Assessment of the Risk of Coronary Heart Event Based on Data Mining

M. Karaolis, *Member, IEEE*, J A. Moutiris, FESC, C.S. Pattichis, *Senior Member, IEEE*

Abstract -- Coronary heart disease (CHD) is a major cause of morbidity and mortality in the western world. Although significant progress has been made in the diagnosis and treatment of CHD, further investigation is still needed. The objective of this study was to develop a data mining system for the assessment of heart event related risk factors. The risk factors investigated were: i. clinical: sex, age, smoking, systolic blood pressure, family history for premature CHD, history of hypertension, and diabetes; and ii. biochemical: cholesterol, triglycerides, and glucose. The events investigated were: myocardial infarction (MI), percutaneous coronary intervention (PCI), and coronary artery bypass graft surgery (CABG). A total of 620 cases were collected from the Paphos district in Cyprus, most of them with more than one event. Data mining analysis was carried out using the C4.5 decision trees algorithms. The most important risk factors, as extracted from the classification rules analysis were: sex, age, smoking, blood pressure, and cholesterol. Most of these risk factors were also extracted by other investigators. It is anticipated that data mining could help in the identification of high and low risk subgroups of patients, a decisive factor for the selection of therapy, i.e. medical or surgical. However, further investigation with larger data sets is still needed.

Keywords: data mining, coronary heart disease, risk of heart event, C4.5, PART, risk factors, rule extraction.

I. Introduction

Coronary Heart Disease (CHD) is one of the major causes of disability in adults as well as one of the main causes of death in the developed countries. Advances in the field of medicine in the past few decades enabled the identification of risk factors that may contribute towards the development of CHD. However, this knowledge has not yet helped in the significant reduction of CHD incidence. There are several factors that contribute to the development of a coronary heart event. These risk factors may be classified into two categories, not-modifiable and modifiable. The first category includes factors that cannot be altered by intervention such as age, gender, operations, family history and genetic attributes.

Modifiable risk factors are those for which either treatment is available or in which alternations in behavior can reduce the proportion of the population exposed. Established, modifiable risk factors for CHD currently include smoking, cholesterol, triglycerides, glucose, blood pressure, hypertension, and diabetes [1]-[2]. There are a number of other 'well-established' risk factors and protective factors that are also modifiable, but there are also a number of other known factors that are not yet considered to be of great importance.

With the advancement of medical tools it is now possible to gather complex clinical data from an individual that could give information regarding the risk of having a heart episode. The data that give information about the risk are complex and multifactorial thus making calculation of the risk by just viewing the complete data an extremely difficult task. Handling of such data could be made possible via data mining.

The objective of this study was to develop a data mining system for the assessment of CHD related risk factors. These risk factors were sex, age, smoking, systolic blood pressure, history of hypertension, family history, diabetes cholesterol, triglycerides, and glucose. A total of 620 cases were analysed, most of them symptomatic. Data mining analysis was carried out using the C4.5 decision trees algorithms for extracting rules based on the aforementioned risk factors.

Many studies were carried out investigating CHD and related risk factors [3]-[18]. Data mining was employed in some of these studies, where different algorithms were used for rule extraction and evaluation like the C4.5, k-means, decision tree models and Apriori.

In this study we investigate how data mining can help for the evaluation of the risk of CHD and the importance of each factor separately. The aim is to identify the most important risk factors based on the classification rules to be extracted. The correct selection of these rules will enable the definition of new patterns, so that the evaluation of an individual's risk to have CHD will be simpler. Evaluation is very important regarding two aspects. First, it will minimize the number of episodes and second, it will reduce the cost of therapy, due to the expected restriction of interventions in the absolutely necessary cases.

The rest of the paper is organized as follows. Section II describes the Material and Methods, Section III the Results and Discussion, and Section IV the Conclusions.

II. Material and Methods

A. Data Collection

Data from 1200 consecutive CHD patients were collected, between the years 2003 – 2006 (300 patients each year) according to a pre-specified protocol, under the supervision of the participating cardiologist (Dr J. Moutiris,

Manuscript received July 5 2008, accepted August 31 2008.

M. Karaolis and C. Pattichis, are with the Department of Computer Science, University of Cyprus, Nicosia, Cyprus (e-mail: karaolis@spidernet.com.cy; pattichi@ucy.ac.cy).

J. Moutiris, is with the Department of Cardiology, Paphos General Hospital, Paphos, Cyprus (email: moutiris@ucy.ac.cy).

second author of this paper) at the Paphos General Hospital of Cyprus. Patients had at least one of the following criteria on enrollment: history of: myocardial infarction (MI), percutaneous coronary intervention (PCI), or coronary artery bypass graft surgery (CABG). Data for each patient were collected under the following groups: i. Clinical factors: Sex, Age, Smoking (SMBEF), systolic blood pressure (SBP) mmHg, history of hypertension (HT), family history (FH), and Diabetes (DM); ii. Biochemical factors: Cholesterol (TC) mg/dL, Triglycerides (TG) mg/dL, and Glucose (GLU) mg/dL.

B. Data Cleaning

The collected data were used to create a structured database system. The fields were identified, duplications were extracted, missing values were filled, and the data were coded. After data cleaning the number of cases was reduced to 620, mainly due to unavailability of biochemical results. The number for MI cases was 416, for PCI 253, and for CABG 257. The database was build and the key fields were identified. The structured data from the above database were used to develop the cubes in an SQL Server. These cubes were further analyzed using data mining tools for the extraction of graphs and rules to evaluate the risk factors.

C. Data Coding

The risk factors collected with their corresponding codings are given in Table I. The criteria for data coding were provided by the participating cardiologist and are as coded by the American and European Heart disease associations.

TABLE I
RISK FACTORS WITH THEIR CORRESPONDING CODINGS

	Risk Factor	Code 1	Code 2	Code 3	Code 4
1	SEX	M: MALE	F: FEMALE		
2	AGE	1: 34 - 50	2: 51 - 60	3: 61 - 70	4: 71 -85
3	SMBEF	Y: YES	N: NO		
4	SBP mmHg	B: <120	H: 121 - 140	O: >140	
5	HT	Y: YES	N: NO		
6	FH	Y: YES	N: NO		
7	DM	Y: YES	N: NO		
8	TC mg/dL	D: <200	B: 201 - 240	H: >240	
9	TG mg/dL	B: <150	H: 151 - 200	O: >200	
10	GLU mg/dL	H: >110	N: <110		

D. Classification by Decision Trees

The C4.5 algorithm [19], [20], the successor and refinement of ID3, which have the divide-and-conquer approach to decision tree induction, was employed. The algorithm uses the information gain criterion and the gain ratio. It works top-down, seeking at each stage an attribute to split on that which best separates the classes, and then recursively processing the sub problems that result from the split. The algorithm uses heuristics for pruning derived based on the statistical significance of splits. The Weka's implementation of this algorithm called J4.8, was ran [21].

E. Pattern Evaluation and Knowledge Representation

The following three different set of models for classifying a patient were investigated: (i) MI versus PCI or CABG, (ii) PCI versus MI or CABG, and (iii) CABG versus MI or PCI. For each of these models, the steps documented in Fig. 1 were carried out for data mining classification and pattern evaluation. Rules were extracted from different combinations of risk factors. A minimum of one to a maximum of ten risk factors were extracted from the different rules (see Fig. 1).

More specifically, selected rules were evaluated according to the importance of each rule. Each extracted rule was further evaluated by inspection of the number of cases from within the database that support the rule. Rules with a small number of records were ignored. We started with the strongest rules; that means the rules that were supported by most records in the database. We initially took the rules with one risk factor. In these rules the hierarchy of the risk of CHD evidently appeared i.e. the higher risk to the lower risk considering the number of cases. As second step we took the rules with two risk factors. Taking the risk factors with the highest percentage that we found in the first step, we checked which risk factor was the second higher for risk of CHD. The same strategy was used for the next step, taking into account rules with 3 risk factors, afterwards with 4, until 10 risk factors. This exercise finally concluded on the hierarchy of the risk factors. More specifically, hierarchy of risk factors based on the values of each risk factor was achieved.

Step 1: Run the C4.5 algorithm
  1.1 for i = 1, …, 10 risk factors, and
  1.2 all their combinations (i.e. for one, 2, 3 and 10 risk factors, 10, 45, 120, and 1 runs were carried out).
  The classification outcome for each rule is the frequency of MI, or PCI, or CABG.
Step 2: Delete duplicated rules.
Step 3: Sort rules for i = 1, …, 10 risk factors, based on the frequency of heart events.
Step 4: Compare rules with consecutive number of risk factors (i.e. i and i+1) and derive importance of i+1 risk factor based on the frequency of MI, or PCI, or CABG.

Fig. 1. Steps for data mining classification

F. Performance Measures

Ten-fold cross validation was used for evaluating the performance of the proposed models. The data were split into ten approximately equal partitions and each in turn was used for testing and the remainder was used for training. This procedure was repeated ten times so that, in the end, every instance has been used exactly once for testing. It is noted that the extracted rules derived from the different sets were similar.

In order to evaluate the performance of our results we used the following measures [21]:

- Correct Classifications (CC): is the percentage of the correctly classified records.
- True positive rate or sensitivity (TP Rate): is the percentage of the positive examples that are correctly classified.

$$Sensitivity = \frac{TP}{TP + FN}$$

- True negative rate or specificity (TN Rate): is the percentage of negative examples that are correctly classified.

$$Specificity = \frac{TN}{TN + FP}$$

- False positive rate (FP Rate): is the percentage of negative examples that are correctly classified.

$$FPRate = \frac{FP}{TN + FP}$$

- Precision (PR): The records that the classifier has classified as positive and are truly positive. The higher the accuracy, the smaller the number of FP.

$$PR = \frac{TP}{(TP + FP)}$$

- TP: Positive records that have been classified as positive.
- Support: is the number of cases for which the rule applies (or predicts correctly; i.e. if we have the rule X & Y → Z, Support is the probability that a transaction contains {X, Y, Z} [22]. Also, called Coverage.
- Confidence: is the number of cases for which the rule applies (or predicts correctly), expressed as a percentage of all intances to which it applies (i.e. if we have the rule X & Y → Z, Confidence is the conditional probability that a transaction having {X, Y} also contains Z [22]. Also, called Accuracy.

TABLE II EVALUATION MEASURES FOR C4.5 MODELS

Model	Correct Classif. %	TP Rate %	FP Rate %	PR %	TP
MI	61	75	68	69	311
PCI	58	66	53	64	241
CABG	57	66	55	63	238

III. RESULTS

A total of 620 patients were investigated (MI = 416, PCI = 253, CABG = 257). Fourteen percent (14%) of the patients were female; 88% of them non-smokers and 12% smokers, 14% within the age range of 51-60 years old, 47.5% 61-70 years old and the remaining 38.5% 71-85 years old. No female patient was under the age of 50 years old; only male patients were found under this age (this finding has also been documented by the American Heart Association) [23]. It is also interesting to note that a number of patients (18.5%) continued smoking after the event in spite of medical advice.

Table II tabulates the evaluation measures for the three different models investigated, for classifying a patient as: (i) MI versus PCI or CABG, (ii) PCI versus MI or CABG, and (iii) CABG versus MI or PCI. The corresponding rules for these models are given in Tables III, IV, and V respectively.

A. MI model

The overall percent of correct classifications score for the MI model was 61%, with 75%, 68%, and 69% for TP Rate, FP Rate, and PR, respectively. Detailed rule analysis for this model is given in Table III. More specifically, the following rules can be extracted:

- Overall, 258 out of 370 smokers (support = 62%, and confidence = 70%) had an MI event [Table III, rule 1.1].
- Among the smokers group a small increase of MI was noted within the group of individuals who also have high cholesterol (no. of cases = 25, support = 6%, and confidence = 76%) [Table III, rules 1.1 and 1.2].
- Interestingly, 40 out of the 44 male smokers in the age 34-50 group (support = 10%, and confidence = 91%) had an MI event [Table III, rule 1.3]. This group corresponds to 31% of all MI patients and to 6,5% of all patients.
- An increase of MI was noted within the non-smokers group for individuals who had high systolic blood pressure (no. of cases = 84, and 22, support = 20%, and 5%, and confidence = 73%, and 82%, respectively) [Table III, rules 2.1 and 2.2].
- Out of the men between 61-70 years of age with SBP higher than 140 mmHg, and glucose higher than 110 mg/dL, 92% had an MI event (no. of cases = 11, support = 11%, and confidence = 71%) [Table III, rule 3.2].

B. PCI model

The overall percent of correct classifications score for the PCI model was 58%, with 66%, 53%, and 64% for TP Rate, FP Rate, and PR, respectively. Detailed rule analysis for this model is given in Table IV. More specifically, the following rules can be extracted:

TABLE III EXTRACTED RULES FOR MI EVENTS MODEL (FOR RISK FACTOR CODINGS SEE TABLE I)

Rule	SEX	AGE	SMBEF	SBP	HT	DM	TC	TG	GLU	FH	MI event	# of cases	Support	Confidence
Rules with risk factor smoking, SMBEF = Y														
1.1			Y								Y	258	62%	70%
1.2			Y				H				Y	25	6%	76%
1.3	M	1	Y								Y	40	10%	91%
1.4			Y			N	B				Y	31	7%	77%
1.5			Y				B		H		Y	15	4%	79%
1.6	M		Y	H			D		H		Y	23	6%	85%
Rules with risk factor smoking, SMBEF = N														
2.1	M		N								Y	84	20%	73%
2.2			N	H							Y	22	5%	82%
2.3			N				D		N		Y	20	5%	80%
2.4	M	2	N								Y	13	3%	87%
2.5	M	1	N								Y	11	3%	92%
2.6	F		N	B			D				Y	11	3%	100%
Rules with more than 3 risk factors														
3.1	M	3		H	Y						Y	16	4%	84%
3.2	M	3		O					H		Y	11	3%	92%
3.3	M	2		O		N		O			Y	10	2%	63%

TABLE IV EXTRACTED RULES FOR PCI EVENTS (FOR RISK FACTOR CODINGS SEE TABLE I)

Rule	SEX	AGE	SMBEF	SBP	HT	DM	TC	TG	GLU	FH	PCI event	# of cases	Support	Confidence
Rules with 2 risk factors														
1.1		1		O							Y	19	7%	76%
Rules with 3 risk factors														
2.1	F		N	H							Y	9	4%	90%
2.2		3	Y			Y					Y	10	4%	67%
2.3		2	Y	O							Y	29	11%	64%
2.4		3	Y	B							Y	16	6%	64%
Rules with more than 3 risk factors														
3.1		2	Y	H					N		Y	12	5%	75%
3.2	M	3	N	H							Y	10	4%	71%
3.3	M	2	Y				B		N		Y	12	5%	75%
3.4	M	2	Y			Y	D				Y	10	4%	71%
3.5		2			Y	N	D		N	N	Y	11	4%	79%

TABLE V EXTRACTED RULES FOR CABG EVENTS (FOR RISK FACTOR CODINGS SEE TABLE I)

Rule	SEX	AGE	SMBEF	SBP	HT	DM	TC	TG	GLU	FH	CABG event	# of cases	Support	Confidence
Rules with 3 risk factors														
1.1	F		N	O							Y	16	6%	80%
Rules with 4 risk factors														
2.1		4	Y	O		N					Y	12	5%	100%
2.2		4		H			D	O			Y	10	4%	91%
2.3		3	N	O						N	Y	10	4%	83%
Rules with 5 risk factors														
3.1	M	4				N				N	Y	33	13%	73%
3.2	M	4				N		O		Y	Y	10	4%	71%
3.3			Y	H	Y			H		Y	Y	6	2%	60%

- Non-smoker women with high SBP have PCI (no. of cases = 9, support = 4%, confidence = 90%) [Table IV, rule 2.1].

- Female non-smokers with high systolic blood pressure had a higher risk for PCI than male within the same group (no. of cases = 9, and 10, support = 4%, and 4%, confidence = 90%, and 71%, respectively) [Table IV, rules 2.1 and 3.2].

C. CABG model

The overall percent of correct classifications score for the CABG model was 57%, with 66%, 55%, and 63% for TP Rate, FP Rate, and PR, respectively. Detailed rule analysis for this model is given in Table V. More specifically, the following rules can be extracted:

- Non-smoker women with high SBP have a confidence level of 80% (with the no. of cases = 16, support = 6%,) for the occurrence of a CABG event [Table V, rule 1.1].

- The majority of patients who had CABG belong to the age group 71 – 85 [Table V, rules 2.1, 2.2, 3.1, and 3.2; with no. of cases = 12, 10, 33, and 10, support = 5%, 4%, 13%, and 4%, confidence = 100, 91%, 73%, and 71%, respectively].

- In the men group aged 71-85 years old that do not have history of diabetes, family history does not play a role for CABG (no. of cases = 33, and 10, support = 13%, and 4%, and confidence = 73%, and 71%, respectively) [Table V, rules 3.1 and 3.2].

- Our data indicate that triglyceride levels play a very important role as a risk factor for CHD [Table III, rule 3.3 with no. of cases = 6, support = 2%, and confidence = 60%; Table V, rules 2.2 and 3.2, with no. of cases = 10, and 10, support = 4%, and 4%, and confidence = 91%, and 71%, respectively].

IV. CONCLUDING REMARKS

In this study a data mining system for the assessment of heart event related risk factors was carried out. The events investigated were: MI, PCI, and CABG. Data mining analysis was carried out using the C4.5 decision trees algorithms. The most important risk factors, as extracted from the classification rules analysis were: sex, age, smoking, blood pressure, and cholesterol. It should be noted that the latter three risk factors could be modified; therefore the CHD risk of a patient may be reduced through a proper control of these factors. Furthermore, the importance of smoking in CHD risk was clearly illustrated. These findings and risk factors were also extracted by other investigators [23].

It is anticipated that data mining could help in the identification of high and low risk subgroups of patients, a decisive factor for the selection of therapy, i.e. medical or surgical. Moreover, the extracted rules could help to reduce CHD morbidity and possibly, mortality. However, further investigation with larger data sets and other rule extraction algorithms and criteria are still needed.

REFERENCES

[1] T. Marshall, "Identification of patients for clinical risk assessment by prediction of cardiovascular risk using default risk factor values", *BMC Public Health*, 8: 25, 2008.

[2] W.B. Kannel, "Contributions of the Framingham Study to the conquest of coronary artery disease", *Am. J. Cardiol.*, 62:1109–1112, 1988.

[3] N.J. Andrew, F.M. Fesmire, D. Sonnemaker, "Mining Bayesian Networks to Forecast Adverse Outcomes Related to Acute Coronary Syndrome", in *Proc. of the 17th international FLAIRS conference*, pp. 1-6, 2004.

[4] K. Matoušek and P. Aubrecht, "Data Modelling and Pre-processing for Efficient Data Mining in Cardiology", in *Proc. IEEE ITAB*, 2006.

[5] Euroaspire study group, "A European Society of Cardiology survey of secondary prevention of coronary heart disease: Principal results", *European Heart Journal*, vol. 18, pp. 1569-1582, 1997.

[6] D. Gamberger, R. Bošković Institute, Zarageb, Croatia, "Medical prevention: Targeting high-risk groups for coronary heart disease", Sol-EU-Net: Data Mining and Decision Support, [Online]. Available: http://soleunet.ijs.si/website/other/case_solutions/CHD.pdf.

[7] C. Ordonez *et al.*, "Mining Constrained Association Rules to Predict Heart Disease", in *Proc. International Conference on Data Mining, ICDM, IEEE*, pp. 431-440, 2001.

[8] University of Waikato, Hamilton, New Zealand, "Data Mining Analysis (request cleveland-14-heart-disease)", AKM Server, May 3, 2003, [Online]. Available: http://www.auknomi.com/webFiles/cleveland-14-heart-disease.pdf.

[9] J.C. Prather, D.F. Lobach, L.K Goodwin, J.W. Hales, M.L. Hage, W.E. Hammond, "Medical data mining: knowledge discovery in a clinical data warehouse", in *Proc. of the AMIA Annual Fall Symposium*, pp. 101-105, 1997.

[10] P.W. Wilson, R.B. D'Agostino, D. Levy, A.M. Belanger, H. Silbershatz, W.B.Kannel, "Prediction of coronary heart disease using risk factor categories**",** *Circulation*; pp. 1837-1847, 1998.

[11] N. Allahverdi, S. Torun, I. Saritas, " Design of a Fuzzy expert system for determination of coronary Heart Disease risk", *ACM International Conference Proceeding Series*; vol. 285: 36, 2007.

[12] T.D. Rea *et al.*, "Smoking Status and Risk for Recurrent Coronary Events after Myocardial Infraction", *Ann Intern Med.*, vol. 137, pp. 494-500, 2002.

[13] Euroaspire II Study Group, "Lifestyle and risk factor management and use of drug therapies in coronary patients from 15 countries", *European Heart Journal*, vol. 22, pp. 554-572, 2002.

[14] Z. Wang and W.E. Hoy, "Is the Framingham coronary heart disease absolute risk function applicable to aboriginal people?", *The Medical Journal of Australia*, vol. 2, pp. 66-69, 2005.

[15] J. Li *et al.*, "Mining Risk Patterns in Medical Data", in *Proc. of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 770-775, 2005.

[16] S.K.Wasan *et al.*, "The impact of data mining techniques on medical diagnostics", *Data Science Journal*, vol. 5, pp. 119-126, 2006.

[17] Y.-T. Kuo *et al.*, "Domain Ontology Driven Data Mining, A Medical Case Study", *ACM SIGKDD Workshop on Domain Driven Data Mining*, pp. 11-17, 2007.

[18] C. Ordonez, "Comparing Association Rules and Decision Trees for Disease Prediction", *HIKM'06*, 2006.

[19] J.R. Quinlan, *"Boosting first order learning,"* in Algorithmic Learning Theory, Springer-Verlag, pp. 143-55, 1996.

[20] J.R. Quinlan, "Bagging, boosting, and C4.5," in *Proc. of the 13th American Assoc. for AI National Conference on Artificial Intelligence*, pp. 725-30, AAAI Press, Menlo Park, CA, 1996.

[21] I.H. Witten, E. Frank, *Data Mining – Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers, 2005.

[22] J. Han and M. Kamber, *Data Mining, Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.

[23] TA. Pearson *et al.*, "AHA guidelines for primary prevention of cardiovascular disease and stroke", *Circulation*, 106(3): 388–391, 2002.