31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Association Rule Analysis for the Assessment of the Risk of Coronary Heart Events

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

Abstract -- **Although significant progress has been made in the diagnosis and treatment of coronary heart disease (CHD), further investigation is still needed. The objective of this study was to develop a data mining system using association analysis based on the apriori algorithm for the assessment of heart event related risk factors. The events investigated were: myocardial infarction (MI), percutaneous coronary intervention (PCI), and coronary artery bypass graft surgery (CABG). A total of 369 cases were collected from the Paphos CHD Survey, most of them with more than one event. The most important risk factors, as extracted from the association rule analysis were: sex (male), smoking, high density lipoprotein, glucose, family history, and history of hypertension. Most of these risk factors were also extracted by our group in a previous study using the C4.5 decision tree algorithms, and by other investigators. Further investigation with larger data sets is still needed to verify these findings.**

I. Introduction

In coronary heart disease (CHD), the coronary arteries that supply the heart muscle with oxygen and nutrients become narrowed by atherosclerotic stenotic lesions. This restricts the supply of blood and oxygen to the heart, particularly during exertion when the myocardial metabolic demands are increased [1].

Extensive clinical and statistical studies have identified several factors that increase the risk of coronary heart disease including acute myocardial infarction [2], [3]. The more risk factors one might have, the greater the risk of developing coronary heart disease. Also, the greater the severity of each risk factor, the greater the overall risk. 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 [4]. The first category includes factors that cannot be altered by intervention such as age, gender, 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, elevated cholesterol and triglycerides, elevated LDL and low HDL, hypertension, and diabetes [2], [5]. There is 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.

The objective of this study was to develop a data mining system for the assessment of CHD related risk factors using the apriori algorithm for extracting rules. A previous study by our group on the same dataset showed that important risk factors could be modified [6]; therefore the risk of CHD of a patient may be reduced through a proper control of these factors as it has already been published by several very important studies, including the EUROASPIRE I, II, and III surveys [7]-[10].

The first and second EUROASPIRE surveys showed high rates of modifiable cardiovascular risk factors in patients with coronary heart disease, and indicated that preventive measures might decrease cardiovascular risk [8], [9]. The third EUROASPIRE survey that investigates the situation in Europe 10 years later (that was done in 2006—07 in 22 countries) to see whether preventive cardiology had improved showed that the major risk factors (smoking, hypertension, and obesity) have not decreased [10]. It is interpreted in the Euroaspire III study that despite a substantial increase in antihypertensive and lipid-lowering drugs, blood pressure management remained unchanged, and almost half of all patients remain above the recommended lipid targets, reflecting a natural reluctance for people to change their lifestyles [10].

Data mining was also employed in several studies, where different algorithms were used for rule extraction and evaluation like the C4.5 decision trees [6], [11] and the Apriori [12], [13] algorithms.

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.A. Moutiris) 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 (see also Table I): i. Clinical factors: Age, Sex, Smoking (SMBEF), systolic blood pressure (SBP) mmHg, diastolic blood pressure (DBP) mmHg, history of hypertension (HT),

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

J.A. Moutiris, is a cardiologist at the Department of Cardiology, Paphos General Hospital, Paphos, Cyprus and coordinator of the Paphos CHD Survey (email: moutiris@ucy.ac.cy).

family history (FH), and Diabetes (DM); ii. Biochemical factors: Cholesterol (TC) mg/dL, high density lipoprotein (HDL) mg/dL, low density lipoprotein (LDL) 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 369, mainly due to the unavailability of biochemical results. The number for MI cases was 265, for PCI 160, and for CABG 152. 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 [7], [14].

TABLE I
RISK FACTORS WITH THEIR CORRESPONDING CODINGS

	Risk Factor	Code 1	Code 2	Code 3	Code 4
Clinical factors					
1	AGE	1: 34-50	2: 51-60	3:61-70	4: 71-85
2	SEX	M: MALE	F:FEMALE		
3	SMBEF	Y: YES	N: NO		
4	SBP*	L<90	N:90-120	H>120	
5	DBP *	L<60	N:60-80	H>80	
6	FH	Y: YES	N: NO		
7	HT	Y: YES	N: NO		
8	DM	Y: YES	N: NO		
Biochemical factors					
9	TC **	L <200	N:201 –240	H>240	
10	HDL**				
	Women	L<50	M:50-60	H>60	
	Men	L<40	M:40-60	H>60	
11	LDL**	N<130	H:131-160	D>160	
12	TG**	N<150	H:151-200	D>200	
13	GLU**	H>110	N <110		

L: Low, N: Normal, H: High, D: Dangerous
* in mmHg ** in mg/dL

D. Association Rule Analysis Using Apriori Algorithm

Apriori is a classic algorithm for learning association rules. The task in Association Rules mining involves finding all rules that satisfy user defined constraints on minimum *support* and *confidence* with respect to a given dataset.

The Apriori algorithm searches for large itemsets during its initial database pass and uses its result as the seed for discovering other large datasets during subsequent passes. Rules having a support level above the minimum are called large or frequent itemsets and those below are called small itemsets. The algorithm is based on the large itemset property which states: Any subset of a large itemset is large and if an itemset is not large and then none of its supersets are large [15]. The Weka's implementation of this algorithm was ran [16].

E. Pattern Evaluation and Knowledge Representation

The following three different set of runs for association analysis 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 runs, the steps were carried out for data mining association and pattern evaluation. Rules were extracted from different combinations of risk factors. A minimum of one to a maximum of 13 risk factors were extracted from the different rules.

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. Since similar support and confidence values were achieved, we used another measure, the *distance*, to give us the most reliable rules (see next section for definition).

F. Performance Measures

Hold-out validation was used for evaluating the performance of the proposed runs. The data were split into training and testing partitions representing 70% and 30% of the records respectively. This procedure was repeated five times. It is noted that the extracted rules derived from the different sets were very similar.

In order to evaluate the performance of our results we used the following measures [15]:

- *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 record contains {X, Y, Z} [15].
- *Confidence:* is the number of cases for which the rule applies (or predicts correctly), expressed as a percentage of all instances to which it applies (i.e. if we have the rule X & Y → Z, Confidence is the conditional probability that a record having {X, Y} also contains Z [15].
- *Distance:* is the absolute difference between the training and testing confidence for a rule. It is a figure of merit that as if approaches to very small values or to zero shows the greatest coverage and reliability of a rule (as it covers both the training and testing sets).

III. RESULTS

Three different set of runs were investigated, for extracting rules: (i) MI versus PCI or CABG, (ii) PCI versus MI or CABG, and (iii) CABG versus MI or PCI. The corresponding rules for these runs are given in Table II. Five different runs were carried out for each of the above sets, where similar performance was obtained.

A. MI Events

Rules 1.1 – 1.13 for this run are given in Table II. More specifically, the following rules can be extracted:
Rules 1.1-1.5 cover male smoker patients. Nearly all patients with MI are men (rules 1.2, 1.3). Rules 1.4, and 1.5 show that factors high systolic blood pressure and history of hypertension have similar importance. Seventy two percent of men with high glucose levels are smokers (rules 1.7, 1.9). Fifty four out of 84 patients with high levels of glucose have high blood pressure (rules 1.6, 1.8). Ninety percent of men with high blood pressure have history of hypertension (rules 1.11, 1.13). Seventy percent of patients with abnormal HDL levels have history of hypertension (rules 1.10, 1.12).

B. PCI Events

Rules 2.1 – 2.17 for this run are given in Table II. More specifically, the following rules can be extracted: 2.1 to 2.4 show rules with high systolic blood pressure. Twenty six out of 36 patients with high blood pressure, family history and history of hypertension, are smokers (rules 2.4, 2.5). Nearly all smoker patients with history of hypertension and a positive family history are male (rules 2.6, 2.7). Sixty eight percent of patients in the age range of 61 to 70 years old have history of hypertension (rules 2.8, 2.10). Only 7.7% of patients in the age range of 51 to 60 years old are non-smokers (rules 2.9, 2.11). Fifty percent of patients with high blood pressure and history of hypertension have abnormal HDL levels (rules 2.12, 2.13). Also men smokers have abnormal HDL levels (rules 2.14, 2.15). Fifty nine percent of men smokers with high blood pressure have a positive family history (rules 2.16, 2.17).

C. CABG Events

Rules 3.1 – 3.15 for this run are given in Table II. More specifically, the following rules can be extracted: 3.1 to 3.7 show rules with high systolic blood pressure and history of hypertension. Thirty two of 39 men smokers with systolic blood pressure have history of hypertension (rules 3.7, 3.12). Rules 3.8 and 3.10 show that sex does not play an important role for smoker patients with high glucose levels and high blood pressure since 26 out of 28 cases are male. Ninety three percent of patients that are smokers a have history of hypertension are male (rules 3.9, 3.11). Rules 3.13 to 3.15 show that 34 out of 41 patients are male and 41 out of 89 patients are in the age range of 61 to 70 years old. Nearly all smoker patients with high blood pressure and history of hypertension are men (rules 3.4, 3.5) and comparing them with rules 3.7 and 3.12 we observe that if cholesterol and LDL levels are within normal range then there is a 13% decrease of the event.

Considering results for all events we observed that smoking is one of the main risk factors that directly affect the coronary heart disease events. Rea et al. [17] also concluded that smoking had an increase effect for recurrent coronary events.

A thorough investigation of the association rules extracted in this study is still needed in order to verify the importance of our findings in the clinical practice.

IV. Concluding Remarks

In this study, a data mining system for the assessment of heart event related risk factors was carried out using association analysis based on the apriori algorithm. The events investigated were: MI, PCI, and CABG. Rules with risk factors like sex (male), smoking, high density lipoprotein, glucose, family history, and history of hypertension, were extracted. The modifiable risk factors can be monitored / lowered with the doctor's advice and medications so that the incidence of heart episodes can be lowered. 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.

References

[1] Heart disease and cardiovascular disorders, http://heart-disease.health-cares.net/, last access May 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] 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.

[4] S.Yusuf et al, "Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study" *Lancet*, Vol. 364, Issue 9438, pp. 937-952, 2004.

[5] 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.

[6] M. Karaolis, J.A. Moutiris, C.S. Pattichis, "Assessment of the risk of coronary heart event based on data mining", in 8th IEEE International Conference on BioInformatics and BioEngineering, BIBE 2008, Issue , pp.:1 – 5, 2008.

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

[8] 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.

[9] 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.

[10] K. Kotseva, D. Wood, G. De Backer, D. De Bacquer, K. Pyörälä, U. Keil, "Cardiovascular prevention guidelines in daily practice: a comparison of EUROASPIRE I, II, and III surveys in eight European countries", The Lancet, Vol. 373, Iss. 9667, pp. 929 - 940, 2009.

[11] Z. Zhou and Y. Jiang. "Medical diagnosis with c4.5 rule preceded by artificial neural network ensemble". *IEEE Transactions on Information Technology in Biomedicine*, pp. 37–42, 2003.

[12] 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.

[13] M. J. Zaki. "Mining non-redundant association rules". *Data Mining and Knowledge Discovery Journal*, pp.: 223–248, 2004.

[14] European Society of Cardiology (ESC), http://www.escardio.org/Pages/index.aspx/, last access May 2008.

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

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

[17] 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.

TABLE II
EXTRACTED RULES FOR MI (RULES NO 1.1-1.18), PCI (RULES NO 2.1-2.17) AND CABG (RULES NO 3.1-3.15) EVENTS
(FOR RISK FACTOR CODINGS SEE TABLE I)

Rule no.	AGE	SEX	SMBEF	SBP	FH	HT	TC	HDL	LDL	GLU	Class	Num Of Attributes	Support Count	Support	Confidence	Testing Confidence	Distance
Extracted Rules for MI Events																	
Rules with risk factor sex and smoking, SEX = M, SMBEF=Y																	
1.1		M									MI	1	159	0,6	0,72	0,74	0,02
1.2			Y								MI	1	120	0,5	0,71	0,77	0,06
1.3		M	Y								MI	2	118	0,5	0,71	0,76	0,05
1.4		M	Y			Y					MI	3	63	0,2	0,66	0,72	0,06
1.5		M	Y	H							MI	3	64	0,3	0,63	0,7	0,07
Rules with risk factor glucose, GLU = H																	
1.6										H	MI	1	84	0,3	0,69	0,72	0,03
1.7		M								H	MI	2	72	0,3	0,69	0,73	0,04
1.8				H						H	MI	2	54	0,2	0,64	0,67	0,03
1.9		M	Y							H	MI	3	52	0,2	0,65	0,76	0,11
Rules with risk factors systolic blood pressure, history of hypertension, and high density lipoprotein, SBP=H, HT=Y, HDL=L																	
1.10								L			MI	1	76	0,3	0,7	0,78	0,08
1.11		M		H							MI	2	88	0,3	0,66	0,68	0,02
1.12						Y		L			MI	2	53	0,2	0,71	0,75	0,04
1.13		M		H		Y					MI	3	61	0,2	0,66	0,67	0,01
Extracted Rules for PCI Events																	
Rules with risk factor systolic blood pressure, SBP = H																	
2.1				H							PCI	1	75	0,3	0,47	0,45	0,02
2.2			Y	H							PCI	2	51	0,2	0,47	0,45	0,02
2.3			Y	H		Y					PCI	3	40	0,2	0,45	0,32	0,13
2.4			Y	H	Y	Y					PCI	4	26	0,1	0,59	0,23	0,36
Rules risk factors smoking, family history, and history of hypertension SMBEF=Y , FH=Y, HT=Y																	
2.5				H	Y	Y					PCI	3	36	0,1	0,55	0,37	0,18
2.6			Y		Y	Y					PCI	3	31	0,1	0,58	0,35	0,23
2.7		M	Y		Y	Y					PCI	4	30	0,1	0,58	0,35	0,23
Rules with risk factor AGE=2 and 3																	
2.8	3										PCI	1	41	0,2	0,4	0,5	0,1
2.9	2										PCI	1	39	0,2	0,56	0,42	0,14
2.10	3					Y					PCI	2	28	0,1	0,38	0,45	0,07
2.11	2		Y								PCI	2	36	0,1	0,65	0,5	0,15
2.12				H		Y					PCI	2	59	0,2	0,45	0,36	0,09
2.13				H		Y		L			PCI	3	28	0,1	0,44	0,36	0,08
Rules with risk factors sex and smoking, SEX=M, SMBEF=Y																	
2.14		M	Y								PCI	2	76	0,3	0,45	0,47	0,02
2.15		M	Y					L			PCI	3	34	0,1	0,43	0,48	0,05
2.16		M	Y	H							PCI	3	49	0,2	0,46	0,46	0
2.17		M	Y	H	Y						PCI	4	29	0,1	0,6	0,33	0,27
Extracted Rules for CABG Events																	
Rules with risk factors systolic blood pressure, and history of hypertension, SBP = H, HT=Y																	
3.1				H							CABG	1	67	0,3	0,43	0,47	0,04
3.2						Y					CABG	1	71	0,3	0,44	0,47	0,03
3.3				H		Y					CABG	2	52	0,1	0,48	0,42	0,06
3.4			Y	H		Y					CABG	3	39	0,2	0,48	0,39	0,09
3.5		M	Y	H		Y					CABG	4	37	0,1	0,48	0,39	0,09
3.6		M	Y	H		Y			N		CABG	5	32	0,1	0,57	0,41	0,16
3.7		M	Y	H		Y	L		N		CABG	6	32	0,1	0,6	0,44	0,16
Rules with risk factor smoking, SMBEF=Y																	
3.8			Y	H						H	CABG	3	28	0,1	0,47	0,42	0,05
3.9			Y			Y				H	CABG	3	30	0,1	0,52	0,5	0,02
3.10		M	Y	H						H	CABG	4	26	0,1	0,46	0,42	0,04
3.11		M	Y			Y				H	CABG	4	28	0,1	0,51	0,5	0,01
3.12		M	Y	H			L		N		CABG	5	39	0,2	0,56	0,42	0,14
Rules with risk factors age and sex, AGE=3, SEX=M																	
3.13	3										CABG	1	41	0,2	0,41	0,46	0,05
3.14		M									CABG	1	89	0,3	0,4	0,43	0,03
3.15	3	M									CABG	2	34	0,1	0,41	0,44	0,03