literature outline new very interesting ideas concerning the general cardiovascular diseases, **in particular for the women**:

- Cardiovascular affections are the *main cause of mortality and main cause of hospitalization in both genders*
- Because of the amplitude of cardiovascular diseases (CVD), in 1997, the European Commission has started and published the **European Initiative for Cardiovascular Health**; also, the European Consensus for the European cardiology and preventive medicine has published in 1998 **recommendations and guides for cardiovascular prevention** with the intention of stimulating the development and reviewing of the national guides for CVD prevention. *The prevention and treatment of CVD is a national and international priority matter*.
- **Cardiovascular affections kill more women than men each year;**

♦ There are important differences between the appearance and the presentation way of CVD in women as compared to men (outlined by very large studies at Framingham, SAVE) and **still, there is a false perception that CVD are not a real threat for women.**

♦ Therefore, the women behave differently [12] [13] concerning the affection because of biological risk, a far more active health monitoring for men, studies conducted predominantly on males, the risk attained by the lifestyle, the way of reacting to the workplace, psychological status, because of the way in which they perceive the symptoms and the treatment of the disease [14] and the behavior related to the disease; the women's associated co-morbidities; previous medical assistance, the differences between women's cardiovascular system as compared with that of men's: physiological, in the CVD physiopathology – regarding the constituents of the atherosclerotic plaque, the endothelial dysfunction - as well as differences in the pharmakinetics and pharmadynamics;

♦ *For improving prevention and optimizing the CVD treatment for the women at menopause it is necessary to be conscious of the **cardiovascular risk factors**:*

♦ **age** and the cardiovascular affections history;

♦ **overweight or obesity** (a high corporal mass coefficient is usually associated with the dramatically increase in the cardiovascular risk factors [15];

♦ **arterial hypertension**;

♦ **smoking**;

♦ **the lipid profile** for women is affected by the hormonal status (therefore, the **estrogen** maintains a high level of HDL cholesterol, increase the level of triglycerides and decrease the level of LDL-cholesterol);

♦ **diabetes mellitus** as the enhancing factor for other risk factors (obesity, hypertension, and especially the hypertriglyceridemia, endothelial dysfunction), as well as an independent risk factor;

♦ **Physical activity** – physical inactivity is more frequent in women, being a reverse association between this, the obesity and cardiovascular events.

♦ **Menopause** – despite the increase in life expectancy to almost 80 years, the age of menopause has remained relatively constant. Also, there is an abrupt increase in the rate of mortality for post-menopause women. During a visible span of time, of almost several years, the estrogen deprivation effects appear, as well as the specific modifications for genital, osteoarticular, cardiovascular, psychosomatic, ocular, digestive and neurological systems. Although up to now the **cardiovascular risk determined by the menopause could not be quantified**, the alterations in vascular tonus, in the coagulation – fibrinolysis ratio for the seric lipoproteins are well known aterogenetic modifications. The Framingham study has shown that the **10 year cardiovascular disease incidence for women at post-menopause, 50 – 59 years has been four times greater than that of women at pre-menopause (women of the same age).**

♦ **Other biochemical factors** – like the cellular adhesion molecules (ICAM 1, VCAM 1), endothelin, plasminogen-activator inhibitor 1 (PAI 1), IL 6, MCP 1, E-selectin, P-selectin, myeloperoxidase, Lipoprotein-associated phospholipase A2 (Lp-PLA2), sCD40L, PCR–(far less dosed in our country) are associated with the increase in cardiovascular risk [16]. **Endothelial cells are not inert but rather are highly metabolically active** [17] **The endothelium plays an important role in many physiological functions**, including the control of vasomotor tone, blood cell trafficking, haemostatic balance, permeability, proliferation, survival, and innate and adaptive immunity. The vascular endothelium is a

highly active tissue regulating vascular tone, leucocytes and platelet activation, angiogenesis, inflammatory response, permeability and the metabolism of vascular mediators [17-19]. **There are many factors released by endothelium :** ★ vasoactive substances – **vasodilatating substances (**relaxing factor derived from endothelium, EDHF, prostacyclin, bradykinin, serotonin, P-substance**), but also vasoconstrictive (**endothelin, prostaglandin H2, reactive oxygen species ROS) ★ **growth's modulators – with positive** (Insulin-like growth factor-1– Interleukin 1, growth factor derived from platelets) or **negative influence** (heparin sulphate) ★inflammation's modulators – **adhesion molecule** (Endothelial leukocyte adhesion molecule –ELAM- , intracellular adhesion molecule –ICAM–, vascular adhesion molecule -VCAM-), P-selectin and **antibodies** (major histocompatibility complex) ★ **haemostatic and fibrinolysis factors .** [18,20]. Endothelial dysfunction is generally considered a key initial event in the atherosclerosis process, preceding the development of structural changes in the coronary arteries [19] The endothelial dysfunction (meaning **unbalance between vasodilators and vasoconstrictive factors, and factors produced by endothelial cells or affecting them** [21]) sets in before the atherosclerosis and is an essential factor in it progression. [20]. A current tendency is the early identification of the endothelial dysfunction, completing the characterization of the patients from a cardiovascular risk point of view. The endothelium is not amenable to traditional physical diagnostic maneuvers of inspection, palpation, percussion and auscultation. From a laboratory standpoint, the endothelium continues to elude convenient diagnostic interrogation. **The endothelium has enormous untapped potential as a therapeutic target** [17].

With all this exulted attention related to the link between endothelial dysfunction and atherogenesis, it is well known that the present studies and dates are not enough for a better therapeutically approach.

Due to the different impact of cardiovascular risk factors in men and women, the strategies for prevention should be treated differently for the two genders. [33].

4. Methodology of the research

The fast development of research imposes keeping a balance between the progress of science and aspects related to research ethics, or ethics in general.

The study will follow **more direction**. Subjects will be assigned to 1 of 2 groups:

1. patients **without cardiovascular disease** (men and women). The patients will be followed according to their age (under and over 55 years of age). Will no be included the patients in which we have no capacity of sure exclusion of this condition (using usual tests such as cardiologic exam, ECG, and sometimes, treadmill test or Holter ECG/24 hours).
2. patients **with coronary heart disease** (stable angina, old myocardial infarct, cardiac heart failure)- no matter gender and age.

We will **collect the information** using representative sample as method. The participation of patients to the research activities will be made only **after the individual informative agreement** will be signed**.**

The patients will **be prospectively followed** – successive clinical, paraclinical (classical cardiovascular risk factors, endothelial dysfunction markers), life's quality evaluation. Regarding endothelial dysfunction will be studied proinflamator status,

protrombtic status and vasomotor impairment. The **studies will be made in the same time** (the data from the first study will be used as starting point for the others-relation between cardiovascular risk factors and for the survival free of disease estimation). In the same time we will create the **informatic platform for efficient, secure, and orthogonal management of data**. It will be launched a new site not only for advertising purposes, but also with active communication features. The software that will support our research project are Windows (Vista, XP) as operating systems, MS Office (2007, 2003) package, MySQL, PHP, XML for databases and site.At the end of first medical and QoL related evaluation, we will start the **mathematical model conception and creation** in order to estimate cardiovascular risk score, as a function of endothelial dysfunction. The last part of project is represented by the estimation, on the basis of the previously made model, **of a cardiovascular risk score, taking** into account all these parameters.

The main purpose of the research project is cardiovascular risk evaluation, using new medical and modeling tools, but also creation of a cardiovascular risk score by introducing new variables.

5. Results

The big stages followed in the elaboration of our model are shown in figure 1.

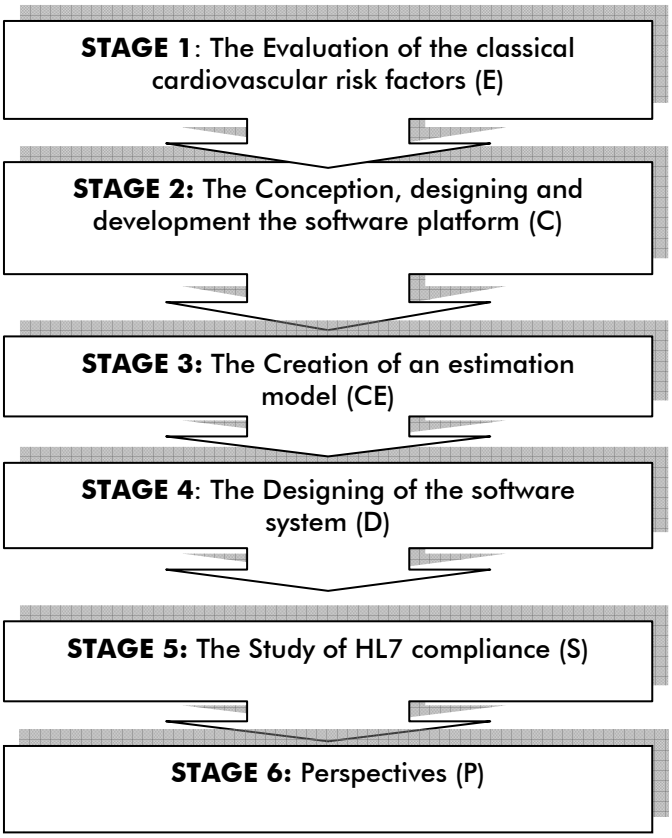

Figure 1. The levels followed in the elaboration of the model

In the following paragraph we detail the objectives followed at each phase:

Stage 1(E): The evaluation of the classical cardiovascular risk factors (overweight/obesity, smoking, dyslipidemia, hypertension, diabetes mellitus), the endothelial dysfunction (intracellular adhesion molecule ICAM 1, vascular adhesion molecule VCAM 1, P-selectin, plasminogen-activator inhibitor 1, endothelin, tumor necrosis factor TNF ά interleukin 6 IL-6) in:

- patients without cardiovascular disease (men and women), under and over 55 years of age;
- patients with coronary heart disease (stable angina, old myocardial infarct, cardiac heart failure)- no matter gender and age. A descriptive, observational and longitudinal study will be run in order to investigate the presence of risk factors, the modifications of the endothelial dysfunction markers.

A prospective, observational will be made in order to investigate cardiovascular risk factors and endothelial dysfunction. We will collect the information using representative sample as method.

Stage 2 (C): The conception, designing and development of a software platform for:

- The control of accuracy and confidentiality of the medical data
- The secure management and controlled disposal of study-specific data
- The promoting, through WEB technologies, of the goals and the achievements of this study
- The achievement of a communication bridge in the national and international medical academic community
- Discovering possible unknown patterns in cardiovascular diseases by using data mining methods or tools, data support and new approaches in modeling, homogenized data integration related to medical issues; Support for updating impressive data amount using almost automated procedures.

Stage 3 (CE): The creation of an estimation model for the cardiovascular risk

Stage 4 (D): The designing of a software system with a role of substitution for the medical experiment (by taking into account the cardiovascular risk score).

On the basis of previous studies will be estimated a cardiovascular risk score by taking into consideration various parameters (personnel and familial disease, used medication, age, smoking, weight, height, body mass index, blood pressure, glycemia, cholesterol, tri-glycerides, HDL, LDL, but also ICAM 1, VCAM 1, endothelin, PAI 1, P-selectin, tumor necrosis factor TNF ά, interleukin 6 IL-6)).

Stage 5 (S): The study of Health Level 7 (HL7) compliance in the perspective of integration in the European virtual space.

Stage 6 (P): Perspectives and new challenges occurrence identification, not only at the fundamental research stage, but also at applicative level.

The start of the project

e-ProCord starts in May 2009. Due to the financial and time constraints this year we will be able to evaluate a smaller patients sample than we proposed before. Also, the most expensive analysis will be delayed. In these circumstances the results will be used as the grounding for the initial modeling purposes considering also the theoretical and previous similar studies' results. The model will be confirmed or infirmed by the future investigations.

On the other hand, the obtained data will be considered as training tests for data mining tools as classifiers, clusterers, and associate makers. In our study we will compare the results provided by classifiers such Decision Table, NaïveBayse, ID3, and C4.5 (J48) [28], with different parameter settings, after data preprocessing and attribute quality evaluation (e.g. InfoGainAttributeEval or GainRatioAtributeEval from Weka) [10]. Then, the results will be compared by calculating the following performance indicators:

True Positive (TP) Rate $\frac{TP}{TP+FN}$ – proportion of patients classified as ill from all the persons who are proved being ill.

True Negative (TN) Rate $\frac{TN}{TN+FP}$ – proportion of patients classified as not being ill from all the persons who are not ill indeed.

The previous indicators show the proportion of the instances classified in a class from all the instances confirmed being in that class by the "gold standard". In medicine, the first one presents more interest (TP rate). It is also called **Sensitivity (Sensibility) –** probability of an ill subject to have a positive test – and the second one is **Specificity**, which is the probability for a not ill subject to have a negative result on examination.

False Positive (FP) Rate $\frac{FP}{FP+TN}$ – indicates how many healthy persons are considered ill using the test

False Negative (FN) Rate $\frac{FN}{FN+TP}$ – how many ill persons are considered healthy by our classifier.

Precision – proportion of the instances from a class which are really classified as that class.

Attributable risk (risk excess) – measures the connection's specificity between risk factor and disease, being calculated as difference between the risk of expose persons and those of non expose person

Success Rate (Accuracy) $\frac{TP+TN}{n}$ (n represents instances) – the number of correct classified instances from the total number of instances

ROC (Receiver-Operating-Characteristics) – represents a diagram on which every point has two coordinates (1-specify, sensibility). The indicator that resumes ROC is area under curve (the bigger values show that investigated test is more valuable, bringing more useful information).

Kappa Statistics (Cohen's Kappa) – measures the agreement between the prediction and reality:

$$K=\frac{(tp+tn)-[(tp+fn)*(tp+fp)+(fp+tn)*(fn+tn)]}{1-[(tp+fn)*(tp+fp)+(fp+tn)*(fn+tn)]},$$

$$tp=\frac{TP}{n}, tn=\frac{TN}{n}, fp=\frac{FP}{n}, fn=\frac{FN}{n}$$

For the beginning, we want to create an initial model, using data from an important study –INTERHEART **[0]**. The main purpose of this was to investigate the

relationship between cardiovascular risk factors and risk of myocardial infarction. In this study (which included approximately 20,000 people, mean age of 58 years), the investigators reported:

- the prevalence of cardiovascular risk factors in controls (subjects without myocardial infarction) and in cases (subjects with myocardial infarction)
- population attributable risk for every cardiovascular risk factor.

We used this information to simulate datasets in order to evaluate the prediction capacity in one subject the disease's presence or absence. Using Monte Carlo simulation, we generated tables with pseudo-random values, according to the risk factor's prevalence and attributable risk and with the respect of independency between attributes. We selected as risk factors current smoking, diabetes mellitus, hypertension, abdominal obesity, no diet with fruits or vegetables, no physical exercise, and no consume of alcohol in small quantities. We considered that a subject without risk factor's presence has an attributable risk equal to one, and a subject in which presence of risk factor is identified has a attributable risk equal with 1 plus the value reported (for the respective risk factor) in INTERHEART (transforming the percentage values in values between 0 and 1). After this, we calculated the individual aggregated attributable risks as a product of attributable risk for each factor. We created four tables with 100, 1,000, and 30,000 (two times) respectively instances. We computed for every dataset – using MedCalc 10.4.0.0 – the capacity of compound attributable risk factor to predict the disease presence in a period of time similar to referred study. The results are presented in Table 1.

Table 1. Testing results

	100 instances	***1,000 instances***	***30,000 instances V1***	***30,000 instances V2***
Sensibility (%)	82.7	58	60	61.4
Specificity (%)	52	71	65	65.5
Cut-off value	1.337	1.55	1.5	1.52
AUC (confidence interval)	0.689 (0.588-0.778)	0.676 (0.646-0.705)	0.675 (0.670-0.680)	0.677 (0.672-0.682)
p	**0.0003**	**0.0001**	**0.0001**	**0.0001**

As we can see, no matter the number of instances generated, the AUC was significantly different from 0.5. A cut-off value more than 1.5 is capable, in approximately 60 percent of the cases, to identify ill people. In approximately 65% of the cases with no disease the test will be negative. If we consider (on the third simulation), a cut-off value equal to 1.91, then the specificity is increasing to 90%.

We also compared the area under curve for different risk factors with those ones obtained for the compound attributable risk. As we can see in Figure 2 and in Table 2, compound attributable risk had a bigger area under curve. In the same time, significant differences were registered between AUC for compound attributable risk and AUCs obtained for others risk factors ($p<0.001$).

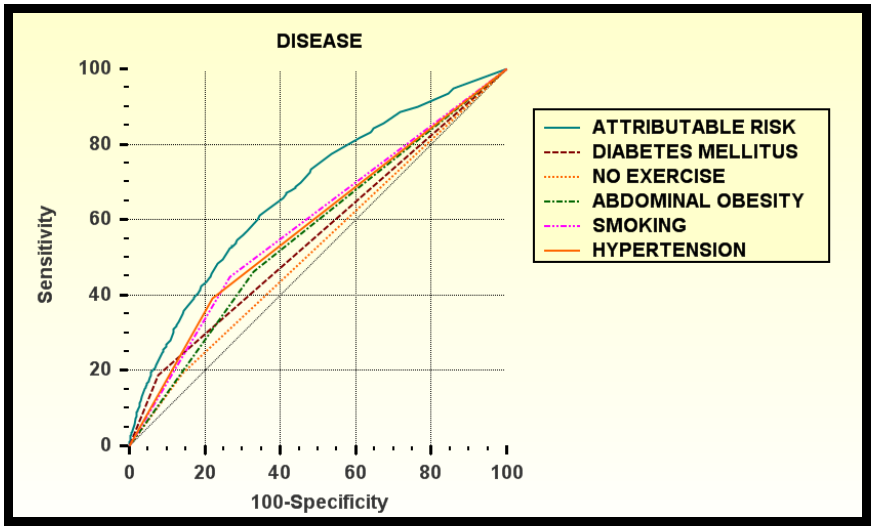

Figure 2. Comparative AUCs

Table 2. Risk factors comparison indicators (30,000 instances V2)

RISK FACTOR	AUC	STANDARD ERROR	CONFIDENCE INTERVAL
ABDOMINAL OBESITY	0.566	0.00332	0.560 to 0.572
DIABETES MELLITUS	0.555	0.00333	0.549 to 0.561
HYPERTENSION	0.586	0.00330	0.580 to 0.592
NO_EXERCISE	0.526	0.00334	0.520 to 0.531
SMOKING	0.590	0.00329	0.585 to 0.596
ATTRIBUTABLE RISK	**0.677**	**0.00311**	**0.672 to 0.682**

This model will be tested using the real data obtained during e-ProCord study.

Discussion

In Romania, not only that compared to the EU countries, it has a **higher mortality rate** due to the cardiovascular affections, but what is even more alarming is the fact that in other countries the rate is decreasing while **in our country is increasing**. We don't have real information about cardiovascular disease evolution and tendency.

National Program regarding Evaluation of Health Status in Romanian Population had as main targets to identify the prevalence of cardiovascular risk factors in general population, to diagnose and to monitories severe diseases in order to avoid premature deaths [30]. What results do we have? The survey showed that almost 15 percents of Romanian people are at risk for cardiovascular diseases development, that cardiovascular risk factors had been identified in a large number of subjects (for example 29.26% of subjects had been diagnosed with diabetes mellitus [31]). Unlike the National Program, which, due to the financial aspects, did not succeed to evaluate in a complete way the cardiovascular risk factors, our project has as the main goal to "heal this deficiency", even if using a small sample. In the same time, we don't have studies about relationships between molecular dysfunction and classical cardiovascular risk factors (depending on age, gender and pathology). Thus we have to try identifying the molecular differences between patients, in order to consider also the classical risk factors, to be able to implement an efficient preventive policy.

Conclusions

So, in consideration with Romanian presented situation, we need "aggressive" measures for cardiovascular disease's prevention, evaluation, control, and research projects like e-ProCord representing just one of the firs steps made in this field.

References

1. Aird, W.C. **Endothelium in health and disease,** Pharmacological reports, 2008, pp. 139-143
2. Albert, M., Glynn, R., Buring, J. E. and Ridker, P. M. **Relation between soluble intercellular Adhesion Molecule-1, homocisteine and fibrinogen levels and race/ethnicity in women without cardiovascular disease,** 99, Am J Cardiol, 2007, pp. 1246-1251
3. Albert, M.A. and Ridker, P. M. **Reactive Protein as a Risk Predictor - Do Race/Ethnicity and Gender Make a Difference,** Circulation, 114, 2006, pp. e67-e74
4. Anghel, T. **Programarea in PHP. Ghid practic,** Ed. Polirom, Iassy, 2005
5. Balaita, R., Hulea, M. and Olariu, C. **Data Mining - Inteligenta artificiala,** Universitatea Tehnica Gh. Asachi Iassy, Facultatea de Automatica si Calculatoare, 2003, Biblioteca virtuala de Inteligenta artificiala, http://eureka.cs.tuiasi.ro/~fleon/bvia.htm
6. Benıtez, M.B. *et all*. **Endothelial and leukocyte adhesion molecules in primary hypertriglyceridemia,** Atherosclerosis, 2007
7. Campisi, R. **Noninvasive assessment of coronary microvascular function in women at risk for ischaemic heart disease,** Int J Clin Pract, 62, 2, 2008, pp. 300–307
8. Codd, E. F. **A Relational Model of Data for Large Shared Data Banks,** CACM 13, 1970
9. Cooper, B., Burger, N., Toth, M., Cushman, M. and Sites, C. **Insulin Resistence with hormone replacement therapy: associations with markers of inflammation and adiposity,** Am J obstet Gynecol, 196, 2007, pp. 123.e1-123e7
10. Daconta M.C., Obrts, L.J. and Smith, K.T. **The Semantic Web - A Guide to the Future of XML,** Web Services and Knowledge Management, Wiley Publishing, 2003
11. Davis, S.R., Dinatale, I., Rivera-Woll, L. and Davison, S. **Obese women and the relation between cardiovascular risk profile and hormone therapy, glucose tolerance, and psychosocial conditions,** Diabetes Care, 29(11), 2006, pp. 2477-2482
12. De Ferrari, L. and Aitken, S. **Mining housekeeping genes with a Naive Bayes classifier,** BMC Genomics, 2006, 7:277
13. DeSouza, C. A., Dengel, D. R., Macko, R. F., Cox, K. and Seals, D.R. **Elevated Levels of Circulating Cell Adhesion Molecules in Uncomplicated Essential Hypertension,** Am J Hypertens, 10, 1997, pp. 1335–1341
14. Erl, Th. **Service-Oriented Architecture, A field guide to Integrating XML and Web Services,** Prentice Hall PTR, 2004
15. Kublickiene, K. and Luksha, L. **Gender and the endothelium,** Pharmacological reports, 60, 2008, pp. 49-60
16. Marchesi, S., Vaudo, G., Lupattelli, G., Lombardini, R., Roscini, A.R., Brozzetti, M., Siepi, D. and Mannarino, E. **Fat distribution and endothelial function in normal-overweigh menopausal women,** Journal of Clinical Pharmacy and Therapeutics, 32, 2007, pp. 477–482
17. Milner, K.A., Funk, M., Richards, S., Wilmes, R.M., Vaccarino, V. and Krumholz, H.M. **Gender differences in symptom presentation associated with coronary heart disease,** Am J Cardiol, 84, 1999, pp. 396-399
18. Nitchi, St. and Racovitan, D.M. **Bazele prelucrarii informatiilor si tehnologie informationala,** Ed. Intelcredo, Deva, 1996
19. Obrenović-Kirćanski, B. **Endothelial dysfunction reversibility,** Vojnosanit Pregl, 64 (5), 2007, pp. 337-343
20. Omran, S. and Al-Hassan, M. **Gender differences in signs and symptoms presentation and treatment of Jordanian myocardial infarction patients,** Int J Nurs Pract, August, 12 (4), 2006 pp. 198-204

21. Sitar-Taut, D.A. **Baze de date distribuite,** Ed. RISOPRINT, Cluj Napoca, 2005
22. Stollberger, C. **Gender and mortality in myocardial infarction -more similarities than differences,** Eur. Heart J, 22 (4), 2001, pp. 279 –80
23. Sumino, H., Ichikawa, S., Ohyama, Y., Takahashi, T., Saito, Y., Nakamura, T., Kanda, T. and Kurabayashi, M. **Effect of transdermal hormone replacement therapy on the monocyte chemoattractant protein-1 concentration and other vascular inflammatory markers an don endotelial function in postmenopausal women,** Am j Cardiol, 96, 2005, pp. 148-153
24. Ullman, J.D. **Data Mining Lecture Notes,** Stanford University, 2003, http://infolab.stanford.edu/~ullman/
25. Vitale, C., Miceli, M. and Rosano, C. **Gender-specific characteristics of atherosclerosis in menopausal women risk factors, clinical course and strategies for prevention,** Climacteric, 10, Suppl 2, 2007, pp. 16–20
26. Witten, I.H. and Frank, E. **Data mining: practical machine learning tools and techniques, 2nd edition,** Morgan Kaufmann series, Elsevier Inc., 2005
27. Yusuf, S., Hawken, S., Ôunpuu, S., Dans, T., Avezum, A., Lanas, F., McQueen, M., Budaj, A., Pais, P., Varigos, J. and Lisheng, L. **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, 364, 2004, pp. 937–952
28. * * * **Health Level 7 (HL7)**, http://www.hl7.org
29. * * * http://www.Europe.eu.int/scadplus/leg/en/lvb/l24226f.htm
30. * * * http://www.ms.ro/fisiere/programe_nationale/13_51_334_h.htm
31. * * * http://www.romedic.ro/diabetul-zaharat-afecteaza-cei-mai-multi-romani-0N5786

[1] **Acknowledgement**
The current research was financed from The National University Research Council*, project no.2246/2008, entitled "E-PROCORD – New Medical and Modeling Approaches in IT&C Age Applied on Cardiovascular Profile Evaluation at Molecular Level. Differences Implied by Gender, Age, and Existing Pathology"**, IDEI Program, PNCDI PN II.
*CNCSIS – Consiliul Național al Cercetarii Stiintifice din Învatamantul Superior (in Romanian)
**E-PROCORD – Noi abordari medicale si de modelare in era TI & C in evaluarea profilului cardiovascular la nivel molecular. Diferente determinate de sex, varsta si patologie prezenta (in Romanian)

[2] Dan-Andrei SITAR-TAUT 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, about 30 papers in the field of Databases, ERP, Data mining, and Web.

[3] Loredana MOCEAN has graduated Babes-Bolyai University of Cluj – Napoca, the Faculty of Computer Science in 1993, she holds a PhD diploma in Economics from 2003 and she had gone through didactic position of assistant and lecturer, since 2000 when she joined the staff of the Babes-Bolyai University of Cluj- Napoca, Faculty of Economics and Business Administration.
She is the author of more than 10 books and over 35 journal articles in the field of Databases, Data mining, Web Services, Web Ontology, ERP Systems and much more.

[4] Adela-Viviana SITAR-TAUT has a Bachelor's degree in Business Information Systems from Faculty of Economics and Business Administration, Babeş-Bolyai University of Cluj-Napoca and a Bachelor's degree in General Medicine from "Iuliu Haţieganu" Medicine and Pharmacy University from the same city. Currently she is a resident medical staff in Internal Medicine and also Ph. D. candidate in Cardiology.
She participated and published papers to many national and international events on Medicine, Business Information Systems, and Bioinformatics related topics.

[5] Codification of references:

[1]	* * * http://www.Europe.eu.int/scadplus/leg/en/lvb/l24226f.htm
[2]	Nitchi, St. and Racovitan, D.M. **Bazele prelucrarii informatiilor si tehnologie informationala,** Ed. Intelcredo, Deva, 1996
[3]	Codd, E. F. **A Relational Model of Data for Large Shared Data Banks,** CACM 13, 1970
[4]	Sitar-Taut, D.A. **Baze de date distribuite,** Ed. RISOPRINT, Cluj Napoca, 2005
[5]	Daconta M.C., Obrts, L.J. and Smith, K.T. **The Semantic Web - A Guide to the Future of XML,** Web Services and Knowledge Management, Wiley Publishing, 2003

[6]	Erl, Th. **Service-Oriented Architecture, A field guide to Integrating XML and Web Services,** Prentice Hall PTR, 2004
[7]	Anghel, T. **Programarea in PHP. Ghid practic,** Ed. Polirom, Iassy, 2005
[8]	* * * **Health Level 7 (HL7)**, http://www.hl7.org
[9]	Ullman, J.D. **Data Mining Lecture Notes,** Stanford University, 2003, http://infolab.stanford. edu/~ullman/
[10]	Witten, I.H. and Frank, E. **Data mining: practical machine learning tools and techniques, 2nd edition,** Morgan Kaufmann series, Elsevier Inc., 2005
[11]	Balaita, R., Hulea, M. and Olariu, C. **Data Mining - Inteligenta artificiala,** Universitatea Tehnica Gh. Asachi Iassy, Facultatea de Automatica si Calculatoare, 2003, Biblioteca virtuala de Inteligenta artificiala, http://eureka.cs.tuiasi.ro/~fleon/bvia.htm
[12]	Milner, K.A., Funk, M., Richards, S., Wilmes, R.M., Vaccarino, V. and Krumholz, H.M. **Gender differences in symptom presentation associated with coronary heart disease,** Am J Cardiol, 84, 1999, pp. 396-399
[13]	Stollberger, C. **Gender and mortality in myocardial infarction -more similarities than differences,** Eur. Heart J, 22 (4), 2001, pp. 279 –80
[14]	Omran, S. and Al-Hassan, M. **Gender differences in signs and symptoms presentation and treatment of Jordanian myocardial infarction patients,** Int J Nurs Pract, August, 12 (4), 2006 pp. 198-204
[15]	Davis, S.R., Dinatale, I., Rivera-Woll, L. and Davison, S. **Obese women and the relation between cardiovascular risk profile and hormone therapy, glucose tolerance, and psychosocial conditions,** Diabetes Care, 29(11), 2006, pp. 2477-2482
[16]	Albert, M.A. and Ridker, P. M. **Reactive Protein as a Risk Predictor - Do Race/Ethnicity and Gender Make a Difference,** Circulation, 114, 2006, pp. e67-e74
[17]	Aird, W.C. **Endothelium in health and disease,** Pharmacological reports, 2008, pp. 139-143
[18]	Kublickiene, K. and Luksha, L. **Gender and the endothelium,** Pharmacological reports, 60, 2008, pp. 49-60
[19]	Campisi, R. **Noninvasive assessment of coronary microvascular function in women at risk for ischaemic heart disease,** Int J Clin Pract, 62, 2, 2008, pp. 300–307
[20]	Obrenović-Kirćanski, B. **Endothelial dysfunction reversibility,** Vojnosanit Pregl, 64 (5), 2007, pp. 337-343
[21]	Vitale, C., Miceli, M. and Rosano, C. **Gender-specific characteristics of atherosclerosis in menopausal women risk factors, clinical course and strategies for prevention,** Climacteric, 10, Suppl 2, 2007, pp. 16–20
[22]	Albert, M., Glynn, R., Buring, J. E. and Ridker, P. M. **Relation between soluble intercellular Adhesion Molecule-1, homocisteine and fibrinogen levels and race/ethnicity in women without cardiovascular disease,** 99, Am J Cardiol, 2007, pp. 1246-1251
[23]	Sumino, H., Ichikawa, S., Ohyama, Y., Takahashi, T., Saito, Y., Nakamura, T., Kanda, T. and Kurabayashi, M. **Effect of transdermal hormone replacement therapy on the monocyte chemoattractant protein-1 concentration and other vascular inflammatory markers an don endotelial function in postmenopausal women,** Am j Cardiol, 96, 2005, pp. 148-153
[24]	Cooper, B., Burger, N., Toth, M., Cushman, M. and Sites, C. **Insulin Resistence with hormone replacement therapy: associations with markers of inflammation and adiposity,** Am J obstet Gynecol, 196, 2007, pp. 123.e1-123e7
[25]	Marchesi, S., Vaudo, G., Lupattelli, G., Lombardini, R., Roscini, A.R., Brozzetti, M., Siepi, D. and Mannarino, E. **Fat distribution and endothelial function in normal-overweigh menopausal women,** Journal of Clinical Pharmacy and Therapeutics, 32, 2007, pp. 477–482
[26]	Benıtez, M.B. *et all.* **Endothelial and leukocyte adhesion molecules in primary hypertriglyceridemia,** Atherosclerosis, 2007
[27]	DeSouza, C. A., Dengel, D. R., Macko, R. F., Cox, K. and Seals, D.R. **Elevated Levels of Circulating Cell Adhesion Molecules in Uncomplicated Essential Hypertension,** Am J Hypertens, 10, 1997, pp. 1335–1341
[28]	De Ferrari, L. and Aitken, S. **Mining housekeeping genes with a Naive Bayes classifier,** BMC Genomics, 2006, 7:277
[29]	Yusuf, S., Hawken, S., Ôunpuu, S., Dans, T., Avezum, A., Lanas, F., McQueen, M., Budaj, A., Pais, P., Varigos, J. and Lisheng, L. **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, 364, 2004, pp. 937–952
[30]	* * * http://www.ms.ro/fisiere/programe_nationale/13_51_334_h.htm
[31]	* * * http://www.romedic.ro/diabetul-zaharat-afecteaza-cei-mai-multi-romani-0N5786