Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees

Minas A. Karaolis, *Member, IEEE*, Joseph A. Moutiris, Demetra Hadjipanayi, and Constantinos S. Pattichis, *Senior Member, IEEE*

***Abstract*—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. 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 targeting in the reduction of CHD events. The risk factors investigated were: 1) before the event: a) nonmodifiable—age, sex, and family history for premature CHD, b) modifiable—smoking before the event, history of hypertension, and history of diabetes; and 2) after the event: modifiable—smoking after the event, systolic blood pressure, diastolic blood pressure, total cholesterol, high-density lipoprotein, low-density lipoprotein, triglycerides, and glucose. The events investigated were: myocardial infarction (MI), percutaneous coronary intervention (PCI), and coronary artery bypass graft surgery (CABG). A total of 528 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 tree algorithm for the aforementioned three events using five different splitting criteria. The most important risk factors, as extracted from the classification rules analysis were: 1) for MI, age, smoking, and history of hypertension; 2) for PCI, family history, history of hypertension, and history of diabetes; and 3) for CABG, age, history of hypertension, and smoking. Most of these risk factors were also extracted by other investigators. The highest percentages of correct classifications achieved were 66%, 75%, and 75% for the MI, PCI, and CABG models, respectively. It is anticipated that data mining could help in the identification of high and low risk subgroups of subjects, a decisive factor for the selection of therapy, i.e., medical or surgical. However, further investigation with larger datasets is still needed.**

***Index Terms*—Coronary heart disease (CHD), data mining, decision trees, risk factors.**

I. Introduction

CORONARY heart disease (CHD) is the single most common cause of death in Europe, responsible for nearly two million deaths a year [1]. Advances in the field of medicine over the past few decades enabled the identification of risk factors that may contribute toward 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 [2]–[4]. 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, hypertension, diabetes, cholesterol, high-density lipoprotein, low-density lipoprotein, triglycerides [5], [6].

The objective of this study was to develop a data mining system based on decision trees for the assessment of CHD-related risk factors targeting in the reduction of CHD events. Data-mining analysis was carried out using the C4.5 decision tree algorithm using five different splitting criteria for extracting rules based on the aforementioned risk factors. Preliminary results of this study were previously published [7].

Many studies have been carried out investigating CHD and related risk factors. Some of them used the Framingham equation to describe the population in a region or country [8], [9], whereas other studies examined the features of available Framingham-based risk calculation [10]. The American Heart Association (AHA) assessed multiple risk factors and also developed new guidelines for CHD [11], [12]. Furthermore, results from the European Action on Secondary and Primary Prevention by Intervention to Reduce Events (EUROASPIRE) revealed the important risk factors through various surveys across European countries [2]–[4].

Data mining facilitates data exploration using data analysis methods with sophisticated algorithms in order to discover unknown patterns. Such algorithms include decision trees that have been used extensively in medicine. According to Podgorelec *et al.* [13] decision-tree-based algorithms give reliable and effective results that provide high-classification accuracy with a simple representation of gathered knowledge, and are especially appropriate to support decision-making processes in medicine.

Several studies have been carried out that investigated the usefulness of decision tree models in CHD-related problems. Ordonez [14], [15] investigated decision trees and association rules to predict CHD based on the risk factors sex, smoking, cholesterol, and age. Gamberger *et al.* [16] used a decision support method to target high-risk groups for CHD using risk factors like smoking, cholesterol and hypertension. Tsien

Manuscript received August 5, 2009; revised November 24, 2009. First published January 12, 2010; current version published June 3, 2010.

M. A. Karaolis, D. Hadjipanayi, and C. S. Pattichis are with the Department of Computer Science, University of Cyprus, Nicosia 1678, Cyprus (e-mail: karaolis@spidernet.com.cy; demetra.hadjipanayi@gmail.com; pattichi@ucy.ac.cy).

J. A. Moutiris is with the Department of Cardiology, Paphos General Hospital, Paphos 8100, Cyprus (e-mail: moutiris@ucy.ac.cy).

Digital Object Identifier 10.1109/TITB.2009.2038906

et al. [17] used also classification trees and logistic regression building three different models for myocardial infarction (MI) and examining also the significance of these models. Decision-tree-based software tools were developed in [18] and [19] to aid in the diagnosis of CHD. Rao *et al.* [18] presented a framework to create structured clinical data for CHD. Zavrsnik *et al.* [19] used decision trees and created the ROSE tool for use in cardiology. Furthermore, Polat *et al.* [20] developed decision tree based models for the classification of CHD, achieving a correct classification score of 82%. Moreover, Pavlopoulos *et al.* [21] used the C4.5 algorithm decision trees to analyze the different heart sound features, which assist clinicians to make a better diagnosis in CHD.

Several other studies investigated different technologies for the assessment of CHD, including logistic regression [17], association rules [7], [15], fuzzy modeling [22], [23], neural networks [24], and other.

In this study, we investigate how data mining based on decision trees can help for the evaluation of the risk of CHD. The aim is to identify the most important risk factors based on the classification rules to be extracted. These rules will enable the better management of the patient targeting in the reduction of events, as well as, reduction of the cost of therapy, due to the expected restriction of interventions in necessary cases only.

The rest of the paper is organized as follows. Section II describes the material and methods, section III the results and Section IV the discussion.

TABLE I
CODING OF RISK FACTORS

	Risk Factor	Code 1	Code 2	Code 3	Code 4
Risk Factors Before The Event: **non modifiable**					
1	AGE	1: 34-50	2: 51-60	3:61-70	4: 71-85
2	SEX	M: MALE	F:FEMALE		
3	FH	Y: YES	N: NO		
Risk Factors Before The Event: **modifiable**					
4	SMBEF	Y: YES	N: NO		
5	HχHTN	Y: YES	N: NO		
6	HχDM	Y: YES	N: NO		
Risk Factors After The Event: **modifiable**					
1	SMAFT	Y: YES	N: NO		
2	SBP (mmHg)	L<100	N:100-130	H>=130	
3	DBP (mmHg)	L<60	N:60-85	H>=85	
4	TC (mg/dL)	N<190	H>=190		
5	HDL (mg/dL)				
	Women	L<50	N:50-60	H>=60	
	Men	L<40	N:40-60	H>=60	
6	LDL (mg/dL)	N<100	H>=100		
7	TG (mg/dL)	N<150	H>=150		
8	GLU (mg/dL)	N <110	H>=110		

L: low; N: normal, H: high, D: dangerous.

TABLE II
NO. OF CASES PER SET OF RULES/MODELS INVESTIGATED

	Model	MI	PCI	CABG
		N/Tr/Ev	N/Tr/Ev	N/Tr/Ev
Event	Yes	378/75/75	72/36/36	86/43/43
	No	150/75/75	274/36/36	307/43/43
	Total	528/150/150	346/72/72	392/86/86

N: total no. of cases, Tr and Ev give the number of cases in training and evaluation, respectively.

II. MATERIALS AND METHODS

A. Data Collection, Cleaning, and Coding

Data from 1500 consecutive CHD subjects were collected between the years 2003–2006 and 2009 (300 subjects each year) according to a prespecified protocol, under the supervision of the participating cardiologist (Dr. J. Moutiris, second author of this paper) at the Department of Cardiology, at the Paphos General Hospital in Cyprus. Subjects had at least one of the following criteria on enrollment, history of MI, or percutaneous coronary intervention (PCI), or coronary artery bypass graft surgery (CABG). Data for each subject were collected as given in Table I: 1) risk factors before the event, a) nonmodifiable—age, sex, and family history (FH); 2) modifiable—smoking before the event (SMBEF), history of hypertension (HxHTN), and history of diabetes (HxDM); and 2) risk factors after the event, modifiable—smoking after the event (SMAFT), systolic blood pressure (SBP) in mmHg, diastolic blood pressure (DBP) in mmHg, total cholesterol (TC) in mg/dL, high-density lipoprotein (HDL) in mg/dL, low-density lipoprotein (LDL) in mg/dL, triglycerides (TG) in mg/dL, and glucose (GLU) in mg/dL.

To clean the data, the fields were identified, duplications were extracted, missing values were filled, and the data were coded as given in Table I. After data cleaning, the number of cases was reduced as given in Table II, mainly due to unavailability of biochemical results.

B. Classification by Decision Trees

The C4.5 algorithm [25], which uses the divide-and-conquer approach to decision tree induction, was employed. The algorithm uses a selected criterion to build the tree. 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.

Algorithm Generate Decision Tree [25], [26]:
Input:

1) Training dataset D, which is a set of training observations and their associated class value.
2) Attribute list A, the set of candidate attributes.
3) Selected splitting criteria method.

Output: A decision tree.
Method:

1) Create a node *Nd*.

2) If all observations in the training dataset have the same class output value C, then return Nd as a leaf node labeled with C.
3) If attribute list is empty, then return Nd as leaf node labeled with majority class output value in training dataset.
4) Apply selected splitting criteria method to training dataset in order to find the ''best" splitting criterion attribute.
5) Label node Nd with the splitting criterion attribute.
6) Remove the splitting criterion attribute from the attribute list.
7) For each value j in the splitting criterion attribute.
   a) Let D_j be the observations in training dataset satisfying attribute value j.
   b) If D_j is empty (no observations), then attach a leaf node labeled with the majority class output value to node Nd.
   c) Else attach the node returned by generate decision tree (D_j, attribute list, selected splitting criteria method) to node Nd.
8) End for.
9) Return node Nd.

In this study, the following splitting criteria were investigated that are briefly presented shortly: information gain, gini index, likelihood ratio chi-squared statistics, gain ratio, and distance measure.

1) Information Gain (IG): Information gain is based on Claude Shannon's work on information theory. InfoGain of an attribute A is used to select the best splitting criterion attribute. The highest InfoGain is selected to build the decision tree [27]

$$\mathrm{InfoGain}(A) = \mathrm{Info}\,(D) - \mathrm{Info}_A\,(D) \tag{2.1}$$

where A is the attribute investigated.

$$\mathrm{Info}\,(D) = -\sum_{i=1}^{m} p_i \log_2\,(p_i) \tag{2.2}$$

where

$p_i = \mathrm{probability}(\mathrm{class}\, i \,\mathrm{in\ dataset}\, D);$
$m = \mathrm{number\ of\ class\ values}.$

$$\mathrm{Info}_A\,(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \mathrm{Info}\,(D_j) \tag{2.3}$$

where

$|D_j|$ = number of observations with attribute value j in dataset D;

$|D|$ = total number of observations in dataset D;

D_j = sub dataset of D that contains attribute value j;

v = all attribute values.

Although information gain is usually a good measure for deciding the relevance of an attribute, it is not perfect. A problem occurs when information gain is applied to attributes that can take on a large number of distinct values. When that happens, then gain ratio is used instead.

2) Gini Index (GI): The Gini index is an impurity-based criterion that measures the divergence between the probability distributions of the target attributes values [28]

$$\mathrm{GiniIndex}(D) = \mathrm{Gini}(D) - \sum_{j=1}^{v} p_j \times \mathrm{Gini}(D_j) \tag{2.4}$$

$$\mathrm{Gini}\,(D) = 1 - \sum_{i=1}^{m} p_i^2. \tag{2.5}$$

3) Likelihood Ratio Chi-Squared Statistics (χ^2): The likelihood ratio chi-squared statistic is useful for measuring the statistical significance of the information gain criterion [29]

$$G^2\,(A, D) = 2 \times \ln\,(2) \times |D| \times \mathrm{InfoGain}\,(A)\,. \tag{2.6}$$

4) Gain Ratio (GR): Gain ratio biases the decision tree against considering attributes with a large number of distinct values. So it solves the drawback of information gain [25]

$$\mathrm{GainRatio}\,(A) = \frac{\mathrm{InfoGain}(A)}{\mathrm{SplitInfo}_A\,(D)} \tag{2.7}$$

$$\mathrm{SplitInfo}_A\,(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right). \tag{2.8}$$

5) Distance Measure (DM): Distance measure, like GR, normalizes the impurity criterion (GI). It suggests normalizing it in a different way [30]

$$\mathrm{DM}\,(A) = \frac{\mathrm{Gini}\,(D)}{-\sum_{j=1}^{v}\sum_{i=1}^{m} p_{ij} \times \log_2\,(p_{ij})}. \tag{2.9}$$

A data-mining tool was developed by our group that supports the C4.5 decision tree algorithm using the aforementioned criteria. Overfitting is a significant practical difficulty for decision tree learning. Therefore, pruning is implemented to avoid overfitting. We implemented the bottom-up pruning algorithm using Laplace error estimation. While the decision tree is built and a leaf node is created, then the Laplace error [31] is estimated as follows:

$$E\,(D) = \frac{N - n + m - 1}{N + m} \tag{2.10}$$

where

$C = \mathrm{class\ value\ majority\ class\ in}\ D;$

$N = \mathrm{number\ of\ observations\ in}\ D;$

$n = \mathrm{number\ of\ observations\ has\ class\ value}\ C.$

As the algorithm returns to the root node, the error of the leaf node is passed to the father node. The father node calculates the total error of all of its children and its own error. If the father's error is less than the total error of the children, then the father node is pruned and replaced by a leaf node with the majority class value. If the father's error is greater than the total error of the children, then no more pruning is done to the path and the returned error is zero.

C. Classification Models Investigated

The following sets of models were investigated as given in Table II.

1) MI: MI versus non-MI. Subjects having myocardial infraction were marked as symptomatic and the rest as asymptomatic.
2) PCI: PCI versus non-PCI. Subjects having only PCI were marked as symptomatic and the rest as asymptomatic. Subjects having both PCI and MI were excluded.
3) CABG: CABG versus non-CABG. Subjects having only CABG were marked as symptomatic and the rest as asymptomatic. Subjects having both CABG and MI were excluded.

For each set of models, three different subsets of runs were carried out as given in the following:

1) with risk factors before the event (B);
2) with risk factors after the event (A); and
3) with risk factors before and after the event (B + A).

For each model, for each splitting criterion, 20 runs were carried out with random sampling [32] of equal number of cases used for training and evaluation as given in Table II. A total of 300 runs were carried out for each set of models [i.e., 20 runs $\times$ 5 splitting criteria $\times$ 3 (for B, A, and B +A datasets)].

The Wilcoxon rank sum test [33] was also carried out to investigate if there was or not significant difference between the five splitting criteria used as well as between the B, A, and B + A decision tree models at $p < 0.05$.

D. Performance Measures

In order to evaluate the performance of our results we used the following measures [34].

1) *Correct classifications* (%CC): is the percentage of the correctly classified records; equals to (TP + TN)/*N*.
2) *True positive rate* (%TP): corresponds to the number of positive examples correctly predicted by the classification model.
3) *False positive rate* (%FP): corresponds to the number of negative examples wrongly predicted as positive by the classification model.
4) *True negative rate* (%TN): corresponds to the number of negative examples correctly predicted by the classification model.
5) *False negative rate* (%FN): corresponds to the number of positive examples wrongly predicted as negative by the classification model.
6) *Sensitivity*: is defined as the fraction of positive examples predicted correctly by the model, equals to TP/(TP + FN).
7) *Specificity*: is defined as the fraction of negative examples predicted correctly by the model, equals to TN/(TN + FP).
8) *Support*: is the number of cases for which the rule applies (or predicts correctly; i.e., if we have the rule $X \rightarrow Z$, Support is the probability that a transaction contains $\{X, Z\}$ [26]

$$\text{Support} = P(XZ) = \frac{\text{no of cases that satify } X \text{ and } Z}{|D|}.$$

9) *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 $X \rightarrow Z$, Confidence is the conditional probability that a transaction having *X* also contains *Z*) [26]

$$\text{Confidence} = P(Z|X) = \frac{P(XZ)}{P(X)}.$$

E. Calculation of the Risk

For each subject, we used the Framingham equation [8]–[10] to calculate the risk for an event to occur. We separated the subjects into two categories, those who have had an event and those who have not had an event. Then, for each extracted rule, we found out the subjects matching that rule and computed the average event risk per rule based on the risk value of each subject (see last two columns of Table V). It is noted that values of risk lower than 5%, between 5–10%, and higher than 10% classify a subject as low, intermediate, and high risk, respectively.

III. Results

Table III tabulates the classification results of the three set of models investigated for the five different splitting criteria using risk factors before the event (B), after the event (A), and before and after (B + A). The median (Me), minimum (*m*), and maximum (*M*) for 20 runs are given for %CC, %TP, and %FP, whereas for sensitivity and specificity only the median values are given. Table IV gives the three most important risk factors obtained from the classification decision tree models. Also, selected rules of the models obtained in Table III are given in Table V as well as the risk per rule computed using the Framingham equation.

A. MI Models

There was no significant difference for the different splitting criteria investigated for %CC using the Wilcoxon rank sum test at $p < 0.05$. As shown in Table III, comparable performance in the region of 60% for %CC was obtained for the B, A, and B + A risk factor models for all splitting criteria. Better performance was obtained for the B + A models, where the median of the %CC ranged from 62% to 63%, respectively. The best model was obtained when using the GI splitting criterion for the B + A risk factor codings with a maximum %CC = 66%.

The most important risk factors as given in Table IV were for the B models, age, history of hypertension, and smoking before the event, for the A models, systolic blood pressure, smoking after the event, and diastolic blood pressure, and for the B + A models, age, systolic blood pressure, smoking, and history of hypertension.

Based on the decision tree model, sample rules could be extracted. For example, as given in Table IV:

Rule 1.3 and 1.4:

1) The percentage of subjects aged 51–60 with history of hypertension who are non smokers and have event is

TABLE III
CLASSIFICATION RESULTS OF THE THREE SET OF MODELS INVESTIGATED FOR THE FIVE DIFFERENT SPLITTING CRITERIA USING RISK FACTORS BEFORE THE EVENT (B), AFTER THE EVENT (A), AND BEFORE AND AFTER (B+A)

	%CC			%TP			%FP			Sensitivity			Specificity		
	B	A	B+A	B	A	B+A	B	A	B+A	B	A	B+A	B	A	B+A
	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me	Me	Me	Me	Me	Me
MI															
IG	58(57,64)	61(60,63)	62(61,65)	64(60,76)	68(61,73)	67(53,68)	48(44,55)	45(41,49)	37(25,47)	58	60	63	60	64	63
GI	61(59,63)	61(59,63)	63(61,66)	67(55,71)	59(55,71)	63(57,76)	47(41,59)	36(33,48)	39(25,51)	59	60	62	61	62	64
X2	58(57,60)	61(59,63)	63(62,65)	65(63,73)	63(59,76)	64(59,72)	49(47,53)	39(35,59)	36(35,47)	57	62	64	59	61	64
GR	60(58,61)	59(59,59)	62(61,64)	65(53,72)	59(55,67)	65(53,67)	45(37,53)	41(36,49)	41(38,45)	59	59	62	61	59	62
DM	60(58,62)	59(58,62)	63(61,65)	71(57,67)	61(57,69)	65(57,71)	47(39,54)	43(40,45)	40(27,45)	59	59	65	63	59	64
PCI															
IG	63(61,65)	67(64,75)	67(65,70)	64(53,72)	72(67,78)	58(56,64)	36(31,42)	39(28,50)	22(22,31)	63	65	71	63	69	65
GI	61(61,64)	67(65,68)	67(63,70)	67(50,86)	69(50,75)	67(56,69)	39(28,64)	42(14,50)	31(22,42)	63	64	69	65	64	64
X2	63(60,64)	65(63,72)	65(63,65)	69(56,69)	72(58,78)	72(58,78)	36(33,44)	36(33,42)	42(28,53)	61	64	63	65	65	68
GR	63(61,70)	64(64,65)	65(64,67)	67(56,82)	67(53,83)	72(53,72)	44(31,50)	39(25,56)	39(22,44)	65	63	65	63	65	67
DM	64(63,65)	65(61,71)	65(64,68)	69(64,78)	72(67,78)	69(64,75)	42(33,47)	42(36,56)	39(33,47)	63	62	64	66	67	67
CABG															
IG	69(67,73)	66(63,69)	70(70,71)	70(63,77)	74(65,79)	65(63,65)	35(23,40)	42(33,47)	23(11,26)	67	67	73	70	68	68
GI	69(69,71)	63(61,65)	69(67,71)	67(58,74)	67(56,72)	74(72,74)	28(21,35)	42(30,42)	37(33,40)	70	63	67	68	64	70
X2	69(67,73)	63(61,65)	69(67,72)	72(63,81)	72(63,79)	74(72,77)	33(21,44)	47(42,58)	37(30,42)	67	61	67	69	66	71
GR	69(66,71)	63(61,66)	69(69,75)	67(65,74)	70(61,74)	74(65,77)	35(26,37)	44(28,49)	30(26,40)	67	62	69	68	65	71
DM	71(70,72)	61(59,67)	69(69,71)	67(63,72)	77(58,81)	70(58,74)	28(19,30)	49(40,58)	33(21,35)	73	59	70	71	67	70

The median (Me), (minimum (*m*), and maximum (*M*) for 20 runs are given for %CC, %TP, and %FP, whereas for sensitivity and specificity only the median values are given.

TABLE IV
THREE MOST IMPORTANT RISK FACTORS OF THE THREE SET OF MODELS INVESTIGATED GIVEN IN TABLE III FOR THE FIVE DIFFERENT SPLITTING CRITERIA USING RISK FACTORS BEFORE THE EVENT (B), AFTER THE EVENT (A), AND BEFORE AND AFTER (B + A)

	B			A			B+A		
MI									
IG	AGE	SMBEF	HxHTN	SBP	SMAFT	DBP	AGE	SMAFT	SBP
GI	AGE	HxHTN	SMBEF	SBP	SMAFT	DBP	AGE	SBP	SMBEF
X2	AGE	HxHTN	SMBEF	SMAFT	SBP	DBP	AGE	DBP	HxHTN
GR	AGE	HxHTN	SMBEF	SBP	SMAFT	DBP	SBP	SMAFT	HxHTN
DM	AGE	HxHTN	SMBEF	SBP	DBP	SMAFT	AGE	SBP	SMBEF
PCI									
IG	FH	AGE	HxDM	DBP	LDL	SMAFT	HxDM	DBP	FH
GI	AGE	HxHTN	FH	DBP	LDL	SMAFT	DBP	FH	HxHTN
X2	FH	HxHTN	HxDM	DBP	LDL	SMAFT	DBP	HxHTN	AGE
GR	FH	HxHTN	HxDM	DBP	SMAFT	LDL	HxDM	FH	DBP
DM	FH	HxHTN	HxDM	DBP	LDL	SMAFT	FH	DBP	HxDM
CABG									
IG	AGE	HxHTN	SMBEF	SMAFT	SBP	DBP	AGE	SMBEF	HxDM
GI	AGE	HxDM	SMBEF	SMAFT	SBP	DBP	AGE	SMBEF	HxDM
X2	AGE	SMBEF	HxDM	SMAFT	SBP	DBP	AGE	SMBEF	SMAFT
GR	AGE	HxDM	SMBEF	SMAFT	SBP	DBP	AGE	SMAFT	HxDM
DM	AGE	HxDM	SMBEF	SMAFT	DBP	SBP	AGE	SMAFT	HxDM

approximately the same with those who were smokers and did not have an episode.

For the MI models, there were 0/0 (0/0%), 28/7 (5.3/1.3%), and 330/163 (62.5/30.9%) subjects with event yes/no, with low, intermediate, and high risk, respectively. Moreover, the average event risk per rule ranged from 11.8% to 15.0%, i.e., all rules were classified as high risk (see Table IV). Also, there was no difference between the rule event risk for an MI event to occur versus not to occur.

B. PCI Models

For the PCI models, slightly better performance was obtained compared to the MI models. Better performance was obtained for the A and B + A models, with the median of %CC ranging from 65% to 67%. Again, similar performance was obtained for all splitting criteria with no significant difference.

The most important risk factors were for the before risk factors models, age, family history, history of hypertension and history of diabetes, for the after risk factors models, diastolic blood pressure, low density lipoprotein, and smoking after the event, and for the before and after risk factors models, history of diabetes, diastolic blood pressure, family history, history of hypertension, and age.

Based on the rules given in Table IV:

Rules 2.5–2.8 for diabetes subjects the number of PCI events increase with age (support increases from 2% to 20%).

For the PCI models, there were 0/0 (0/0%), 20/15 (3.8/2.8%), and 193/300 (36.6/56.8%) subjects with event yes/no, with low, intermediate, and high risk respectively. The average event risk per rule ranged from 11.7 to 13.9%, i.e all rules were classified as high risk (see Table IV). Also, there was no difference between the rule event risk for a PCI event to occur vs not to occur.

TABLE V
SELECTED RULES FROM MODELS GIVEN IN TABLE III (BASED ON THE CODING OF THE RISK FACTORS GIVEN IN TABLE II)

	SEX		AGE			FH		SM		HxHTN		HxDM		CLASS		SUP %	CONF %	EVENT RISK	
	M	F	1	2	3	4	Y	N	Y	N	Y	N	Y	N	Y	N			Y	N
	non Modifiable							Modifiable												
Risk factors before the event (MI)																				
1.1			+												+		19	79	11,8	12,6
1.2				+								+			+		22	76	12,4	11,4
1.3				+						+	+				+		10	67	12,6	12,4
1.4				+					+		+					+	17	68	13,5	13,2
1.5					+			+				+			+		20	63	12,7	12,9
1.6					+			+			+					+	23	59	12,8	13,3
1.7		+				+										+	11	69	12,5	13,2
1.8	+					+		+								+	24	61	12,1	13,4
1.9	+					+	+			+					+		7	64	12,6	12,9
1.10	+					+	+		+							+	10	67	15,0	14,3
Risk factors before the event (PCI)																				
2.1										+		+		+	+		29	71	11,7	12,1
2.2									+			+		+		+	35	64	12,3	12,4
2.3											+			+	+		72	65	12,8	13,0
2.4		+											+		+		13	67	13,1	12,9
2.5	+		+										+		+		2	100	13,1	12,0
2.6	+			+									+		+		10	86	13,1	13,8
2.7	+				+								+		+		21	67	13,1	13,3
2.8	+					+							+		+		20	93	13,3	13,9
Risk factors before the event (CABG)																				
3.1			+													+	20	94	11,5	11,9
3.2				+										+		+	34	79	12,7	12,4
3.3				+									+		+		14	67	13,8	13,2
3.4					+							+		+	+		16	64	13,0	12,5
3.5					+			+			+			+	+		16	57	12,7	12,8
3.6					+		+				+			+		+	19	69	13,3	12,7
3.7					+								+		+		28	71	13,0	13,3
3.8						+									+		53	70	13,4	12,9

C. CABG Models

Highest performance was obtained for the CABG models, with median of %CC in the region of 70%. As in the aforementioned two set of models, there was no significant difference in the models obtained with the different splitting criteria. The highest performance was obtained for the GR splitting criterion, for the B + A model, where the maximum %CC = 75%.

The most important risk factors based on Table IV were for the before risk factors models, age, history of hypertension, history of diabetes, and smoking before the event, for the after risk factors models, smoking after the event, systolic blood pressure, and diastolic blood pressure, and for the before and after risk factors models, age, smoking before the event, smoking after the event, and history of diabetes. Based on the rules given in Table V:

Rules 3.2 and 3.3:

1) CABG occurs usually in subjects aged between 51 and 60 years old when they have history of diabetes.

Rules 3.5 and 3.6:

1) Family history is not an important risk factor for CABG.

For the CABG models, there were 0/0 (0/0%), 9/26 (1.7/4.9%), and 206/287 (39/54.4%) subjects with event yes/no, with low, intermediate, and high risk respectively. Similar to the previous two models, the average event risk per rule varied very little, ranging from 11.5 to 13.3%, and all rules were classified as high risk (see Table IV). Also, there was no difference between the rule event risk for a CABG event to occur vs not to occur.

IV. DISCUSSION

The events investigated through this study were: MI, PCI, and CABG. Three classification models were developed based on decision trees for classifying MI, PCI, and CABG patients, where the highest percentage of correct classifications obtained were 66%, 75%, and 75%, respectively. Although different risk factors were obtained for the MI, PCI, and CABG models investigated, the most important risk factors, as extracted from the classification rule analysis were: sex, age, smoking, blood pressure, and cholesterol. It is important to note that the latter three risk factors can be modified; therefore the CHD risk of a subject may be reduced through a proper control of these factors. Furthermore, the importance of smoking in increased CHD risk was clearly illustrated.

The above findings and risk factors were also extracted by other investigators [35]. The EUROASPIRE study with EUROASPIRE surveys (I, II, III) involved various European populations and also included additional risk factors such as obesity. All Euroaspire surveys were reviewed together and combined results were extracted [4]. A general outcome was the fact that patients do not follow the advice and recommendations of their physicians. In comparison with the EUROASPIRE survey, our findings concerning the modifiable risk factors after the event are the following [4]:

1) 14% of subjects smoke after the event (16% in EUROASPIRE);
2) 22% of subjects had high blood pressure (26% in EUROASPIRE);
3) 34% of subjects had high total cholesterol (31% in EUROASPIRE); and
4) 45% of subjects had low-density lipoprotein (31% in EUROASPIRE).

In the EUROASPIRE survey, smoking, blood pressure, and cholesterol were found to be important risk factors [2], [4]. It was concluded that wide variations exist between 15 countries in the risk factor prevalence's and the use of cardioprotective drug therapies [3]. Also, there is still considerable potential throughout Europe to raise standards of preventive care in order to reduce the risk of recurrent disease and death in patients with CHD.

Furthermore, additional observations that could be extracted from the database investigated in this study regarding the nonmodifiable risk factors in comparison with EUROASPIRE survey [4] are the following:

1) 14% of subjects were female (24.7% in EUROASPIRE);
2) 9% of subjects were $\leq$50 years old (23.1% in EUROASPIRE);
3) 28% were between 51 and 60 years old (33.8% in EUROASPIRE);
4) 39% of subjects were between 61 and 70 years old (43.1% in EUROASPIRE); and
5) 24% of subjects were between 71 and 84 years old.

No female subject was under the age of 50 years old; only male subjects were found under this age.

Rea *et al.* [35] concluded that smoking was associated with an elevated risk for recurrent coronary events, whereas Gamberger

et al. [16] mention the relationship between the risk factors cholesterol and overweight.

Wang *et al.* [8] used the risk factors age, sex, cholesterol, HDL, blood pressure, diabetes, and smoking to predict CHD. They used the Framingham function and concluded that the traditional risk factors have different degrees of impact and/or than other factors are contributing to risk.

It should be noted that the results of our study based on a small city in the island of Cyprus are comparable with other studies, as it is known that traditional risk factors have different degrees of impact and/or that other factors are contributing to risk. A population-specific risk function is needed as also indicated by other investigators [8].

The values of risk computed for each subject were between 7% and 15.5% that fall into the range of none (0%) for low risk, 35 (6.6%) for intermediate risk, and 493 (93.4%) for high risk.

Although an average rule risk was computed for each rule, the values extracted for an event to occur or not are very close, not making possible the differentiation between high and low risk subgroups of subjects. This finding should somehow be expected, given that almost all of the subjects used for deriving the proposed models fall into the high, risk group. Thus, the proposed methodology should be further investigated by using a more heterogenous group of subjects, covering numerous cases of low and medium risk.

Ordonez [14] using the C4.5 decision tree algorithm and association rules for the prediction of cardiac disease based on 25 risk factors documented that association rules generally include simpler predictive rules than decision tree rules [15]. The usefulness of association rules in the analysis of CHD risk factors was also investigated by our group on a similar database with this study [36]. The results regarding the most important risk factors were similar.

Tsien *et al.* [17] in their study indicated that classification trees, which have certain advantages over logistic regression models, may perform similar to logistic regression models in the diagnosis of patients with MI.

The following five different criteria were investigated, information gain, gini index, likelihood ratio chi-squared statistics, gain ratio, and distance measure, that resulted in models with similar performance, with no significant difference between them. Thus any one of the splitting criteria investigated could be used for the datasets in this study. This finding is in agreement with this study for developing the decision tree models, that documented that the choice of splitting criteria does not make much difference on the tree performance [24], [32]. Also, the different splitting criteria, agreed on the most important risk factors. To the best of our knowledge no similar study was found in the literature comparing the five different criteria investigated in this study for the problem of CHD.

Concluding, comparing our findings with other studies: 1) a data mining system was proposed to extract rules for CHD events, 2) the rules extracted facilitated the grouping of risk factors into high and low risk factors, and 3) the rules extracted are associated with an event risk, however, this needs further investigation.

It is anticipated that data mining based on decision trees could help in the identification of risk subgroups of subjects for developing future events and it might be a decisive factor for the selection of therapy, i.e., angioplasty or surgery. Moreover, the extracted models and rules could help to reduce CHD morbidity and possibly, mortality. However, further investigation with larger datasets and other rule extraction algorithms and criteria are still needed.

REFERENCES

[1] British Heart Foundation. (2008, Mar. 8). European Cardiovascular Disease Statistics. [Online]. Available: http://www.heartstats.org/datapage.asp?id=7683

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

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

[4] Euroaspire stusy group, "Euroaspire III: A survey on the lifestyle, risk factors and use of cardioprotective drug therapies in coronary patients from 22 European countries," *Eur. J. Cardiovasc. Prev. Rehabil.*, vol. 16, no. 2, pp. 121–137, 2009.

[5] T. Marshall, "Identification of patients for clinical risk assessment by prediction of cardiovascular risk using default risk factor values," *Br. Med. Assoc. Public Health*, vol. 8, p. 25, 2008.

[6] W. B. Kannel, "Contributions of the Framingham Study to the conquest of coronary artery disease," *Amer. J. Cardiol.*, vol. 62, pp. 1109–1112, 1988.

[7] M. Karaolis, J. A. Moutiris, and C. S. Pattichis, "Assessment of the risk of coronary heart event based on data mining," in *Proc. 8th IEEE Int. Conf. Bioinformatics Bioeng.*, 2008, pp. 1–5.

[8] Z. Wang and W. E. Hoy, "Is the Framingham coronary heart disease absolute risk function applicable to Aboriginal people?" *Med. J. Australia*, vol. 182, no. 2, pp. 66–69, 2005.

[9] P. Brindle, J. Emberson, F. Lampe, M. Walker, P. Whincup, T. Fahey, and S. Ebrahim, "Predictive accuracy of the Framingham coronary risk score in British men: Prospective cohort study," *Br. Med. Assoc.*, vol. 327, pp. 1267–1270, 2003.

[10] S. Sheridan, M. Pignone, and C. Mulrow, "Framingham-based tools to calculate the global risk of coronary heart disease: A systematic review of tools for clinicians," *J. Gen. Intern. Med.*, vol. 18, no. 12, pp. 1060–1061, 2003.

[11] T. A. Pearson, S. N. Blair, S. R. Daniels, R. H. Eckel, J. M. Fair, S. P. Fortmann, B. A. Franklin, L. B. Goldstein, Ph. Greenland, S. M. Grundy, Y. Hong, N. H. Miller, R. M. Lauer, I. S. Ockene, R. L. Sacco, J. F. Sallis, S. C. Smith, N. J. Stone, and K. A. Taubert, "AHA guidelines for primary prevention of cardiovascular disease and stroke," *Circulation*, vol. 106, no. 3, pp. 388–391, 2002.

[12] S. M. Grundy, R. Pasternak, P. Greenland, S. Smith, and V. Fuster, "Assessment of cardiovascular risk by use of multiple-risk-factor assessment equations," *Amer. Heart Assoc.*, vol. 100, pp. 1481–1492, 1999.

[13] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, "Decision trees: An overview and their use in medicine," *J. Med. Syst.*, vol. 26, no. 5, pp. 445–463, 2002.

[14] C. Ordonez, "Comparing association rules and decision trees for disease prediction," in *Proc. Int. Conf. Inf. Knowl. Manage., Workshop Healthcare Inf. Knowl. Manage*. Arlington, VA, 2006, pp. 17–24.

[15] C. Ordonez, E. Omiecinski, L. de Braal, C. A. Santana, N. Ezquerra, J. A. Taboada, D. Cooke, E. Krawczvnska, and E. V. Garcia, "Mining constrained association rules to predict heart disease," in *Proc. IEEE Int. Conf. Data Mining (ICDM 2001)*, pp. 431–440.

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

[17] C. L. Tsien, H. S. F. Fraser, W. J. Long, and R. L. Kennedy, "Using classification trees and logistic regression methods to diagnose myocardial infraction," in *Proc. 9th World Congr. Med. Inf.*, vol. 52, pp. 493–497, 1998.

[18] R. B. Rao, S. Krishan, and R. S. Niculescu, "Data mining for improved cardiac care," *ACM SIGKDD Explorations Newslett.*, vol. 8, no. 1, pp. 3–10, 2006.
[19] J. Zavrsnik, P. Kokol, I. Maleiae, K. Kancler, M. Mernik, and M. Bigec, "ROSE: Decision trees, automatic learning and their applications in cardiac medicine," *Medinfo*, vol. 8, no. 2, p. 1688, 1995.
[20] K. Polat, S. Sahan, H. Kodaz, and S. Guenes, "A hybrid approach to medical decision support systems: combining feature selection, fuzzy weighted pre-processing and AIRS," *Comput. Methods Programs Biomed.*, vol. 88, no. 2, pp. 164–174, 2007.
[21] S. A. Pavlopoulos, A. Ch. Stasis, and E. N. Loukis, "A decision tree-based method for the differential diagnosis of aortic stenosis from mitral regurgitation using heart sounds," *Biomed. Eng. OnLine*, vol. 3, p. 21, 2004.
[22] C. A. Pena-Reyes, "Evolutionary fuzzy modeling human diagnostic decisions," *Ann. NY Acad. Sci.*, vol. 1020, pp. 190–211, 2004.
[23] K. Boegl, K.-P. Adlassnig, Y. Hayashi, T. E. Rothenfluh, and H. Leitich, "Knowledge acquisition in the fuzzy knowledge representation framework of a medical consultation system," *Artif. Intell. Med.*, vol. 30, no. 1, pp. 1–26, 2004.
[24] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, *Machine Learning, Neural and Statistical Classification*. West Susser, England: Ellis Horwood, 1994.
[25] J. R. Quinlan, in *C4.5 Programs for Machine Learning*, C. Schaffer, Ed. San Mateo, CA: Morgan Kaufmann, 1993.
[26] J. Han and M. Kamber, *Data Mining, Concepts and Techniques*, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2001.
[27] J. R. Quinlan, "Simplifying decision trees," *Int. J. Man-Mach. Stud.*, vol. 27, pp. 221–234, 1987.
[28] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Belmont, CA: Wadsworth Int. Group, 1984.
[29] F. Attneave, *Applications of Information Theory to Psychology*. New York: Holt, Rinehart, and Winston, 1959.
[30] R. Lopez de Mantras, "A distance-based attribute selection measure for decision tree induction," *Mach. Learn.*, vol. 6, pp. 81–92, 1991.
[31] T. Niblett, "Constructing Decision trees in noisy domains," in *Proc. 2nd Eur. Working Session Learn.*, 1987, pp. 67–78.
[32] L. Rokach and O. Maimon, *Data Mining with Decision Trees, Theory and Applications*. Singapore: World Scientific, 2008.
[33] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics*, vol. 1, pp. 80–83, 1945.
[34] P.-N. Tan, *Introduction to Data Mining*. Reading, MA: Addison-Wesley, 2006.
[35] T. D. Rea, S. R. Heckbert, R. C. Kaplan, N. L. Smith, R. N. Lemaitre, and B. M. Psaty, "Smoking status and risk for recurrent coronary events after myocardial infraction," *Ann. Int. Med.*, vol. 137, pp. 494–500, 2002.
[36] M. Karaolis, J. A. Moutiris, L. Papaconstantinou, and C. S. Pattichis, "Association rule analysis for the assessment of the risk of coronary heart events," in *Proc. 31st Annu. Int. IEEE Eng. Med. Biol. Soc. Conf.*, Minneapolis, MN, Sep. 2–6, 2009, pp. 6238–6241.

Minas A. Karaolis (M'08) was born in Cyprus, on July 1, 1964. He received the Diploma (Masters' equivalent) degree in computer science from the Technical University Carolo Wilhelmina Braunschweig, Braunschweig, Germany. He is currently working toward the Ph.D. degree in computer science from the Department of Computer Science, University of Cyprus, Nicosia, Cyprus.

He was engaged in the Ministry of Education, Nicosia, Cyprus, hold the positions of a High School Teacher, a Teaching Assistant in the Computer Science Department at the University of Cyprus, Nicosia, Cyprus. He was engaged for four years as an Assistant in the Computer Science Department, Technical University Carolo Wilhelmina Braunschweig, Braunschweig, Germany, where he was involved in the development of the automatic generation of loop plans for industrial buildings under the Preussag Company in Germany, using software engineering and databases. He was involved as a Teacher of computers for the special education programs of the Cyprus Ministry of Education, for five years. He was engaged in the Department of Information Technology, Bank of Cyprus, as a Programmer-Analyst. He was also involved in the technical support and training for Prisma Computers Ltd. and in the technical support and Microsoft training for AKTINA. His current research interests include data-mining applications and development of algorithms in medical diagnostic systems. He is the author or coauthor of more than nine publications in this area.

Joseph A. Moutiris received the M.D. degree from Medical School, University of Cluj, the Diploma in cardiology from Medical School, Imperial College, London, U.K., the M.Sc. degree in cardiology from the University of London, London, and the Ph.D. degree in medicine from Medical School, University of Warsaw, Warsaw, Poland.

He is currently a Consultant Cardiologist and Assistant Director of cardiology, in the Department of Cardiology, Paphos General Hospital, Paphos, Cyprus. He is also a Visiting Lecturer of cardiology at the School of Health Sciences, University of Nicosia, Nicosia, Cyprus. His research interests include management of coronary artery disease and especially secondary prevention, invasive cardiology, and cardiac pacing.

Demetra Hadjipanayi was born in Cyprus in 1983. She received the B.Sc. degree in computer science, and the M.Sc. degree in advanced information technology from the University of Cyprus, Nicosia, Cyprus, in 2009.

She is currently in the Department of Computer Science, University of Cyprus. Her M.Sc. thesis is on rule extraction of cardiovascular database using decision trees. Her final year project is on bounding volume of visual hulls. She has also been a Professional Services Consultant and Software Engineer with NCR since September 2006, where she has been engaged on numerous projects related to banking and insurance. She is also engaged on a project with the Insurance Companies Control Service at the Ministry of Finance.

Constantinos S. Pattichis (S'88–M'88–SM'99) was born in Cyprus on January 30, 1959. He received the Diploma degree as a Technician Engineer from the Higher Technical Institute, Nicosia, Cyprus, in 1979, the B.Sc. degree in electrical engineering from the University of New Brunswick, Fredericton, NB, Canada, in 1983, the M.Sc. degree in biomedical engineering from the University of Texas, Austin, in 1984, the M.Sc. degree in neurology from the University of Newcastle Upon Tyne, Newcastle Upon Tyne, U.K., in 1991, and the Ph.D. degree in electronic engineering from the University of London, London, U.K., in 1992.

He is currently a Professor with the Department of Computer Science, University of Cyprus, Nicosia. His research interests include e-health, medical imaging, biosignal analysis, and intelligent systems. He has been involved in numerous projects in these areas funded by EU, the National Research Foundation of Cyprus, the INTERREG and other bodies, with a total funding managed close to 5 million Euros. He is the author or coauthor of 52 refereed journal and 142 conference papers, and 19 chapters in books in these areas. He is the Co-Editor of the books *M-Health: Emerging Mobile Health Systems* (Springer, 2006) and of the *Information Technology in Biomedicine* (Piscataway, NJ: IEEE Press, to be published in 2011). He is the coauthor of the monograph *Despeckle Filtering Algorithms and Software for Ultrasound Imaging* (San Mateo, CA: Morgan Kaufmann, 2008).

Dr. Pattichis was the Guest Co-Editor of the Special Issues on *Emerging Health Telematics Applications in Europe*, *Biomedical Informatics*, and *Computational Intelligence in Medical Systems* of the IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE. He was the General Co-Chairman of the *Medical and Biological Engineering and Computing Conference* (MEDICON'98), and the IEEE Region 8 *Mediterranean Conference on Information Technology and Electrotechnology* (MELECON'2000), Program Co-Chair of the *IEEE Information Technology in Biomedicine*, ITAB06, and General Co-Chair of ITAB09 organized in Cyprus. Moreover, he has been an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, since 2000, he serves on the Editorial Board of the *Journal of Biomedical Signal Processing and Control*, and served as Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS (2005–2007). He served as Chairperson of the Cyprus Association of Medical Physics and Biomedical Engineering (1996–1998), and the IEEE Cyprus Section (1998-2000).